

2.2 Cosa è il World Wide Web

Trovare una definizione unica che permetta di descrivere il World Wide Web (chiamato anche WWW o Web) è un compito molto difficile, sono state create molte definizioni nel tempo che prendono in considerazione alcuni aspetti del web ma nessuna è esaustiva, una definizione abbastanza completa è:

- il WWW è l'iniziativa che mira a fornire un "wide-area hypermedia information retrieval" (un sistema di reperimento di informazioni su larga scala) per l'accesso ad un vasto insieme di documenti.

Nonostante il WWW sia cambiato molto dal momento della sua invenzione la definizione data nel 1991 da Tim Berners-Lee (l'inventore del Web) è tuttora attuale:

- W3 is a "distributed heterogeneous collaborative multimedia information system"¹.

La definizione di Web fornita dal sito del W3C Consortium è:

- The World Wide Web (known as "WWW", "Web" or "W3") is the universe of network-accessible information, the embodiment of human knowledge².

In effetti indipendentemente dalla definizione che si vuole scegliere il punto chiave è che il World Wide Web è stato in grado di fornire, tramite la rete di computer, un permanente accesso ad una varietà di informazioni, disponibili in diversi formati, tramite un aspetto molto semplice. Usando un interfaccia software sul web, come Internet Explorer, Mozilla, Netscape, Lynx, etc, il WWW ha cambiato il modo con cui la gente crea e visualizza le informazioni, è stata creata la prima vera rete ipertestuale³ globale.

Alla base del successo del WWW ci sono state molte sue caratteristiche, tra le quali la possibilità di raggiungere le informazioni da tutto il mondo, informazioni che possono essere nei formati più diversi, come esempio testo, figure, immagini, suoni, filmati, etc. Inoltre il WWW ha fornito tramite dei collegamenti la possibilità di connettere informazioni, spesso residenti in posti diversi, in maniera molto semplice. La semplicità d'uso è sicuramente un fattore chiave che ne ha permesso la diffusione e lo sviluppo che oggi

¹ <http://www.w3.org/Talks/General.html> (tda) Il W3 (il Web) è un "sistema distribuito, eterogeneo, collaborativo di gestione delle informazioni multimediali"

² <http://www.w3.org/WWW/> (tda) Il World Wide Web (conosciuto come "WWW", "Web" o "W3") è l'universo di informazioni accessibili via rete, l'incorporamento delle conoscenze umane.

³ La definizione di ipertesto verrà data più avanti per ora ci base sapere che l'ipertesto è un testo con dei riferimenti altre informazioni

conosciamo, infatti è sufficiente avere a disposizione un computer collegato ad internet con un browser web per poter navigare. A questa caratteristica è inoltre associata la leggibilità universale del web, ovvero la possibilità di potersi connettere ad internet tramite qualunque tipo di computer indipendentemente dal tipo di computer con cui le informazioni sono state scritte e dal computer dove le informazioni risiedono. Gli elementi che hanno reso il Web molto popolare non sono però nati insieme al web ma derivano da idee precedenti, idee riutilizzate poi a favore dello sviluppo del web stesso, tratteremo queste origini nel prossimo paragrafo.

Approfondimenti

- What are CERN's greatest achievements? - The World Wide Web
<http://public.web.cern.ch/Public/Content/Chapters/AboutCERN/Achievements/WorldWideWeb/WWW-en.html>
- Seminario online sul World Wide Web, 1991 Tim Berners-Lee
<http://www.w3.org/Talks/General.html>

2.3 Le origini del web

I concetti che stanno alla base del Web non sono molto nuovi, infatti alcune idee che ispireranno il WWW nascono nel 1945. Nel 1945, Vannevar Bush, uno studioso americano, pubblica un articolo intitolato "As We May Think" sul giornale "Atlantic Monthly". Nell'articolo, Bush esponeva l'idea di una macchina capace di memorizzare informazioni testuali e grafiche in modo tale che ogni pezzo di informazione potesse essere collegato con gli altri. La macchina immaginaria era chiamata MEMEX. Pensandoci un attimo non è molto diverso da quello che il web ha fatto circa 45 anni dopo.

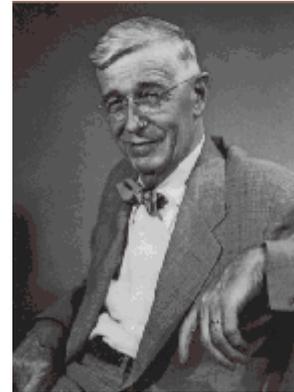


Figura 1-Vannevar Bush

In particolare non essendo ancora disponibili apparecchiature elettroniche MEMEX proponeva un dispositivo basato su un meccanismo foto-elettrico-meccanico per la memorizzazione di documenti, oltre alla possibilità di poter creare e seguire collegamenti tra documenti memorizzati su microfilm.

MEMEX era quindi il primo esempio di information retrieval system (sistema di reperimento delle informazioni), nato ancora prima di poter utilizzare l'elettronica e quindi i computer come strumento di lavoro. Inoltre nell'articolo si proponeva anche di creare dei collegamenti tra documenti, in pratica quelli che oggi chiamiamo collegamenti ipertestuali. Queste sue idee saranno riprese ed ampliate da importanti studiosi negli anni seguenti ed a Vannevar Bush si ispirerà anche Tim Berners-Lee l'inventore del WWW.

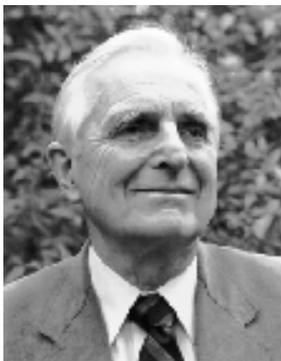


Figura 2-Douglas Engelbart

La storia delle idee che hanno portato al web passa anche da "oNLine System" (NLS) [NLS] un sistema proposto da Douglas Engelbart [ENGELBART] nel 1962. Engelbart fu uno dei seguaci dell'idea di Bush di creare una macchina che potesse aiutare l'uomo nel maneggiare le informazioni a sua disposizione. Engelbart lavorò per l'Ames Aeronautical Lab e per l'Augmentation Research Centre (ARC) a Stanford dove creò con alcuni colleghi (William K. English e John F. Rulifson) On-Line System (NLS), la prima implementazione al mondo di quello che fu poi chiamato ipertesto. Nel 1962 fu pubblicato l'articolo Augmenting Human Intellect: A Conceptual Framework [AHUCF], nel quale veniva proposto "oNLine System (NLS)", un sistema che introduceva importanti innovazioni tra cui: - il mouse - la modifica di documenti su monitor - il collegamento tra

oggetti (link tra documenti) - il protocollo per terminali multipli - la teleconferenza - il word processing e le presentazioni

La demo del progetto NLS sarà pronta nel 1968 ed ancora oggi su internet si trovano i video [NSLDEMO] che mostrano i primi prototipi delle innovazioni sopra citate.



