Introduzione al WWW e agli URI

Fabio Vitali

12 novembre 1999

“[...] un libro è più di una struttura verbale, o di una serie di strutture verbali; è il dialogo che intavola col suo lettore e l'intonazione che impone alla sua voce e le mutevoli e durature immagini che lascia nella sua memoria. [...] La letteratura non è esauribile, per la sufficiente e semplice ragione che un solo libro non lo è. Il libro non è un ente privo di comunicazioni: è una relazione, è un asse di innumerevoli relazioni. Una letteratura differisce da un'altra, successiva o precedente, meno per il testo che per il modo in cui è letta: se mi fosse dato leggere una qualsiasi pagina attuale - questa, per esempio - come la leggeranno nell'anno duemila, saprei come sarà la letteratura nell'anno duemila.”

J. L. Borges, Nota su Bernard Shaw in Altre Inquisizioni***, 1960***

Introduzione

Oggi esaminiamo:

- Una brevissima storia degli ipertesti
- Una brevissima storia del WWW
- Alcune caratteristiche generali del WWW
- Gli Universal Resource Identifier (URI)

Gli ipertesti

L’esigenza di evidenziare visivamente riferimenti e connessioni implicite od esplicite tra testi esiste da sempre.

Commentari a libri importanti (Bibbia, Talmud, Corano fino alla Divina Commedia o agli autori classici), che esistono fin dal tardo impero romano, utilizzano ogni sorta di trucco grafico e visivo per realizzare effetti di collegamento.

Era solo inevitabile che la meccanizzazione prima, e l’informatizzazione dopo, cercasse meccanismi per migliorare quest’esigenza, ed automatizzarne la fruizione.

Breve storia degli ipertesti (1)

Vannevar Bush e il Memex

- Negli anni quaranta, Vannevar Bush (consulente scientifico alla presidenza degli stati uniti) ipotizzò e iniziò a progettare un sistema elettromeccanico basato su microfilm per la memorizzazione e interconnessione di tutte le carte, libri ed informazioni utili per la vita d'ufficio.
- Il sistema (chiamato Memex) era basato su un meccanismo di fotografia e microfilmazione automatica di fogli, sulla pozzibilità di punzonare secondo codici prestabiliti i microfilm, e su un meccanismo di ricerca rapida di microfilm sulla base di queste punzonature.
- Esso era ovviamente troppo lento per un'utilizzo pratico del sistema, e l'elettronica rese obsoleta la tecnologia meccanica. Tuttavia, erano presenti in nuce tutte le idee degli ipertesti.

Breve storia degli ipertesti (2)

Theodore Nelson e Xanadu

- Negli anni sessanta Ted Nelson, visionario e futuro autore di uno dei lbri più importanti per il successo dei pesonal computer, Computer Lib, iniziò a progettare e cercare fondi per realizzare 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 Memex), 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, la modifica e possedevano un sistema efficace di gestione monetaria per i documenti a pagamento.
- Il sistema Xanadu non esistette mai, anche se ci andò vicino ala fine degli anni 80. Su di esso si ispirò Tim Berners-Lee per il World Wide Web.

Breve storia degli ipertesti (3)

Douglas Engelbart e Augment

- Nella fine degli anni sessanta Doug Engelbart iniziò a lavorare sul concetto di personal computing, utilizzando costose workstation IBM per realizzare un sistema di video-conferenza, editing di testi gerarchici (outline processor) ed ipertestuali e di supporto per il lavoro cooperativo dotato di interfaccia a finestre, mouse e altri meccanismi rivoluzionari di input ed output.
- Esiste tuttora un video del 1967 dove si vedono Engelbart ed il suo team usare queste tecnologie che sarebbero diventate d'uso comune 15 anni dopo almeno!
- L'idea di Engelbart era che il supporto di caratteristiche innovative potesse aumentare il potenziale intellettivo degli uomini, e che l'evidente difficoltà tecnica fosse un ostacolo superabile con un training adeguato.

