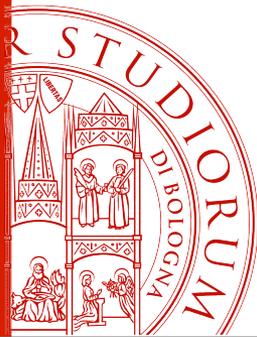


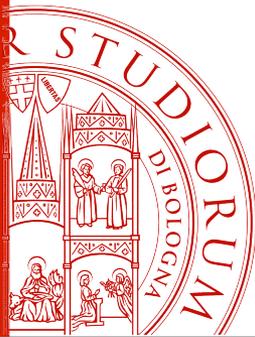
World Wide Web

Angelo Di Iorio
Università di Bologna



Cosa è il World Wide Web?

- Il World Wide Web è un sistema ipertestuale per la presentazione a schermo di documenti multimediali, e per l'utilizzo di link per la navigazione.
- Il sistema è distribuito e scalato su Internet
- Che relazione/differenza tra Internet e il World Wide Web?



Internet e WWW

R.it | Tecnologia

News App Social Network Mobile Videogiochi Sicurezza Prodotti Interattivi Video

Vodafone ADSL solo online a un prezzo speciale Scopri >

Consiglia 139 Tweet 33 +1 16 LinkedIn 6 indoona

Internet compie 25 anni. Tim Berners-Lee: "Il web deve essere un diritto di tutti"

Il www, nato come rete militare e poi come risorsa accademica, esisteva in forme via via più evolute fin dagli anni '60. Il 12 marzo 1989 Tim Berners-Lee depositò al Cern la proposta di creare l'infrastruttura digitale

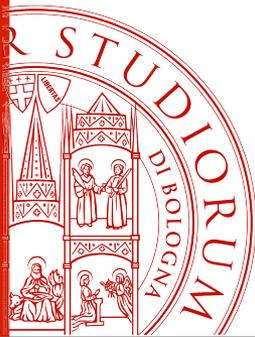
UN PROGETTO DI
la Repubblica
A&F
AFFARI & FINANZA

In collaborazione con:
CONAD

Approfondimenti e analisi:
nielsen

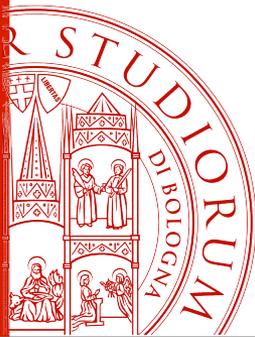
la Repubblica+
Leggi il tuo quotidiano

[Disclaimer: non è un attacco a Repubblica! 😊]



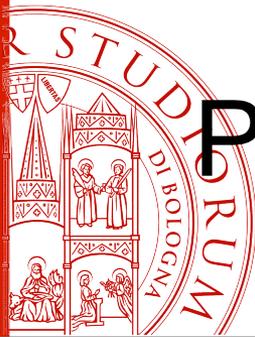
Internet e WWW

- Una rete di calcolatori è struttura di telecomunicazione in cui più calcolatori (in genere *eterogenei*, cioè diversi per hardware e sistema operativo) sono collegati tra loro, allo scopo di condividere risorse e scambiarsi informazioni
- Le reti possono essere collegate tra loro a formare reti di reti (*internets*)
- Chiamiamo **Internet** (con la "I" maiuscola) la rete planetaria di tutte le reti collegate tra loro e che comunicano con lo stesso protocollo
- **“World Wide Web”** non è sinonimo di “Internet”!



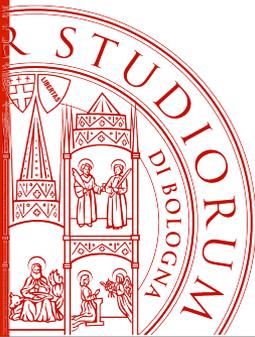
L'architettura del WWW

- L'organizzazione del WWW è basata su servizi di tipo **client-server**:
 - I **server** mettono a disposizione le informazioni, cioè i documenti ipertestuali, che sono residenti sul server stesso.
 - I **client (browser)** accedono ai documenti multimediali memorizzati sui server, attraverso un indirizzo che individua univocamente i documenti multimediali sulla rete.
- Il protocollo di comunicazione tra client e server è **HTTP**, HyperText Transfer Protocol
- Il linguaggio in cui sono scritti i documenti è **HTML**



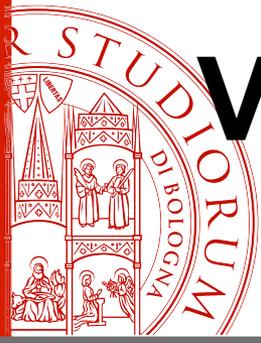
Protocolli, linguaggi e un po' di storia

- I protocolli e i linguaggi alla base del World Wide Web non sono cambiate per molti anni (e ancora oggi sono le stesse!) e la sua storia è abbastanza recente
 - I pionieri: Bush, Engelbart, Nelson (anni 50-60)
 - Gli inizi (1989-1994)
 - La prima guerra dei browser (1995-99)
 - Il predominio di Explorer (1999-2004)
 - La seconda guerra dei browser (2004-2009)
 - Lo strapotere di Google (2009 – oggi)
 - I dispositivi che cambiano e il Web of Data



Inizio non brillante...