Figura 3-Ted Nelson

Nel 1965, quindi qualche anno dopo la presentazione di NLS, Ted Nelson [NELSON] conia il termine hypertext (ipertesto); l'ipertesto secondo la definizione di Nelson è un testo con collegamenti (link). Il termine ipertesto non è, come si potrebbe pensare, un termine prettamente informatico, infatti una enciclopedia con riferimenti ad altri termini è essa stessa un esempio di ipertesto. Gli studi di Nelson sugli ipertesti nascono da un progetto, nato nel 1960, di un sistema integrato di gestione della letteratura, ovvero dei contenuti, dei riferimenti espliciti ed impliciti, e del processo che sostiene la produzione letteraria (creazione, pubblicazione, modello economico).

Il sistema (chiamato Xanadu [XANADU]), come evoluto alla metà degli anni Ottanta, era basato su server che costruivano documenti virtuali basati su pezzi di testo di lunghezza arbitraria, e client che permettevano la creazione di link.

Approfondimenti

- A Little History of the World Wide Web
<http://www.w3.org/History.html>
- As We May Think - by Vannevar Bush
<http://www.theatlantic.com/doc/194507/bush>

2.4 La nascita del WWW

Alla fine degli anni '80 Il CERN (European Organization for Nuclear Research) [CERN] ⁴doveva affrontare il problema della perdita di informazioni dovuta principalmente al fatto che le informazioni erano distribuite su sistemi diversi consultabili con programmi e competenze specifiche. Ogni gruppo di lavoro aveva propri sistemi per la gestione delle informazioni ed i sistemi non potevano interagire tra loro.

Anche se gli scienziati del CERN spendevano molto tempo nella sede del CERN, spesso lavoravano in università e laboratori dislocati nelle loro nazioni. Avere un buon sistema di scambio di informazioni era chiaramente fondamentale. Insieme ai mezzi tradizionali (pubblicazioni su riviste internazionali, conferenze, scambio di idee nelle "coffee room", ...) gli scienziati del CERN avevano cominciato da anni a scambiarsi le informazioni in maniera elettronica, distribuendo in questo formato dati, software e documentazione. Inoltre dal 1980 in poi l'utilizzo della posta elettronica contribuì molto allo scambio di informazioni. Nonostante la possibilità di scambiare informazioni in formato elettronico negli anni '80 c'erano comunque molti ostacoli allo scambio effettivo dei dati. Erano presenti una vasta varietà di tipi di computer e di sistemi di rete con praticamente nessuna caratteristica in comune, inoltre le informazioni erano di tipo diverso e potevano essere accedute in modi molto diversi tra loro, creando una forte sensazione di sconforto ed inefficienza negli utenti. Tutto questo creò terreno fertile all'invenzione del World Wide Web, il cui scopo principale era permettere agli scienziati di accedere alle informazioni derivanti da qualunque tipo di sistema in una maniera coerente e semplice⁵.

Il CERN necessitava quindi di un modello che rispondesse alle problematiche sopra riportate ed in generale che avesse i seguenti requisiti:

- essere accessibile via rete
- permettere l'accesso da sistemi eterogenei (VM/CMS, Macintosh, VAX/VMS, Unix)
- essere non centralizzato
- fornire l'accesso a dati esistenti
- avere aree private
- avere l'accesso a dati anche in fase di "lavori in corso"

⁴ CERN Home page <http://www.cern.ch>

⁵ What are CERN's greatest achievements? - Why from CERN?
<http://public.web.cern.ch/Public/Content/Chapters/AboutCERN/Achievements/WorldWideWeb/WhyCERN/WhyCERN-en.html>

Tim Berners-Lee [Berners-Lee], un giovane fisico che lavorava al CERN, propose in un rapporto interno un sistema distribuito per la gestione delle informazioni che rispondeva ai requisiti sopra citati. Il modello proposto è illustrato nella figura 5 nella quale si vede che l'informazione era posta al centro del modello ed indicata in con la dicitura "This document"; il documento era descritto tramite un ipertesto (la nuvoletta "Hypertext") e descriveva la proposta Mesh (la nuvoletta "Mesh") la quale a sua volta proponeva un sistema di unificazione dei CERNDOC, VAX/NOTES... Il documento ipertestuale (Hypertext) inoltre includeva altre informazioni riferite tramite collegamenti ipertestuali.



Figura 4 – Tim Berners-Lee

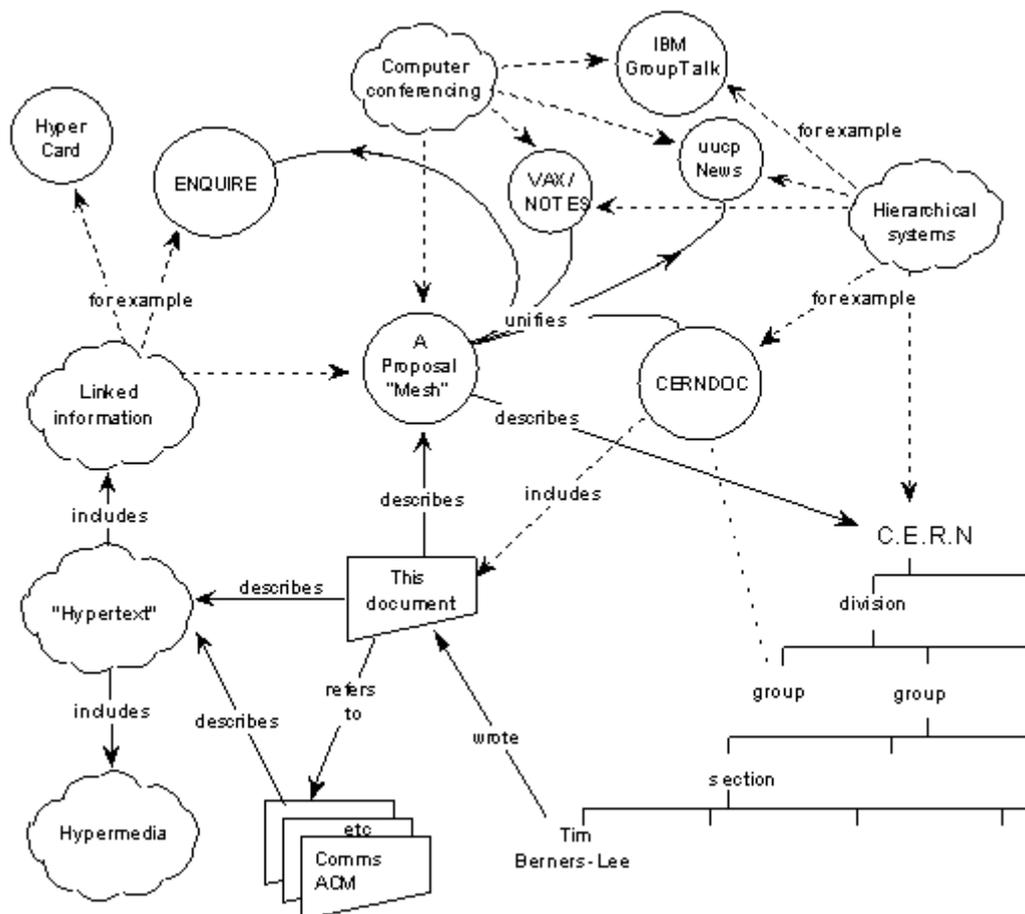


Figura 5 - Tim Berners-Lee propone un sistema distribuito di gestione delle informazioni

Tim Berners-Lee e Robert Cailliau per rafforzare la loro proposta scrivono il primo browser ipertestuale che fu chiamato WorldWideWeb. Il WorldWideWeb aveva integrato sia il browser che l'editor ed aveva quindi anche la possibilità di creare nuovi documenti e di modificare quelli esistenti tramite un'interfaccia visuale. Screenshot del WorldWideWeb sono mostrate sia in Figura 6 che in Figura 7. Si può notare anche come si potessero creare link tra i documenti e modificare lo stile delle pagine.