Storia del WWW (1)

- Nel 1989, un gruppo di ricercatori informatici del CERN (il centro di ricerca in fisica nucleare di Ginevra ricevettero l'incarico da parte della direzione di realizzare un meccanismo per la diffusione rapida di articoli, appunti e opinioni tra i fisici che ruotavano intorno al centro.
- Tim Berners-Lee, Robert Cailliau ed altri identificarono Internet, ipertesti e SGML come elementi chiave per questo meccanismo.
- Nel 1991, alla conferenza sugli ipertesti, Berners-Lee e Cailliau mostrarono (con poco successo) il primo prototipo della loro applicazione, realizzata in client-server su architettura NeXT: World-Wide Web.
- Il prototipo NeXT era composto di:
  - Un server che spediva documenti memorizzati localmente a chiunque lo richiedessesecondo il protocollo stabilito, e che memorizzava documenti spediti da remoto, il tutto senza autorizzazione o verifica
  - Un editor di testi parzialmente WYSIWYG che permetteva di visualizzare documenti ipertestuali e di modificarli, creando link e blocchi di testo.

Storia del WWW (2)

- I principali competitor su Internet erano:
  - FTP, che forniva un meccanismo di scambio di file, senza preoccuparsi della visualizzazione
  - WAIS, che era un server con notevoli potenzialità di ricerca su documenti di solo testo
  - Gopher, che forniva un meccanismo di organizzazione di documenti di testo in gerarchie distribuite su più server
- I principali competitor nel campo degli ipertesti erano:
  - Hypercard (e Toolbook): strumenti di produttività individuale, a metà strada tra il database, il programma di disegno, l'ambiente di fast prototyping e l'ipertesto. Non distribuito, era un'applicazione monolitica.
  - Microcosm: un semplice server su rete locale, con identificazione degli utenti, e una serie di moduli e modifiche (***hack***) su applicazioni comuni (AutoCAD, MS Word …) per permettere la ricerca ed visione dei link.
  - Hyper-G: un sistema client-server complesso e completo, con un protocollo di comunicazione sofisticato, un modello di dati completo, ma senza il concetto di indirizzamento completo.

Storia del WWW (3)

- Nonostante l'accoglienza fredda dei ricercatori sugli ipertesti, i fisici furono entusiasti del WWW: comodo, facile da usare e da imparare, gratuito, privo di alternative realistiche (Hyper-G è di un anno più tardi).
- Nell'ottobre del 1992 il National Centre for Supercomputing Applications (NCSA) esaminò il prototipo di WWW e decise di realizzarne una versione propria.
- L'NCSA é impegnata nella realizzazione o nella re-ingegnerizzazione di soluzioni informatiche che aiutino tutta la comunità scientifica nello svolgimento delle loro ricerche. L'NCSA produce software per la visualizzazione di dati scientifici, e ha realizzato lo standard Internet per la connessione a computer via rete: Telnet.
- Con la realizzazione del server NCSA e del primo browser WWW, chiamato Mosaic, l'NCSA decretò l'inizio del successo esplosivo del sistema.

Storia del WWW (4)

- Mosaic aveva perso la capacità di editing del prototipo NeXT, ma aveva acquistato la capacità di visualizzare semplici immagini GIF.
- In breve, milioni di utenti iniziarono ad usare Mosaic o altri browser WWW, e decine di migliaia installarono server HTTP. Nel frattempo, però, l'NCSA aveva deciso di non spendere più energie di tanto sul WWW.
- Marc Andreessen, realizzatore del prototipo di Mosaic su X-Windows, e capo del gruppo alla NCSA che aveva realizzato le altre versioni ed il server, si trovò la strada bloccata per l'evoluzione del progetto.
- Jim Clark, ex professore a Berkeley e co-fondatore della Silicon Graphics, standard nella grafica professionale (animazioni, effetti speciali, progettazione, ecc.) cercava una nuova idea da finanziare.