- 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 (un precursore di Mac OsX): World-Wide Web.
- Nonostante l'accoglienza fredda gli informatici, i fisici furono entusiasti del WWW: comodo, facile da usare e da imparare, gratuito, privo di alternative realistiche.



Vague but exciting...

Vague but exciting...

CERN DD/OC
 Information Management: A Proposal
 Tim Berners-Lee, CERN/DD
 March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system.

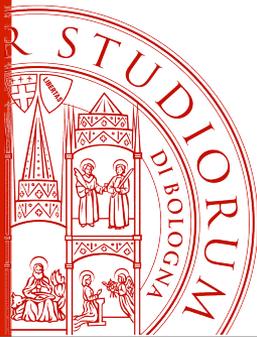
Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

The diagram illustrates a network of concepts and their relationships. At the center is 'A Proposal X'. 'This document' describes 'A Proposal X' and 'Hyper text'. 'Hyper text' includes 'Hypermedia'. 'Computer conferencing' unifies 'IBM GroupTalk', 'uucp News', and 'VAX/NOTES'. 'Hierarchical systems' includes 'uucp News' and 'C.E.R.N.'. 'C.E.R.N.' includes 'DD division', 'MIS', and 'OC group'. 'RA section' is also shown. 'This document' refers to 'Comms' and 'ACM'. 'Tim Berners-Lee' wrote 'This document'. 'Hyper Card' includes 'ENQUIRE'. 'Linked information' includes 'A Proposal X'. 'CERNDOC' describes 'A Proposal X' and 'C.E.R.N.'. 'C.E.R.N.' includes 'DD division', 'MIS', and 'OC group'. 'RA section' is also shown.



I protocolli fondamentali del WWW

- Alla base di WWW ci sono i seguenti protocolli e linguaggi:
 - 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 risorse ipertestuali via rete (chiamato **HTTP**).
 - Un linguaggio per la realizzazione di documenti ipertestuali (**HTML** e ora XHTML) basato su **SGML** (e ora XML) e incentrato sulla possibilità di realizzare connessioni ipertestuali in linea nella descrizione strutturale del documento.



Indirizzamento su WWW: URI e URL

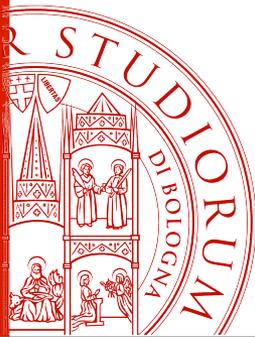
- Un punto fondamentale è capire come indirizzare e recuperare le risorse sul Web
- Ogni risorsa è univocamente individuata da una sequenza di caratteri che ne costituisce l'indirizzo, chiamata **URL (Uniform Resource Locator)** o **URI (Uniform Resource Identifier)**

<http://www.cs.unibo.it/docenti/index.html>

Protocollo

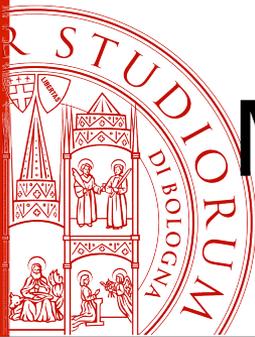
Nome simbolico di dominio

Nome locale del documento



Quando si usano gli URL?

- Gli URL possono essere forniti ad un browser direttamente dall'utente, inserendoli nella barra di navigazione
- Oppure possono essere scelti implicitamente cliccando su un link. Il “click” viene trasformato in una richiesta HTTP al dato URL
- Se un documento include altre risorse, infine, queste sono identificate (e recuperate) tramite i loro URL
 - immagini, video, etc.
 - programmi da eseguire sul browser (applets/javascript)



Ma come è scritta una pagina Web?

- In HTML (HyperText Markup Language), un linguaggio che “marca”:
 - struttura del documento
 - informazioni di presentazione
 - collegamenti ipertestuali (e URL della destinazione)
 - risorse multimediali (e URL)
- Il browser è in grado di interpretare queste informazioni e visualizzare la pagina finale



Il codice sorgente di una pagina HTML...

```
<HTML>
```

```
<HEAD>...</HEAD>
```

```
<BODY>
```

```
<H1> Questo &egrave; il titolo</H1>
```

```
<P>Questo documento descrive con un esempio l'uso di  
HTML. </P>
```

```
<H2> Essenza di HTML </H2>
```

```
<P>Un documento HTML consiste di testo "immerso" in  
comandi HTML (tecnicamente: tags). I comandi  
descrivono la struttura del documento. HTML  
permette:</P>
```

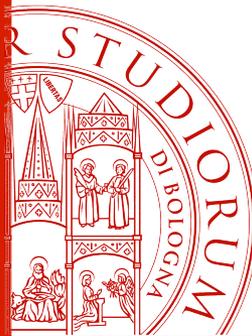
```
<UL>
```

```
<LI> di descrivere la struttura del documento;</LI>
```

```
<LI> inserire elementi non testuali, come questo  
stemma: <IMG src="./logoUnibo.gif"
```

```
ALT="logo dell'universita di Bologna">;</LI>
```

```
[...]
```



...e il risultato sul browser

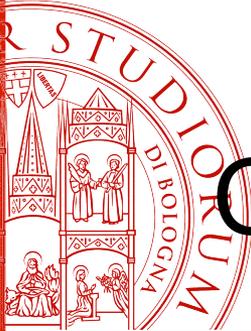
Questo è il titolo

Questo documento descrive con un esempio l'uso di HTML.