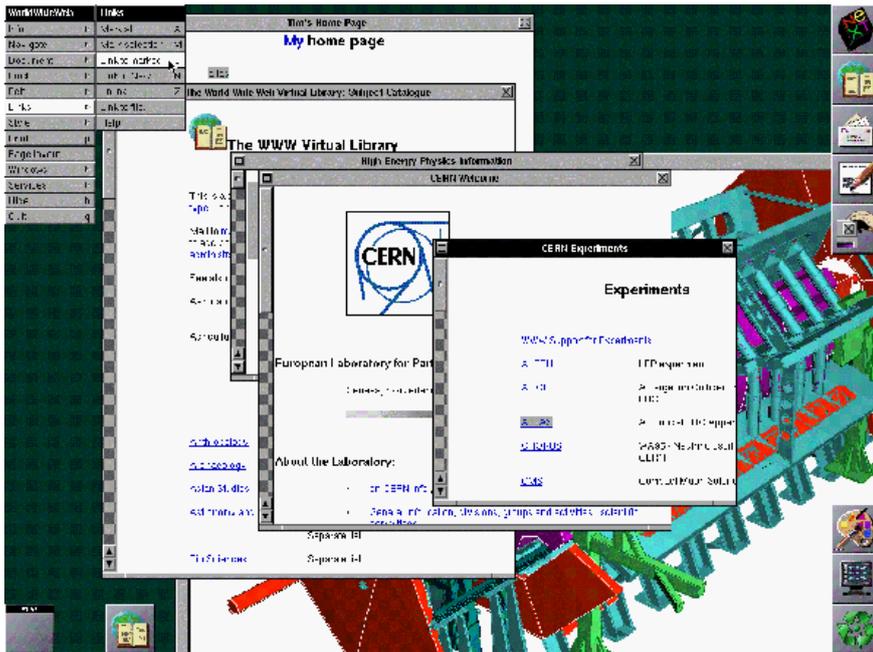


Figura 6 - Screenshot del browser WorldWideWeb



Figura 7 - WorldWideWeb: Browser ed Editor

Approfondimenti

- The original proposal of the WWW, HTMLized <http://www.w3.org/History/1989/proposal.html>

2.5 Storia del WWW

Gli anni 1989 e 1990 sono stati un periodo di divulgazione del progetto del WWW, progetto che finalmente dal 1991 in poi si è evoluto ad una velocità impressionante. Essendo molto difficile descrivere la storia del web dal 1991 in poi di seguito sono riportati gli avvenimenti principali che si sono susseguiti di anno in anno permettendo lo sviluppo del web fino a quello che oggi conosciamo.

1991 nascono i primi browser e server multi-piattaforma; viene organizzato il primo seminario sul WWW;

1992 il browser "Viola" [VIOLA] aggiunge la possibilità di vedere immagini (oltre il testo);

1993 inizio dell'anno: ci sono oltre 50 server web conosciuti; settembre: NCSA [NCSA] rilascia il browser Mosaic; fine dell'anno: incremento dell'utilizzo del WWW, ci sono oltre 200 server web conosciuti;

1994 aprile: fondata la Mosaic Communication Corporation che produce il browser Mosaic Netscape (il primo Netscape); maggio: si tiene la prima conferenza internazionale sul WWW; fine dell'anno: Rilasciata la versione 2.0 di Netscape; (integrato il client di posta ed il newsreader, supporto per i frames, i div, compatibilità con le gif, etc.);

1995 aprile: viene rilasciata la prima versione pubblica (la 0.6.2) di Apache [APACHE]; rilasciato Internet Explorer 2;

1996 Microsoft produce Internet Explorer 3, Netscape Communication produce Netscape 3 (comincia il supporto ai CSS);

1997 nasce XML, il linguaggio di markup che diventerà uno dei mattoni fondamentali del web; rilasciati Netscape 4, Explorer 4, Opera 3 (cresce il supporto ai CSS, nasce il DHTML ed il DOM, Microsoft introduce le ActiveX);

1998 quinta generazione dei browser (Explorer 5, Netscape 5). Nasce il progetto Mozilla [MOZILLA].

Dal 1998 in poi la ricerca sul Web si sposterà verso la famiglia delle tecnologie XML. Le tecnologie HTML e CSS avranno una minore evoluzione ma in compenso si svilupperanno quelle tecnologie che renderanno possibile una maggiore integrazione delle varie tecnologie web; tra le novità che verranno introdotte possiamo per esempio citare i web services, il semantic web, XForm, XQuery, XSL/XSLT, l'iniziativa WAI, etc. Il lettore interessato alle tecnologie basate su XML può sicuramente trovare molta documentazione sul sito del W3C [W3C].

Approfondimenti

- W3C 10th anniversary - timeline graphic
<http://www.w3.org/2005/01/timelines/timeline-2500x998.png>
- quirksmode.org - A history of browsers
<http://www.quirksmode.org/browsers/history.html>
- Extensible Markup Language (XML) <http://www.w3.org/XML/>

2.6 Il web ed internet

Il web ed internet⁶ sono due cose distinte, il web è solamente uno dei servizi di comunicazione basati su internet. La relazione tra le due entità può essere spiegata con una semplice analogia utilizzando la rete stradale. In internet, come sulla rete stradale, sono essenziali tre elementi: la connessione fisica (i cavi e le strade), i comportamenti comuni (il codice stradale e i protocolli di rete) e i servizi (il servizio di posta terrestre e il WWW).

2.6.1 Le connessioni fisiche: cavi e strade

I cavi sono infrastrutture passive, mantenute da enti pubblici e da compagnie di telecomunicazioni. I cavi hanno differenti capacità: una singola linea telefonica come le normali linee telefoniche di casa possono trasportare circa 56 kilobytes al secondo, l'equivalente di una pagina di testo con qualche immagine per secondo. Le fibre ottiche possono trasportare migliaia di milioni di bytes al secondo. Anche se i cavi possono essere di differenti tipi e le loro congiunzioni possono essere molto complicate, questi sono tutti interconnessi tra di loro. Per tornare alla nostra analogia, sulle nostre strade è possibile spostarsi da una città all'altra passando attraverso tipi di strade diverse: strade urbane, extraurbane, autostrade, etc, e tutte queste strade sono interconnesse tra di loro. Tramite le diverse connessioni fisiche è quindi per esempio possibile raggiungere con il nostro computer di casa un computer in Australia o in America, attraverso una serie contigua di collegamenti fisici e di nodi di interscambio che permettono di passare da un collegamento all'altro.