Storia del WWW (4)

- Prendendo Andreessen come condirettore, Clark fonda nel 1993 la Mosaic Corporation, poi rinominata Netscape Corporation per evitare problemi legali con la NCSA.
- Il successo di Netscape Navigator (1994) è immediato e totale. La ditta Netscape ha il più rapido passaggio da fondazione a quotazione in borsa della storia, ed una delle quotazioni iniziali di maggior successo
- Ma fin dall'inizio Clark e soci sanno di essere o di star per entrare nel mirino di Microsoft, e si preoccupano di mantenere competitività e controllo del mercato.
- Nel frattempo, Berners-Lee e Cailliau cercano di mantenere il controllo sull'evoluzione del World Wide Web e fondano il W3C, con fondi della ricerca e dell'università.

Storia del WWW (5)

- Microsoft, dopo una falsa partenza con Microsoft network, abbraccia definitivamente e con energia la tecnologia Internet, ed inizia a reaizzare un browser WWW (Internet Explorer) ed un server HTTP (Microsoft Information Server)
- La lotta è senza esclusione di colpi, inclusa il bundle di IE in Windows 95, l'integrazione del browser come interfaccia di Windows in Second Edition, la causa di concorrenza sleale all'antitrust, ecc.
- Nel marzo 98 Netscape chiede l'aiuto della comunità dei programmatori rilasciando il codice sorgente della versione 5 di Navigator. Tanti progetti nascono, ma le nuove versioni stabili ed utilizzabili latitano.
- A novembre del 98 Netscape viene comprata da America On Line, Clark e Andreessen lasciano (multimiliardari). A febbraio 99 lascia l'ingegnere principale di Mozilla.org.

Cos’è il WWW

Il World Wide Web è un sistema per la presentazione a schermo di documenti multimediali, e per l’utilizzo di link ipertestuali per la navigazione.

Il sistema è distribuito e scalato su tutta Internet, ed è basato su alcuni semplici concetti:

- Il client o browser è un visualizzatore di documenti ipertestuali e multimediali. Può visualizzare testi, immagini e semplici interfacce grafiche, ma non permette di editare documenti.
- Il server è un semplice meccanismo di accesso a risorse locali (file o record di database, ecc.) in grado di trasmettere via socket TCP documenti individuati da un identificatore univoco
- Il server può collegarsi ad applicazioni server-side (tramite protocollo CGI) ed agire da tramite tra il browser e l’applicazione facendo del browser l’interfaccia dell’applicazione.

I protocolli del WWW

Alla base di WWW ci sono i seguenti protocolli:

- Uno standard per identificare in maniera generale risorse di rete e per poterle specificare all'interno di documenti ipertestuali (chiamato URI).
- Un protocollo di comunicazione state-less e client-server per l'accesso a documenti ipertestuali via rete (chiamato HTTP).
- Un linguaggio per la realizzazione di documenti ipertestuali (chiamato HTML) basato su SGML e incentrato sulla possibilità di realizzare connessioni ipertestuali in linea nella descrizione strutturale del documento.

Evoluzioni del WWW (1)

- **Inclusione di oggetti**: Mosaic introdusse il supporto per immagini in-line, e Netscape introdusse poi i plug-in per inserire oggetti di tipi diversi nel documento del browser, ed infine le applet Java. IE ha generalizzato le possibilità offerte da COM, realizzando il protocollo proprietario ActiveX.
- **Scripting**: Netscape introdusse LiveScript, poi ribattezzato Javascript, per realizzare semplici applicazioni client-side con linguaggi di scripting appositi. IE rispose con Jscript e Vbscript.
- **Stili**: L'uso di trucchi per forzare HTML a rese grafiche insolite ha portato a creare linguaggi appositi per gestire gli aspetti di visualizzazione del documento. CSS (livelli 1 e 2) permette di controllare le caratteristiche dei documenti HTML. XSL si occupa di documenti XML.