Essenza di HTML

Un documento HTML consiste di testo "immerso" in comandi HTML (tecnicamente: tags). I comandi descrivono la struttura del documento. HTML permette:

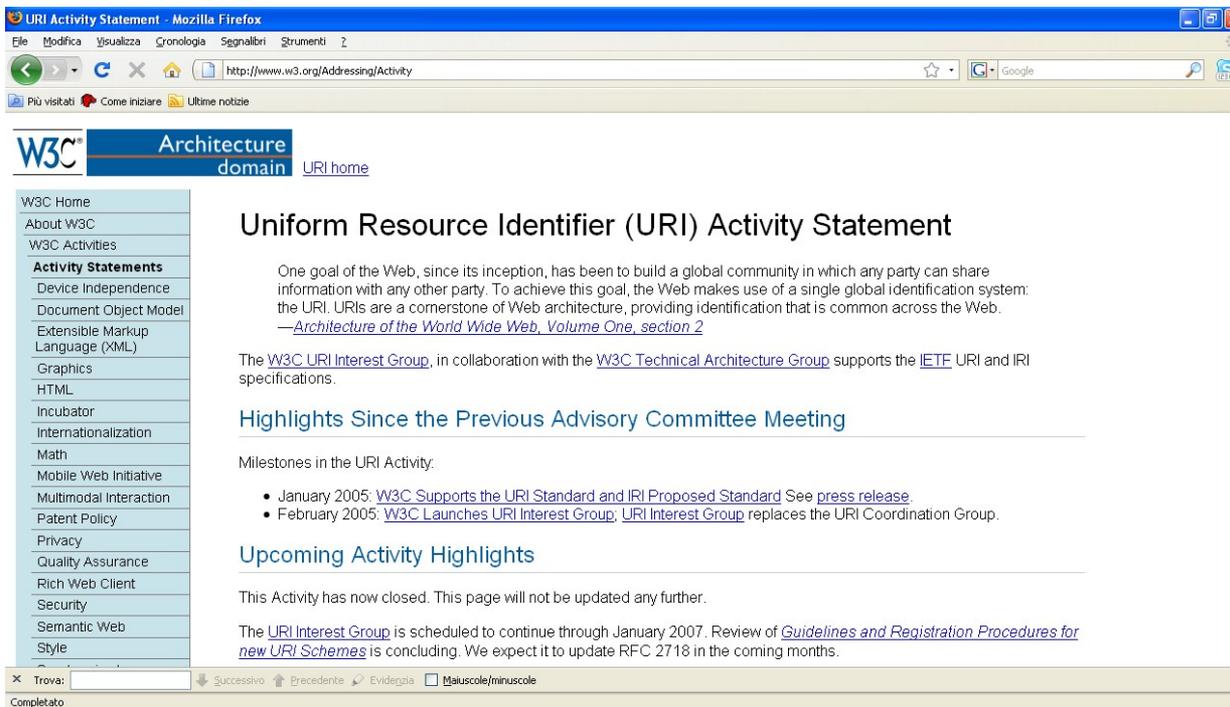
- di descrivere la struttura del documento;
- inserire elementi non testuali, come questo stemma:  ;
- inserire riferimenti ipertestuali: [Dipartimento di Scienze dell'Informazione](#)



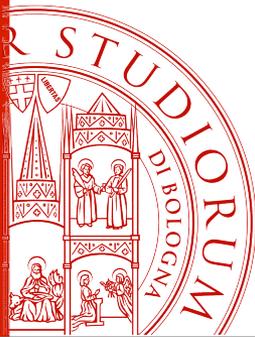
Caricamento di una pagina Web

- Il browser include quindi un client HTTP e un visualizzatore di pagine HTML

<http://www.w3.org/Addressing/Activity>

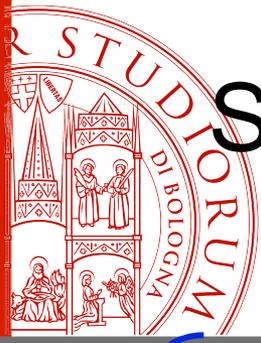


The screenshot shows a Mozilla Firefox browser window with the address bar containing <http://www.w3.org/Addressing/Activity>. The page content includes the W3C logo and navigation links, a sidebar menu, and the main heading "Uniform Resource Identifier (URI) Activity Statement". The main text explains the goal of the Web and mentions the W3C URI Interest Group. A "Highlights" section lists two items from January and February 2005. An "Upcoming Activity Highlights" section states that the activity has now closed and will not be updated further.



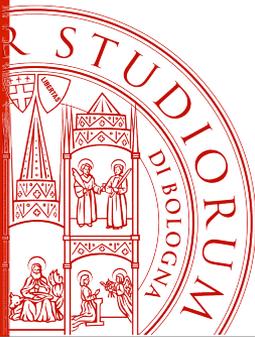
Le fasi del caricamento

1. Il browser determina la correttezza formale dello URL;
2. il browser sfrutta i livelli inferiori di TCP/IP (in particolare: il DNS per trasformare `www.w3.org` in un indirizzo IP; il protocollo TCP per inoltrare una richiesta al server che risiede su `www.w3.org`, ecc.);
3. Il browser inoltra al server di `www.w3.org` la richiesta del documento `/Addressing/Activity`;
4. il server invia il file richiesto e rilascia la comunicazione;
5. il browser interpreta il file ricevuto, rileva che la sua visualizzazione completa richiede la presenza di una figura, contraddistinta dallo URL `http://www.w3.org/Icons/WWW/w3c_home`
6. il browser inoltra la richiesta della figura al server indicato;
7. e così via per ogni altra componente del documento.