2.6.2 Le regole di comportamento comuni

Connettere un computer al sistema di connessioni di internet non è però sufficiente: per poter dialogare con gli altri è necessario concordare con loro delle regole di comportamento (di comunicazione), così come noi facciamo quando guidiamo sulle strade. Internet è come il codice delle strade: i computer devono usare i collegamenti fisici in accordo a delle regole concordate. Migliaia di macchine possono utilizzare le strade anche se hanno destinazioni diverse, non ci saranno problemi fino a quando ognuno

⁶ <http://public.web.cern.ch/Public/Content/Chapters/AboutCERN/Achievements/WorldWideWeb/WebInternet/WebInternet-en.html>

rispetterà le regole del codice stradale. Internet trasferisce le informazioni in piccoli pacchetti scambiati tra computer. Per poter scambiare questi pacchetti i computer sono soggetti a delle regole: devono usare lo stesso protocollo di comunicazione. Un protocollo di comunicazione è qualcosa con cui tutti siamo abituati ad aver a che fare quando abbiamo una conversazione con qualcuno: durante la conversazione le persone sanno quando possono cominciare a parlare, quando fermarsi, come fare per incoraggiare le altre persone a cominciare a parlare e così via; questo è il protocollo di comunicazione "implicito" utilizzato dagli umani. I computer scambiano le informazioni utilizzando un insieme di regole e di comportamento, per essere connesso ad internet un computer deve rispettare i protocolli di internet. Il protocollo più comune utilizzato è il Transmission Control Protocol / Internet Protocol o TCP/IP⁷.

2.6.3 Servizi per tutti

Una volta che si hanno le connessioni fisiche ed i protocolli di rete è possibile comunicare con gli altri computer. Ma cosa possono comunicarsi i computer tra loro? Le strade possono essere utilizzate per guidare da solo, per viaggiare sui bus con le altre persone, per trasportare materie prime, la posta terrestre, etc. Allo stesso modo tramite internet possono funzionare diversi servizi: la posta elettronica, il trasferimento di file, le connessioni a server remoti, etc. il World Wide Web è appunto uno dei servizi che funziona tramite internet, può essere visto come una distribuzione personalizzata delle informazioni: alla tua richiesta, il WWW ti consegna il documento richiesto.

Approfondimenti

- Internet Society - Home page <http://www.isoc.org/>
- The Internet Engineering Task Force - Home page <http://www.ietf.org/>
- Internet Assigned Numbers Authority (IANA) - Home page <http://www.iana.org/>
- Internet Society - All about the internet <http://www.isoc.org/internet/infrastructure>

⁷ A TCP/IP Tutorial, gennaio 1991 <http://www.ietf.org/rfc/rfc1180.txt>

2.7 Come funziona il web

Il web è l'universo di informazioni disponibili con un click del mouse⁸. Per usarlo c'è bisogno di un computer, una connessione a internet ed un browser.

Quando utilizzi il tuo browser, lui trova e mostra pagine di informazioni. Il ruolo del browser è quello di interpretare il linguaggio delle pagine web (HTML, ...) e trasformarlo in testo e immagini che puoi vedere sul tuo schermo. Se hai bisogno di ulteriori informazioni, tutto quello che hai da fare è cliccare su un iperlink.

I documenti sono memorizzati in computer chiamati "server" i quali forniscono i documenti agli utenti che possono visualizzarli richiedendoli tramite il loro computer, chiamato client. Questo modello di computer che effettuano delle richieste e computer che rimangono in attesa di richieste e forniscono le risposte è chiamato modello client/server. I computer che partecipano a questo meccanismo sono naturalmente collegati tramite internet. Quindi quando tramite il tuo browser clicchi su un iperlink, il tuo computer richiede ad un server di restituirgli un documento che ti verrà poi mostrato nel browser.

Per esempio partendo dalla homepage del CERN <http://www.cern.org> in svizzera, puoi cliccare ed ottenere un documento che è in un laboratori di fisica dall'altra parte del mondo. Tutte le informazioni sembrano essere in una piccola scatola davanti a te ma in realtà sono sparse in tutto il mondo.

Il web fa anche un buon uso della rete (internet): quando clicchi su iperlink il tuo browser "ordina un documento" ad un altro computer, lo riceve e poi te lo mostra. Tu sei poi libero di guardarlo subito oppure quando vuoi, senza ulteriore consumo di risorse di rete.

Il web può essere utilizzato per iniziare dei processi sia lato client che lato server. Una richiesta può per esempio avviare un ricerca in un database di un server, il quale restituisce un documento generato tramite i i risultati della ricerca. Un documento restituito ad un client può essere in un formato che induce il browser ad avviare un programma esterno al browser stesso per interpretarlo (come per esempio Acrobat Reader per il file PDF). L'abilità del web di negoziare il formato delle informazioni tra client e server rende possibile poter scambiare informazioni di qualunque tipo se il client possiede il software appropriato per poter maneggiare l'informazione ricevuta. Questo rende possibile che video, suono ed ogni altro formato sia accessibile agli utenti senza la necessità che una singola applicazione (il browser) debba essere in grado di interpretare ogni cosa.

⁸ How the Web works <http://public.web.cern.ch/Public/Content/Chapters/AboutCERN/Achievements/WorldWideWeb/HowWorks/HowWorks-en.html>

Riassumendo gli attori del web sono:

- gli utenti (user) le persone che accedono alle informazioni
- i browser i programmi utilizzati per accedere alle informazioni sul web.
- i client i dispositivi con i quali si accede alle risorse del WWW (computer, palmari, telefoni, etc). Con client, a seconda del contesto di può intendere il browser, oppure il dispositivo.
- server i computer attraverso il quali si mettono a disposizione delle risorse (informazioni o servizi) ad altri server o ai vari client

2.8 Principi fondamentali del WWW

Quello che Tim Berners-Lee fece quando ideò il web fu quello di utilizzare alcuni concetti basi dai suoi predecessori (V. Bush, D. Engelbart, T. Nelson, etc) e di renderli possibili tramite protocolli e tecnologie semplici da utilizzare.

I concetti fondamentali sui quali il web si basa sono:

- La leggibilità universale
- Gli ipertesti
- La ricercabilità dei documenti
- Il modello client-server
- La negoziazione del formato

2.8.1 Leggibilità universale

È stata uno dei principi ispiratori del WWW che ha permesso l'integrazione di informazioni derivanti da fonti diverse. Le informazioni disponibili sul web possono essere reperite:

- da ogni tipo di computer
- da ogni paese nel mondo tramite il web
- da qualunque persona (autorizzata)
- tramite un semplice programma di navigazione

2.8.2 Gli ipertesti

Gli ipertesti secondo la definizione data da Nelson sono dei organizzati in modo non lineare, cioè che danno, tramite gli iperlink, la possibilità di saltare da una parte all'altra del testo senza seguire un percorso lineare. Il web in realtà è rimasto per molto poco tempo solo con testo ma il concetto di ipertesto si è evoluto in quello di ipermedia, quindi non solo testo ma anche immagini, video, etc. Gli ipertesti, e gli ipermedia permettono di:

- navigare tra le informazioni in modo non sequenziale.
- riferire informazioni dislocate in posizioni diverse

2.8.3 La ricercabilità delle informazioni

I documenti messi a disposizione tramite il Web dovevano essere ricercabili. Il web nasce già con le basi per lo sviluppo dei motori di ricerca e questo è permesso da:

- reperibilità permanente dei documenti
- accessibilità dei documenti da parte di indicizzatori di contenuti.

2.8.4 Modello client-server

Come abbiamo visto nel paragrafo precedente il web segue il modello client-server dove le informazioni sono:

- memorizzate sui Server Web
- raggiungibili tramite un client Web

I client effettuano richieste ai server Web per ottenere le informazioni desiderate.