Evoluzioni del WWW (2)

- **Gestione delle transazioni**: meccanismi per la gestione dello stato sono stati introdotti prima da Netscape, e poi standardizzati (cookies). Meccanismi di accesso in scrittura e cooperazione a risorse WWW vengono studiati in questo periodo (WebDAV).
- **Strutturazione dei documenti**: i limiti di HTML non erano soltanto nella visualizzazione, ma anche nella strutturazione. XML permette di definire linguaggi di markup più adatti ai singoli task.
- **Modello di link**: il modello di link di HTML è eccessivamente semplice. Xlink ed XPointer permettono di definire link sofisticati, sia per indirizzamento (blocchi di dimensione e locazione arbitraria), sia per funzionamento (inclusione in-line, memorizzazione esterna, multi-direzione, ecc.)
- **Modello di metainformazioni**: HTML permette di usare dei tag speciali ma limitati per fornire meta-informazioni sui documenti. RDF estende e generalizza questa possibilità.

URI

- Gli URI (Universal Resource Indentifier) sono una sintassi usata in WWW per definire i nomi e gli indirizzi di oggetti (risorse) su Internet.
- Questi oggetti sono considerati accessibili tramite l'utilizzo di protocolli esistenti, inventati appositamente, o ancora da inventare.
- Gli URI si indirizzano a risolvere il problema di creare un meccanismo ed una sintassi di accesso unificata alle risorse di dati disponibili via rete.
- Tutte le istruzioni d'accesso ai vari specifici oggetti disponibili secondo un dato protocollo sono codificate come una stringa di indirizzo

L’esigenza di identificatori (1)

Gli URI sono stati verosimilmente il fattore determinante per il successo del WWW.

Attraverso gli URI, il WWW è stato in grado di identificare risorse accessibili tramite il proprio protocollo, HTTP, e tramite tutti gli altri protocolli esisenti (FTP, Telnet, Gopher, WAIS, ecc.).

Il punto principale a cui gli altri sistemi non erano arrivati era una sintassi universale, indipendente dal protocollo e facilmente memorizzabile o scambiabile con cui identificare le risorse di rete.

L’esigenza di identificatori (2)

Il WWW utilizza gli identificatori in una varietà di modi:

- Link ipertestuali disponibili nel documenti HTML
- Immagini ed altri oggetti inclusi nel documento HTML (che è un formato solo testo)
- Connessioni e relazioni globali tra documenti (ad esempio, script e link possono essere messi esternamente al documento HTML e da sso riferiti globalmente.

In tutti questi casi lo stesso identificatore può essere usato dal protocollo di comunicazione, espresso nella sintassi HTML, o digitato direttamente dall’utente.

Criteri di design degli URI (1)

La sintassi degli URI é progettata per essere

- Estensibile: si possono aggiungere nuovi schemi, al fine di mantenere l'accessibilità delle risorse anche se nuovi protocolli vengono inventati
- Completa: tutti i nomi esistenti sono codificabili e nuovi protocolli sono comunque esprimibili tramite URI
- Stampabile: é possibile esprimere URI con caratteri ASCII a 7-bit, così da permettere scambi lungo qualunque canale, per quanto limitato o inefficiente, inclusi carta e penna.

Lo standard URI definisce alcune regole per la generazione di schemi di naming (insiemi di nomi caratterizzati dalla dipendenza da un protocollo di accesso comune), per la definizione dei caratteri accettabili e del carattere di escape.

Criteri di design degli URI (2)

Gli Universal Resource Identifier (URI) sono, per definizione, o degli Universal Resource Names (URN), o degli Universal Resource Locator (URL).