Struttura di una pagina Web (HTML)

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0//EN">  
<html>  
  <head>  
    <title>Titolo del documento</title>  
  </head>  
  <body>  
    <p class='primo'>Testo di un paragrafo</p>  
  </body>  
</html>
```

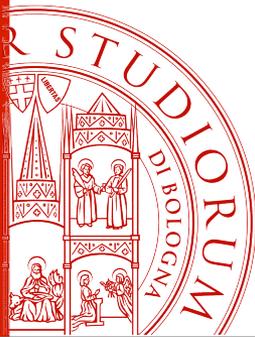


Gli elementi di HTML

Possiamo considerare gli elementi di HTML organizzati secondo alcune categorie:

- Tag della struttura complessiva (HTML, HEAD, BODY)
- Tag inline (B, I, SPAN, ecc.)
- Tag di blocco e liste (P, H1, H2, DIV, UL, OL, ecc.)
- Tag speciali (A, IMG, HR, BR)
- Tabelle (TABLE, TR, TD, TH)

Noi non guardiamo i dettagli degli elementi ma ci interessa avere un'idea generale del funzionamento di HTML



Elementi di blocco

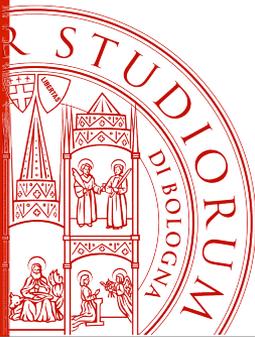
I tag di blocco definiscono l'esistenza di blocchi di testo che contengono elementi inline.

Elementi base:

- P (paragrafo),
- DIV (generico blocco),
- PRE (blocco preformattato),

Blocchi con ruolo strutturale

- H1, H2, H3, H4, H5, H6 (intestazione di blocco)



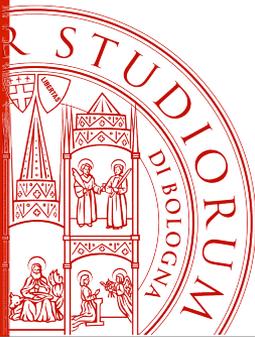
Elementi per liste

Le liste di elementi sono contenitori di elementi omogenei per tipo.

- UL: Lista a pallini di ; Attributo type (disc, square, circle)
- OL: lista a numeri o lettere di ; attributi start (valore iniziale) e type (1, a, A, i, I).

```
<UL>  
  <LI>Primo</LI>  
  <LI>Secondo</LI>  
  <LI>Terzo</LI>  
</UL>
```

```
<OL start="2">  
  <LI>Due</LI>  
  <LI>Tre</LI>  
  <LI>Quattro</LI>  
</OL>
```



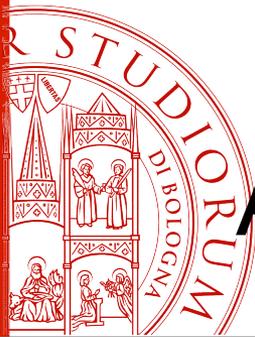
Link

I link sono definiti attraverso l'elemento A (anchor interna al documento). `<A>` è sintatticamente un elemento inline (non interrompe il flusso di testo)

Attributo principale:

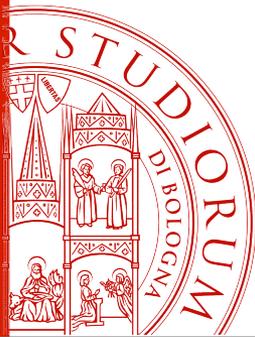
- HREF: specifica l'URI di una destinazione. Quindi ` . . . ` è un *estremo di partenza* di un link.