2.8.5 La negoziazione del formato

Nel modello del Web si doveva permettere lo scambio di documenti memorizzati nei formati più diversi. Il meccanismo della negoziazione del formato permette di scambiare tramite il web file di vario tipo, tra cui: HTML, PDF, immagini (png, jpg, gif), audio, video. La negoziazione del formato è il meccanismo tramite il quale insieme alle informazioni che si scambiano sulla rete (file di vario tipo, png, jpg, etc.) viene anche inviato il tipo dell'informazione che stiamo passando (il content-type) attraverso il quale i vari client possono decidere l'applicazione che devono utilizzare per "trattare" il contenuto ricevuto. Per esempio se il client riceve un file con content-type "application/pdf" probabilmente tenterà di visualizzare il file con un visualizzatore di PDF.

2.8.6 Le tecnologie ed i protocolli fondamentali del WWW

Il WWW si basa su alcuni protocolli e tecnologie fondamentali:

- HTTP (HyperText Transfer Protocol)
- URI, URL (Uniform Resource Identifier, Uniform Resource Locator)

- HTML (HyperText Markup Language)

Queste tre tecnologie stanno alla base del funzionamento del Web e servono per scambiare le informazioni (HTTP), identificare univocamente le informazioni (URI, URL), e rappresentare le informazioni stesse (HTML). L'HTTP è un protocollo e come tale definisce il formato e l'ordine dei messaggi scambiati tra due o più entità, così come le azioni intraprese nella trasmissione o ricezione di un messaggio. Le URI (e URL) identificano le informazioni presenti sul web, mentre l'HTML definisce la rappresentazione dell'informazione. Nei paragrafi che seguono le tre tecnologie/protocolli verranno trattate in maggiore dettaglio.

Approfondimenti

- Wikipedia: Client-server http://en.wikipedia.org/wiki/Client_server

2.9 HTTP

Il web utilizza un proprio protocollo per trasferire le informazioni, l'HTTP (Hypertext Transfer Protocol). HTTP trasmette le proprie informazioni tramite una connessione TCP/IP stabilita tra due computer collegati tra di loro (tipicamente tramite internet, ma potrebbe anche essere una rete diversa come per esempio una rete di computer privata).

HTTP fornisce:

- un sistema veloce di trasferimento dati
- un sistema senza stati (stateless)
- un architettura estendibile
- un sistema per la negoziazione dei differenti tipi di dato trasferiti

Di seguito sono riportate le varie fasi della comunicazione client/server necessarie per comporre una pagina web. Nella prima fase (Figura 8) il client richiede al server il testo del documento (il suo codice HTML) all'interno del quale possono esserci riferimenti ad altri oggetti da richiedere al server (immagini, CSS, file esterni,...). Nella seconda fase (Figura 9) il client ogni qual volta trova nel testo del documento degli oggetti esterni necessari per visualizzare la pagina (come per esempio immagini) le richiede al server con un'altra richiesta HTTP.

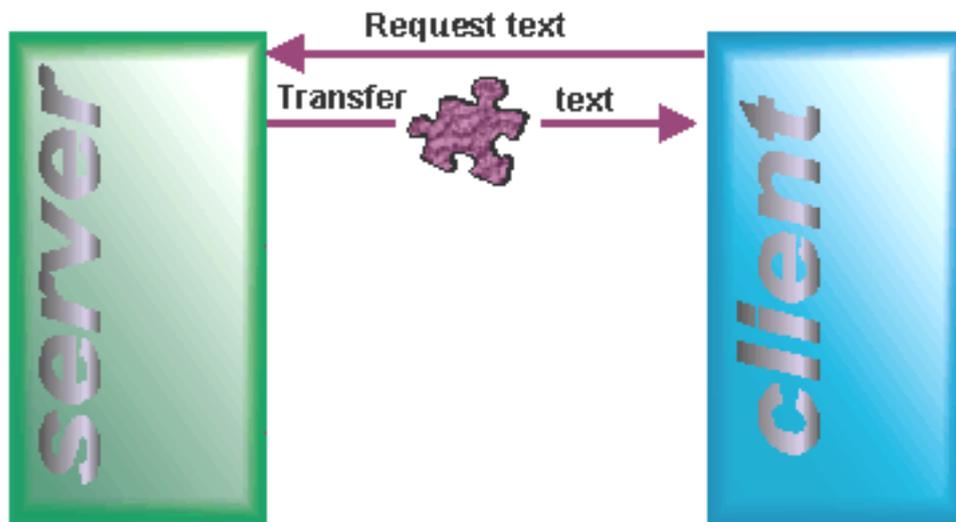


Figura 8 - Richiesta della pagina

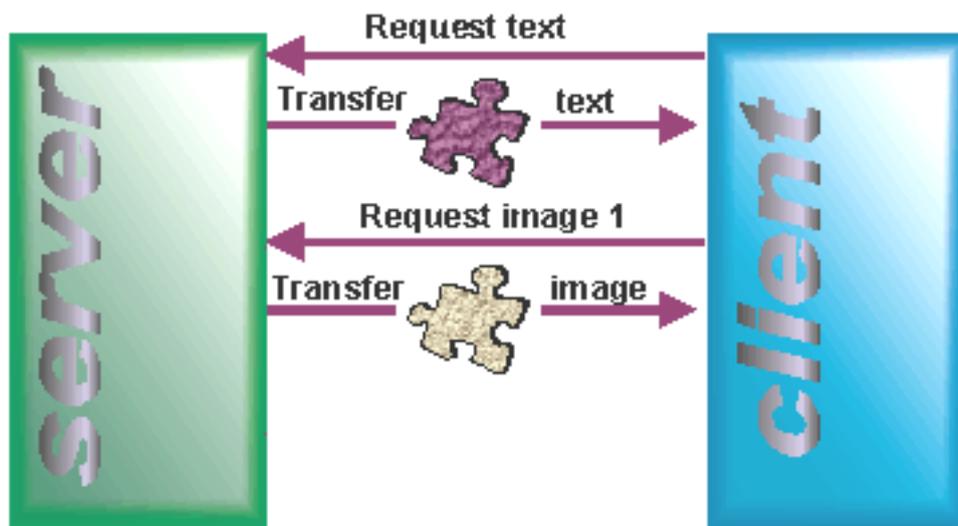


Figura 9 - Invio della pagina e richiesta degli altri elementi della pagina

Nella Figura 10 viene evidenziato che possono essere necessari molti oggetti per formare una pagina web completa. Infine (Figura 11) il client mostra all'utente la pagina composta da tutti gli elementi richiesti.