- Gli URL sono un indirizzo della risorsa che possa essere immediatamente utilizzato da un programma per accedere alla risorsa.
- Gli URL contengono tutte le informazioni necessarie per accedere all'informazione, ma sono fragili a modifiche non sostanziali del meccanismo di accesso (es. cambio del nome di una directory).
- Gli URN sono un nome stabile e definitivo di una risorsa, che possa fornire un informazione certa ed affidabile sulla sua esistenza ed accessibilità.
- Gli URN debbono essere trasformati da un apposito servizio, negli URL attualmente associati alla risorsa. Inoltre la mappa deve essere aggiornata ogni volta che la risorsa viene spostata.

La sintassi degli URI

Un URI è diviso in due parti:

- `uri = schema ":" parte-specifica`

Lo schema di naming (in pratica, il protocollo) é identificato da una stringa arbitraria (ma registrata) usata come prefisso. Il carattere di due punti separa il prefisso dal resto. La decodifica del resto dell'URI é funzione del prefisso.

Ogni schema ha una sua sintassi, ma esistono delle regole che tutti gli schemi debbono rispettare.

Caratteri riservati negli URI (1)

% Il carattere “%” é il codice di escape, e serve per l’utilizzo di caratteri particolari nell’URI, precedendone il codice esadecimale. Ad esempio, per utilizzare un carattere “%” nel’URI bisogna usare la stringa “%25”

/ Il carattere “/” é utilizzato unicamente per l’identificatzione di sottoparti di uno schema gerarchico, e non può essere usato per altri scopi.

. Il punto singolo “.” o il punto punto “..” hanno anch’essi un significato gerarchico riservato, per indicare ovviamente risorse allo stesso livello o al livello superiore.

Caratteri riservati negli URI (2)

- # Il carattere di hash “#” serve per delimitare l’URI di un oggetto da un identificatore di un frammento interno alla risorsa considerata. Questo permette ad un URI di far riferimento non soltanto ad una risorsa (oggetto di interesse del server), ma anche a frammenti interni alla risorsa (che verranno identificati dal client).
- ? Il punto interrogativo “?” serve per separare l’URI di un oggetto su cui é possibile fare una query (un database, per esempio), dalla stringa usata per specificare la query.
- + All’interno della query, il segno più “+” é usato al posto dello spazio (che non é mai usato per nessuna ragione).

Caratteri riservati negli URI (3)

* L’asterisco “*” ha un significato speciale all’interno di schemi specifici.

! Analogamente il punto esclamativo “!” ha un significato all’interno di uno schema.

%XX Caratteri speciali o riservati o in generale non sicuri (es. quelli superiori al codice ASCII 127) possono essere specificati tramite codifica esadecimale introdotta dal carattere di escape.

Caratteri riservati negli URI (4)

Esempio: i due URI

- `http://www.alpha.edu/a/b/c/d`
- `http://www.alpha.edu/a/b/c%2Fd`

non sono uguali, perché, benche il codice esadecimale corrisponda al carattere “/”, nel primo caso esso ha significato gerarchico, e nel secondo fa parte del nome dell’ultima sottoparte della gerarchia, “c/d”.

URN (1)

Gli URN non hanno ancora molto successo. Non esiste ancora nessun meccanismo di URN sufficientemente affermato.

Gli scopi degli URN sono:

- Ambito globale: non viene indicata una locazione, ed ha lo stesso significato da ovunque lo si usi
- Unicità globale: non è possibile assegnare lo stesso URN a risorse diverse
- Persistenza: Non esiste ragione per la sua cessata esistenza a parte la cancellazione della risorsa a cui fa riferimento.
- Scalabilità: ogni risorsa sulla rete deve poter possedere per lungo tempo un URN

URN (2)

- Estensibilità: nuove funzionalità emergeranno. E' necessario che lo schema di URN permetta estensioni per coprire le esigenze delle nuove funzionalità.
- Supporto per i meccanismi esistenti: esistono già dei meccanismi di naming globali: numeri ISBN per i libri, identificatori pubblici ISO per gli standard, codici UPC per i prodotti fisici. Lo schema di naming deve inglobare trasparentemente questi schemi di naming.
- Risoluzione: deve esistere un meccanismo semplice per la mappatura di un URN nell'URL più appropriato
- Indipendenza: ogni suddivisione gerarchica dell'autorità dei nomi deve essere autonoma (cioè gestisce in autonomia i nomi ad essa soggetti).