```
<p>Un link al <a href="B.html"> documento  
B</a></p>
```



Architetture del World Wide Web

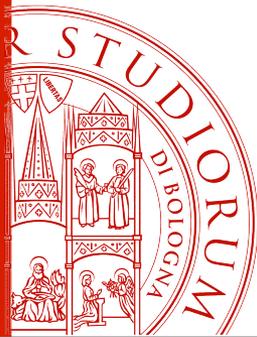
- Sito web statico
- Sito web dinamico: modello a tre livelli
 - Embedded code
 - Full application
- Sito web dinamico: modello a quattro livelli
- Rich client: applicazioni AJAX



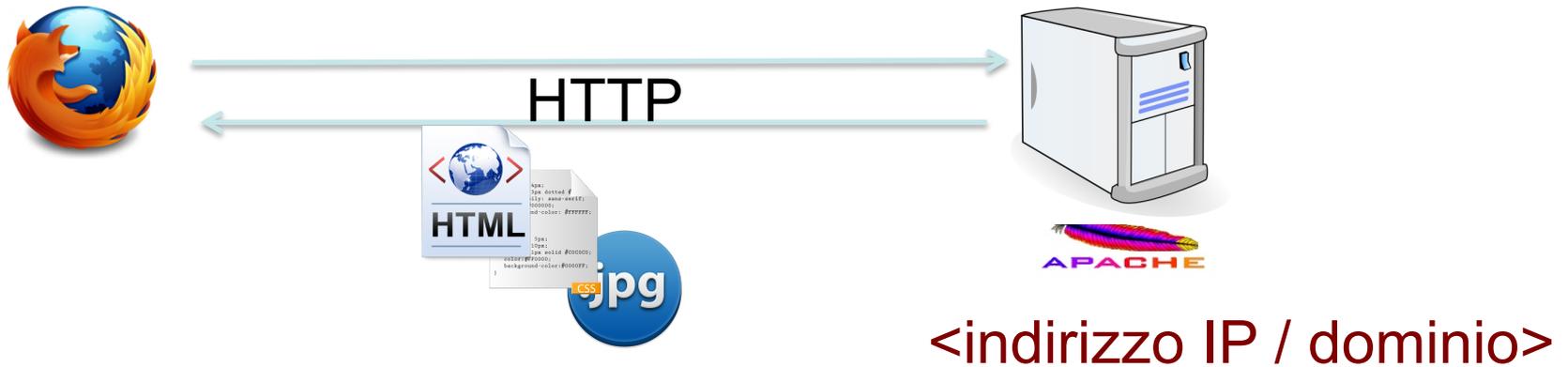
Sito web statico

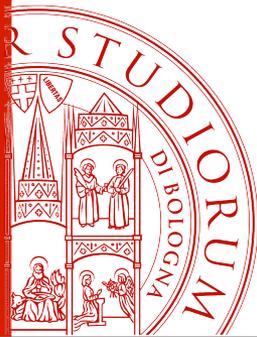
- Il server contiene una quantità di file fisici memorizzati in directory e di formati immediatamente riconoscibili dal browser (HTML, GIF, JPEG ecc.)
- C'è esattamente un file per ogni schermata possibile, ciascuno con un indirizzo (un URL) diverso.
- Il client richiede questi file ad uno ad uno e li riceve per la visualizzazione.
- Nessun contenuto visualizzato cambia rispetto al documento memorizzato su disco

- **Pregi:** facile da realizzare
- **Difetti:** totale mancanza di automazione e integrazione. Ogni file è indipendente dagli altri.



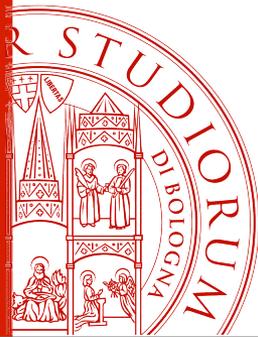
Sito web statico



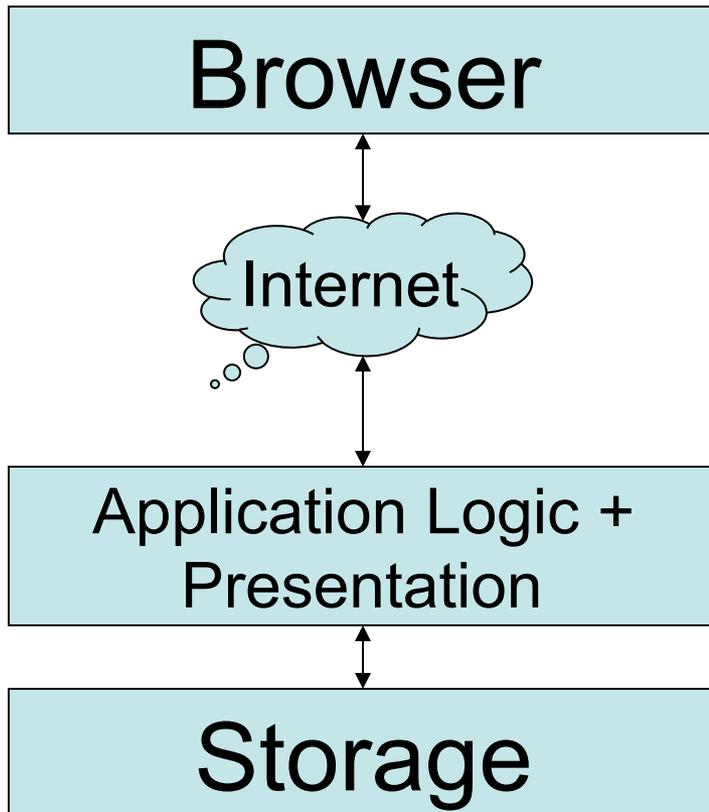


Sito web dinamico

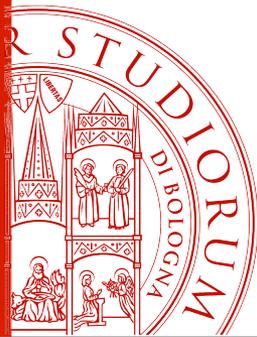