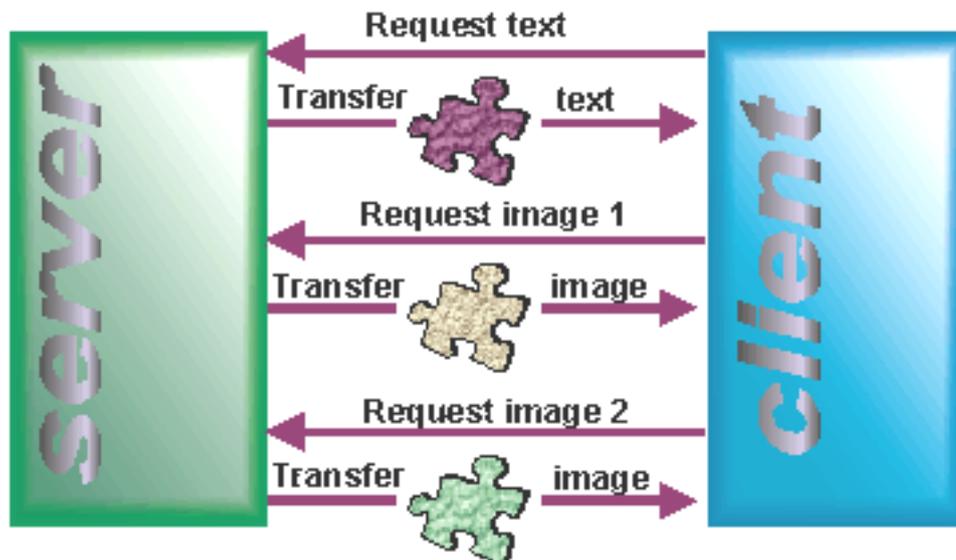


Figura 10 - Richiesta ed invio di tutti gli elementi della pagina

La Figura 11 mostra come, alla fine del processo di trasferimento dei vari file, il client riesce a ricostruire, e visualizzare, la pagina web completa.

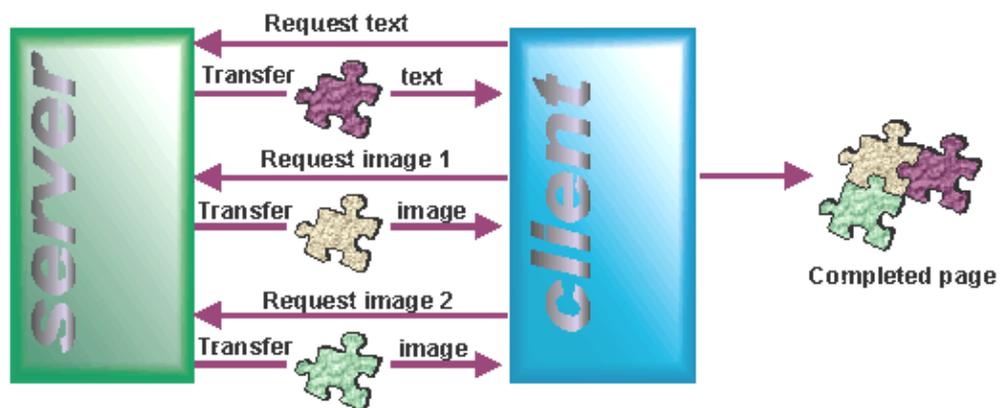


Figura 11 - Visualizzazione della pagina nel client

Approfondimenti

- Hypertext Transfer Protocol - HTTP/1.1, giugno 1999
<http://www.ietf.org/rfc/rfc2616.txt>

2.10 Breve storia dell'HTTP

L'HTTP nella sua versione originaria, del 1990, è conosciuto come HTTP 0.9. La versione HTTP 1.0 pubblicata nel 1995, formalizza l'uso comune del protocollo che si era sviluppato fino a quegli anni. L'attuale versione, HTTP 1.1, formalizzata nel 1996 e riedita nel 1999, è descritta nella RFC 2616 [RFC2616].

2.10.1 Cosa è una RFC?

Le RFC (Request for Comments) sono una serie di documenti tecnici e note organizzative che trattano di Internet (originariamente ARPANET⁹), cominciate nel 1969. Le RFC discutono una serie di aspetti delle reti di computer, tra cui protocolli, procedure, programmi, concetti, così come note ed opinioni. I documenti con le specifiche ufficiali dell'insieme dei protocolli Internet sono definite dall'Internet Engineering Task Force e dal Internet Engineering Steering Group (IESG) e sono registrate e pubblicate con il processo standard delle RFC [RFC]. Il risultato è che il processo di pubblicazione delle RFC gioca un ruolo importante nel processo di standardizzazione di Internet.

2.10.1.1 L'RFC-Editor

L'RFC Editor è il responsabile della pubblicazione delle RFC e della revisione editoriale finale dei documenti. RFC Editor mantiene anche un file chiamato "RFC Index" che può essere trovato online. Per oltre 30 anni, l'RFC Editor è stato Jon Postel[POSTEL], oggi l'RFC Editor è un piccolo gruppo fondato dalla "Internet Society (ISOC) [ISOC]".

⁹ Arpanet è il precursore di Internet, una vasta rete di computer create da United States Defense Advanced Research Project Agency (ARPA). Arpanet nasce nel 1969 come sistema per lo studio di nuove tecnologie di interconnessione di rete tra università e centri di ricerca americani. Maggiori informazioni possono essere trovate sul sito <http://www.darpa.mil>.

2.11 HTTP 1.1

HTTP 1.1 è retro compatibile con HTTP/1.0; un Web server che utilizza HTTP 1.1 può parlare con un browser che utilizza la versione 1.0 e viceversa. Il protocollo HTTP (sia nella versione 1.0 che 1.1) fornisce le regole per le interazioni tra client e server, tra cui:

- Message i messaggi tra client e server (gli header, il corpo del documento)
- Request le regole per fare le richieste al server (Get, Post, Head, etc., i parametri accettati)
- Response le regole per fornire le risposte al client (le date di modifica, la cache, dimensioni, etc.)
- Connection i metodi per stabilire le connessioni tra client e server
- Status Code i codici di stato (200, 404, etc..)

2.11.1 HTTP 1.1: Status Code più comuni

I codici di stato sono molto utili per verificare il corretto dialogo tra client e server e quindi anche la validità delle nostre pagine HTML. I codici di stato (Status Code) più utilizzati sono:

200 - OK

La richiesta è corretta, l'output è restituito nel corpo del messaggio (message body).

404 - Not Found

La risorsa richiesta non esiste.

301 - Moved Permanently

La risorsa non è più disponibile a questo indirizzo.

500 - Server Error

Si è verificato un errore sul server. La causa più comune è uno script sul server che fallisce.

2.11.2 HTTP 1.1: categorie di Status Code

Gli Status code¹⁰ [1] si suddividono in categorie:

1xx

sono messaggi solo di tipo informativo

2xx

indicano una richiesta in qualche modo soddisfatta

3xx

re-direzionano il client ad un altro indirizzo

4xx

indicano un errore da parte del client

5xx

indicano un errore da parte del server

Elenco degli status code in HTTP 1.1

100	Continue	404	Not Found
101	Switching Protocols	405	Method Not Allowed
200	OK	406	None Acceptable
201	Created	407	Proxy Authentication Required
202	Accepted	408	Request Timeout
203	Non-Authoritative Information	409	Conflict
204	No Content	410	Gone
205	Reset Content	411	Length Required
206	Partial Content	412	Precondition Failed
300	Multiple Choices	413	Request Entity Too Large
301	Moved Permanently	414	Request URI Too Long
302	Moved Temporarily	415	Unsupported Media Type
303	See Other	416	Requested Range Not Satisfiable
304	Not Modified	417	Expectation Failed
305	Use Proxy	500	Internal Server Error
306	(not Used)	501	Not Implemented

¹⁰ Trattazione completa degli status code in http può essere reperita all'indirizzo <http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html>.