URL

Lo schema, in un URL, corrisponde al protocollo di accesso da utilizzare per accedere alla risorsa. La parte specifica dello schema dipende dal protocollo specifico.

Vediamo brevemente i seguenti schemi:

- HTTP e HTTPS
- FTP
- NNTP
- SMTP
- Telnet

HTTP e HTTPS

La sintassi della parte specifica è:

`http://`*`host`*`[:`*`port`*`]/`*`path`*`[#`*`fragment`*`][?`*`query`*`]`

`https://`*`host`*`[:`*`port`*`]/`*`path`*`[#`*`fragment`*`][?`*`query`*`]`

dove:

- **host** é l'indirizzo TCP-IP o DNS, dell'host su cui si trova la risorsa
- **port** é la porta a cui il server é in ascolto per le connessioni. In mancanza di specificazione, la porta é quella di default, 80 per HTTP e 443 per HTTPS.
- **path** é un pathname gerarchico (per esempio, un filename parziale) per l'identificazione della risorsa
- **fragment** é un identificativo di una sottoparte dell'oggetto. La definizione e il ritrovamento di queste sottoparti é a carico del client, e quindi la parte di fragment viene ignorata dal server, che restituisce l'intero oggetto.
- **query** é una frase che costituisce l'oggetto di una ricerca sulla risorsa specificata.

FTP

La sintassi della parte specifica è:

```
ftp://[user[:password]@]host[:port]/path [type]
```

dove:

- User e password sono utente e password per l'accesso ad un server FTP. La loro mancanza fa partire automaticamente una connessione anonima
- Host, port e path sono l'indirizzo del server, la porta di connessione ed il nome del file dell'oggetto ricercato, come per HTTP. La porta di default è 21.
- type regola i parametri di connessione FTP, come il tipo di trasferimento (ASCII o binario).

SMTP e Telnet

SMTP

La sintassi della parte specifica è:

```
mailto:user@host
```

dove

- non esiste il prefisso “//” perché lo schema non è gerarchico
- User e host sono i componenti dell’indirizzo di e-mail del destinatario

Telnet

La sintassi della parte specifica è:

```
telnet:host
```


NNTP

La sintassi della parte specifica è:

```
news:group
news:articleID@host
nntp:host/group/digit
```

dove

- l'accesso viene fatto usualmente al news server locale (specificato in varie preferenze).
- La specifica del solo gruppo restituisce l'elenco dei messaggi presenti nel gruppo.
- La specifica nella forma articleID@host permette di specificare l'articolo secondo l'identificativo interno locale al news server identificato.
- La terza sintassi, con specifica esplicita del protocollo nntp, viene usata scarsamente e solo per news server limitati privi di meccanismo di identificazione dei messaggi per articleID.

Conclusioni

Oggi abbiamo parlato di

- Storia degli ipertesti
- Evoluzione ed involuzione del WWW
- I meccanismi di naming stabili ed instabili
- La sintassi degli URI e degli URL

Riferimenti

Wilde's WWW, capitoli 1.1, 1.2, 1.3 e 2

Altri testi:

- D. Lowe, W. Hall, ***Hypermedia and the Web***, Wiley, 1999
- K. Sollins, L. Masinter, ***Functional Requirements for Uniform Resource Names***, RFC 2276, Jan. 1998
- T. Berners-Lee, L. Masinter, M. McCahill, ***Uniform Resource Locator***, RFC 1738, Dec. 1994
- R. Fielding, ***Relative Uniform Resource Locator***, RFC 1808, Jun 1995.