Siti dinamici: modello a tre livelli

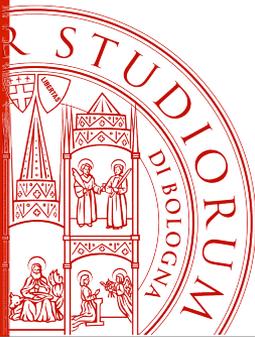


- Parte dei contenuti è statica e memorizzata su file, ma la parte importante è generata in output da un'applicazione server-side.
- Questa spesso raccoglie dati da query su un database, le elabora e le trasforma, e le spedisce come risposta al browser.
- Tipicamente associata a modelli applicativi di tipo LAMP (Linux, Apache, MySQL, PHP)



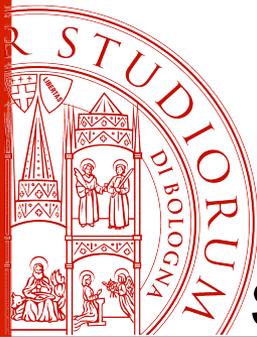
Rich client: Ajax e il web 2.0

- La presenza dell'application logic server-side comporta diversi problemi:
 - ogni passo dell'applicazione richiede di consultare il server, eseguire una funzione dell'application logic, generare l'HTML finale, riceverlo e visualizzarlo. Questo può portare via molto tempo.
 - Ogni passo dell'applicazione ha un proprio URL, che richiede specifiche politiche di naming e di caching piuttosto complesse da gestire
 - Questi problemi sono particolarmente rilevanti in un contesto in cui la quantità di dati a disposizione è in continuo aumento (Big Data, Internet of Things, Mobile devices)



Rich client: Ajax e il web 2.0