100	Continue	404	Not Found
307	Temporay Redirect	502	Bad Gateway
400	Bad Request	503	Service Unavailable
401	Unauthorized	504	Gateway Timeout
402	Payment Required	505	HTTP Version Not Supported
403	Forbidden		

2.11.3 Richieste multiple

Una delle differenza principali tra l'HTTP 1.1 e le precedenti versioni è la possibilità di fare richieste multiple. Effettuando più richieste contemporaneamente si diminuisce il "round-trip time" globale ovvero il tempo che intercorre tra l'inizio della richiesta e la fine della risposta, come si vede dalla Figura 12 il tempo totale per servire tutte le richieste è molto minore in caso di richieste multiple.

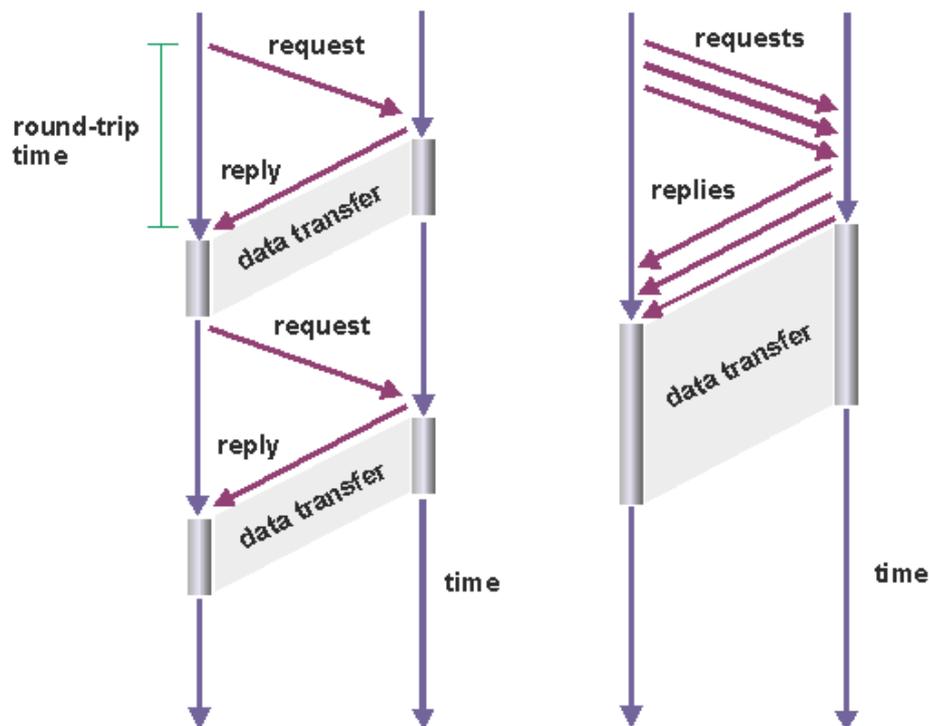


Figura 12 – Differenze nelle richieste multiple tra HTTP1.0 (sinistra) e HTTP1.1 (destra)

2.11.4 Connessione permanente

Un'altra delle innovazioni dell'HTTP 1.1 rispetto alle precedenti versioni è la possibilità di mantenere una connessione permanente. Questa diminuisce notevolmente il tempo dovuto alle connessioni mantenendo la connessione aperta per farvi transitare tutte le altre connessioni virtuali, vedi Figura 13.

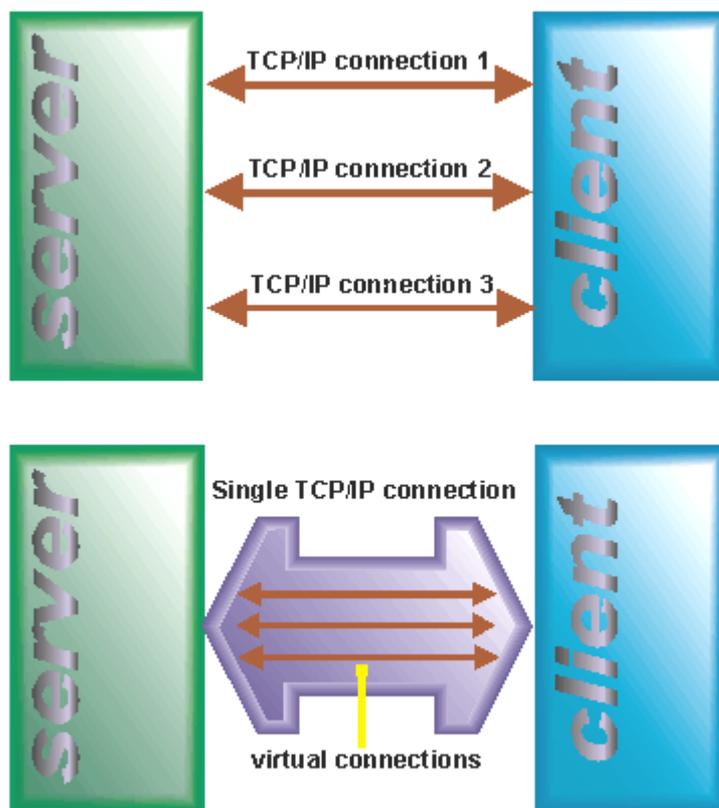


Figura 13 – Differenze tra connessioni singole e connessione permanente. HTTP1.0 (figura in alto) e HTTP1.1 (figura in basso)

2.12 URI

Le risorse sul web sono identificate tramite una URI[RFC3986] (Uniform Resource Identifier). L'URI, come dice il nome stesso, identifica in modo univoco una risorsa sul Web. Le URI possono essere utilizzate per identificare risorse diverse da pagine web, l'esempio di Figura 14 mostra come la risorsa "Oaxaca Weather Info" sia identificata tramite l'URI <http://weather.example.com/oaxaca>, mentre la sua rappresentazione sia la pagina HTML mostrata in figura. Le URI sono un concetto non legato solamente alle risorse Web, per esempio il codice ISBN di un libro è un esempio di un identificatore univoco (un URI) che però non si riferisce ad una risorsa sul web ma ad un libro.

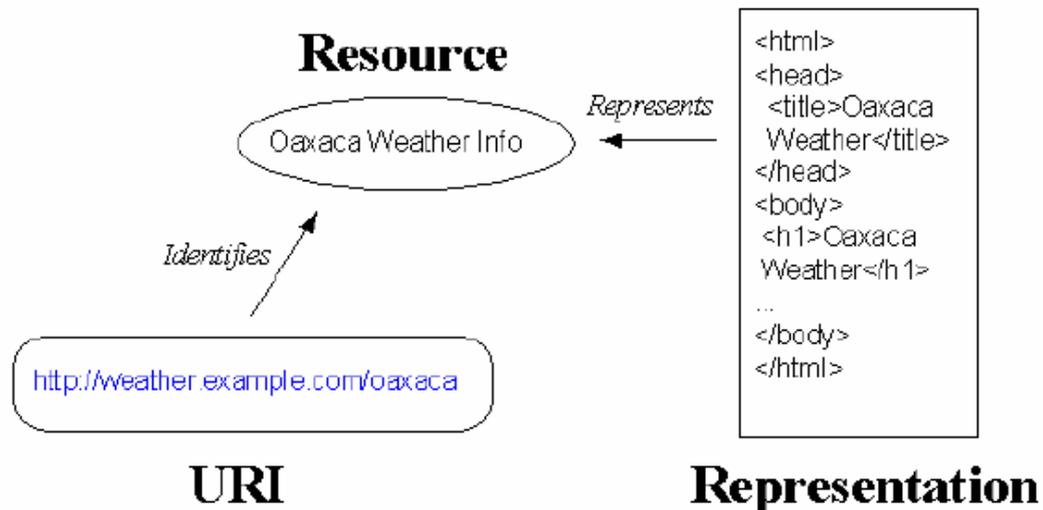


Figura 14 - Ruolo dell'URI nel WEB

2.12.1 Composizione delle URI

Le URI sono composte da:

- schema
- parte specifica dello schema

secondo la sintassi

`<schema>:<scheme-specific-part>`

Alcuni esempi di URI sono:

- schema ftp per File Transfer Protocol services
<ftp://ftp.is.co.za/rfc/rfc1808.txt>

- schema http per Hypertext Transfer Protocol services <http://www.math.uio.no/faq/compression-faq/part1.html>
- schema mailto per indirizzi di posta elettronica <mailto:mduerst@ifi.unizh.ch>

2.12.2 URI Schema

Le URI Schema definiscono la metodologia per individuare la risorsa in modo univoco. La <scheme-specific-part> viene interpretata in funzione di quale schema stiamo utilizzando. Il registro ufficiale dei nomi degli schema delle URI è mantenuto da IANA[IANA]. Esempi di schema registrati sono:

- ftp
- http
- file

Schemi di URI [URISCHEMES]

Nome schema	Descrizione	Riferimento
ftp	File Transfer Protocol	[RFC1738]
http	Hypertext Transfer Protocol	[RFC2616]
gopher	The Gopher Protocol	[RFC1738]
mailto	Electronic mail address	[RFC2368]
news	USENET news	[RFC1738]
nntp	USENET news using NNTP access	[RFC1738]
telnet	Reference to interactive sessions	[RFC1738]
wais	Wide Area Information Servers	[RFC1738]
file	Host-specific file names	[RFC1738]
prospero	prospero Directory Service	[RFC1738]
z39.50s	Z39.50 Session	[RFC2056]
z39.50r	Z39.50 Retrieval	[RFC2056]
cid	content identifier	[RFC2392]
mid	message identifier	[RFC2392]
vemmi	versatile multimedia interface	[RFC2122]
service	service location	[RFC2609]
imap	internet message access protocol	[RFC2192]
nfs	network file system protocol	[RFC2224]
acap	application configuration access protocol	[RFC2244]
rtsp	real time streaming protocol	[RFC2326]
tip	Transaction Internet Protocol	[RFC2371]
pop	Post Office Protocol v3	[RFC2384]
data	data	[RFC2397]
dav	dav	[RFC2518]
opaquelocktoken	opaquelocktoken	[RFC2518]
sip	session initiation protocol	[RFC2543]
tel	telephone	[RFC2806]
fax	fax	[RFC2806]
modem	modem	[RFC2806]

ldap	Lightweight Directory Access Protocol	[RFC2255]
https	Hypertext Transfer Protocol Secure	[RFC2818]
soap	soap.beep	[RFC3288]
soap.beeps	soap.beeps	[RFC3288]
xmlrpc.beep	xmlrpc.beep	[RFC3529]
xmlrpc.beeps	xmlrpc.beeps	[RFC3529]
urn	Uniform Resource Names	[RFC2141]
go	go	[RFC3368]
h323	H.323	[RFC3508]
ipp	Internet Printing Protocol	[RFC3510]
tftp	Trivial File Transfer Protocol	[RFC-lear-tftp-uri-06.txt]

2.12.3 URI HTTP

Le URI http sono così composte¹¹: `http_URI = "http:" "://" [authority "@"] host [":" port] [abs_path]`

- authority informazioni utili per l'autorizzazione
- host un nome a dominio o un indirizzo IP
- port un numero intero
- abs_path identificativo per raggiungere la risorsa sull'host indicato

Esempi

`http://www.ietf.org/rfc/rfc2396.txt`

`http://lucchesi:password@www.iit.cnr.it:80/index.php`

2.12.4 Struttura ad albero degli URI http

Le URI http hanno una struttura ad albero gerarchica delimitata da "/".

Esempio di struttura gerarchica

individua la risorsa tramite lo schema http scomponendo l'identificativo in:

`http://www.w3.org/TR/xhtml1/Overview.html`

- www.w3.org
- TR
- xhtml1
- Overview.html

¹¹ Ricordarsi che le parti fra parentesi quadre ([...]) non sono obbligatorie.

2.12.5 Elementi riservati, e valori predefiniti nelle URI

Alcuni caratteri nelle URI sono riservati poiché coprono particolari esigenze. I caratteri che seguono sono alcuni tra quelli riservati:

";" | "/" | "?" | ":" | "@" | "&" | "=" | "+" | "\$" | ","

Esempio:

"/" è utilizzato come separato negli schema delle URI con struttura gerarchica.

Le URI HTTP hanno dei valori predefiniti:

- se la porta non è esplicitamente indicata allora la porta viene impostata a 80
- se `abs_path` non è esplicitamente indicato allora l'`abs_path` viene impostato a /

2.12.6 URL

La URL è un concetto informale per identificare un tipo di URI.

Le URL sono URI che identificano una risorsa tramite la rappresentazione del metodo di accesso alla risorsa stessa (normalmente l'indirizzo di rete della risorsa).

Esempi di URI:

<http://www.ietf.org/rfc/rfc2396.txt> è una URL (quindi anche una URI !)

ISBN:0-201-59625-3 è una URI, ma **non** è una URL

Nota: Contemporary View

Over time, the importance of this additional level of hierarchy seemed to lessen; the view became that an individual scheme did not need to be cast into one of a discrete set of URI types, such as "URL", "URN", "URC", etc. Web-identifier schemes are, in general, URI schemes, as a given URI scheme may define subspaces. Thus "http:" is a URI scheme. "urn:" is also a URI scheme; it defines subspaces, called "namespaces". For example, the set of URNs, of the form "urn:isbn:n-nn-nnnnnn-n", is a URN namespace. ("isbn" is an URN namespace identifier. It is not a "URN scheme", nor is it a "URI scheme.")

Further, according to the contemporary view, the term "URL" does not refer to a formal partition of URI space; rather, URL is a useful but informal concept. A URL is a type of URI that identifies a resource via a representation of its primary access mechanism (e.g., its network "location"), rather than by some other attributes it may have. Thus, as we noted, "http:" is a URI scheme. An http URI is

a URL. The phrase "URL scheme" is now used infrequently, usually to refer to some subclass of URI schemes which exclude URNs.

Approfondimenti

- Naming and Addressing: URIs, URLs, ...
<http://www.w3.org/Addressing/>
- Uniform Resource Identifiers (URI): Generic Syntax, agosto 1998
<http://www.ietf.org/rfc/rfc2396.txt>
- Internationalized Resource Identifiers (IRIs), gennaio 2005
<http://www.ietf.org/rfc/rfc3987.txt>