- All'inizio dell'esecuzione del servizio, il browser carica una pagina HTML, che contiene codice Javascript e alcune librerie Ajax che spostano l'elaborazione sul client.
- Scopo delle librerie Ajax è duplice:
 - Fornire librerie utili per la realizzazione di applicazioni client side in maniera facile (ad esempio, l'interazione con il server o la generazione di frammenti di output)
 - Velocizzare il caricamento dei dati (ad esempio, pre-caricarne una parte)

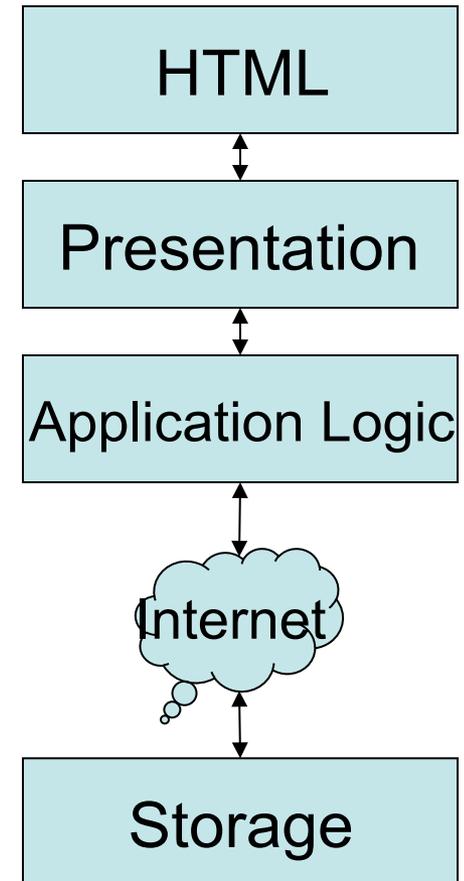
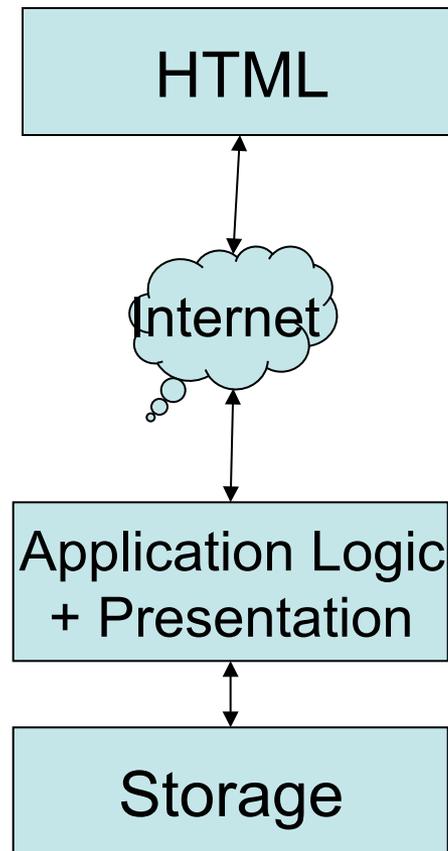
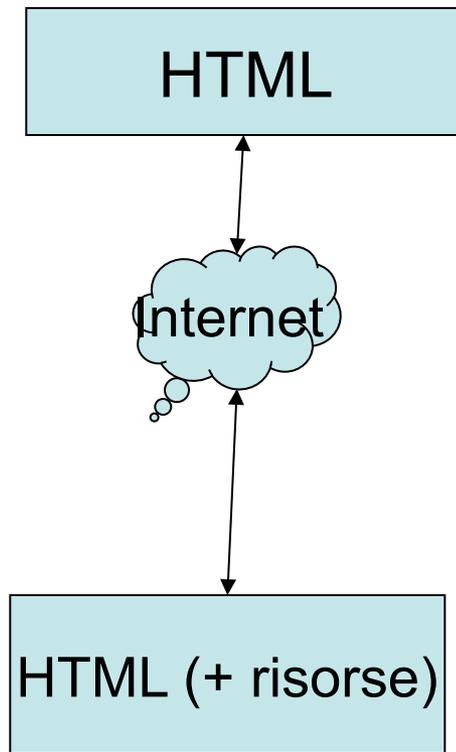


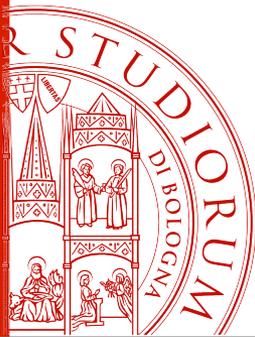
Confronto

Statico

Dinamico (3 livelli)

Ajax

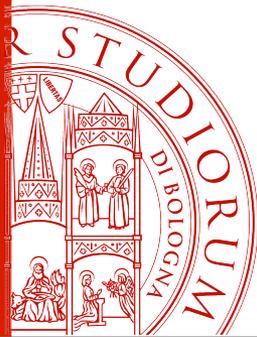




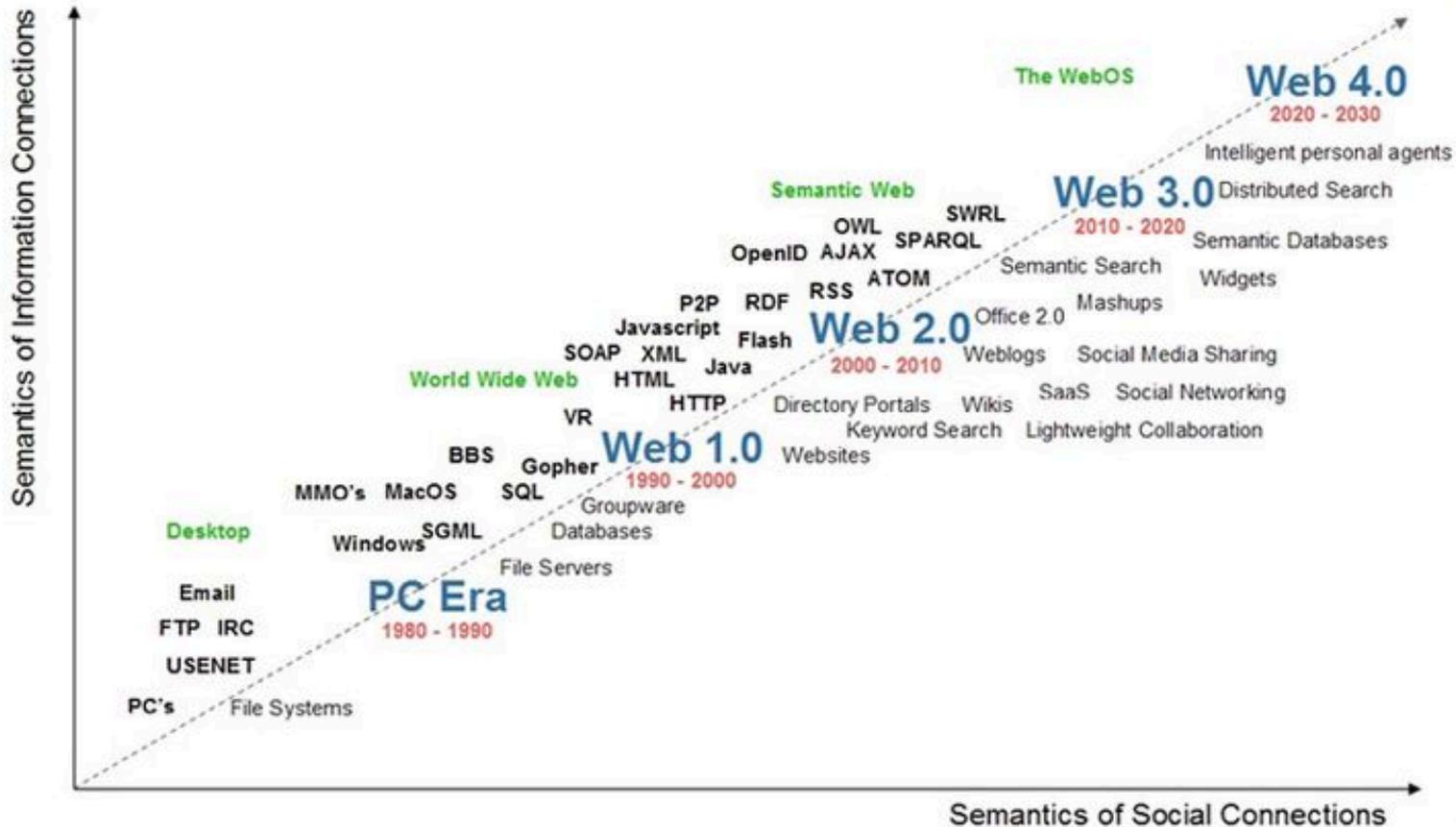
Web 2.0

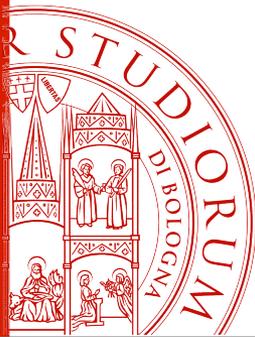
La *buzzword* attualmente più usata. Ha un significato tecnico e uno sociale/economico/marketing

- *Significato tecnico*: l'insieme delle tecnologie, che generano **siti altamente interattivi e dinamici**, in grado di fornire più servizi della semplice erogazione di contenuti
- *Significato sociale*: l'insieme degli strumenti di **partecipazione attiva** alla costruzione di contenuti per il web: blog, wiki, bookmark sociali, podcast, RSS, folksonomie, ecc.



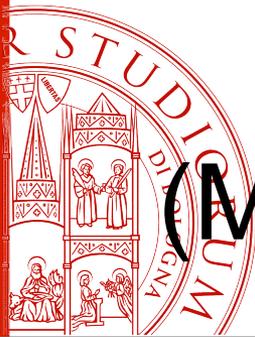
E dopo il Web 2.0?





Human-readable Web

- I contenuti Web sono stati per lungo tempo pensato per essere lettere da essere umani
- Gli agenti software possono recuperare, aggregare, visualizzare questi contenuti ma la maggior parte di questi è stata “consumata” da persone
- Ci sono alcune informazioni machine-readable (ad esempio nell’elemento META di HTML) ma sono molto limitate:
 - si riferiscono alla pagina corrente
 - descrivono la pagina e non le entità a cui si riferisce
 - non escono ad esprimere metadati complessi e **non forniscono una rappresentazione completa della conoscenza**



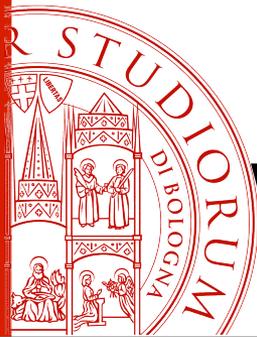
(Machine-readable) Web of Data

- “The Semantic Web will **bring structure to the meaningful content of Web pages**, creating an environment where **software agents** roaming from page to page can readily carry out sophisticated **tasks for users.**”

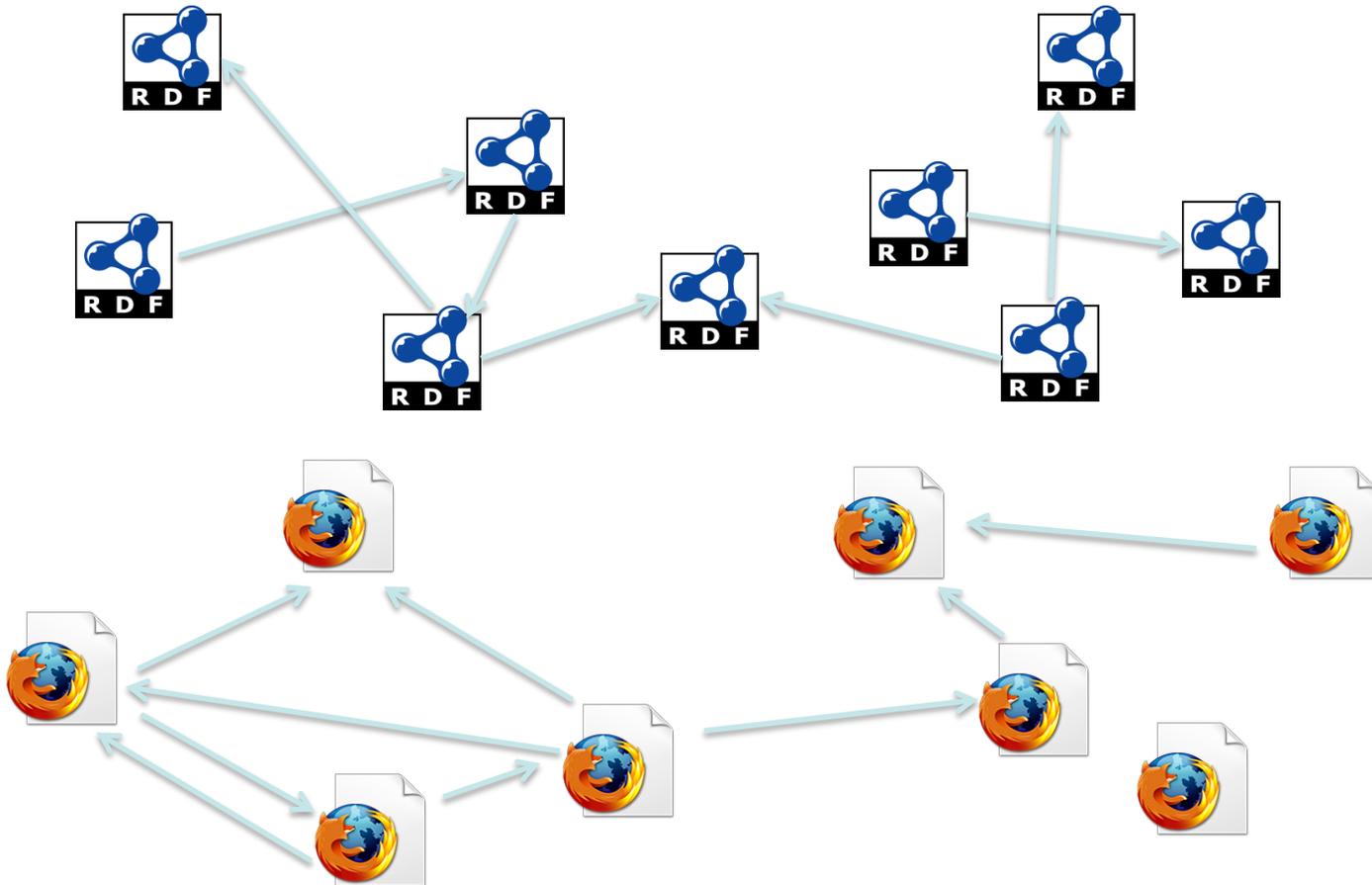
Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web", Scientific American, 2001

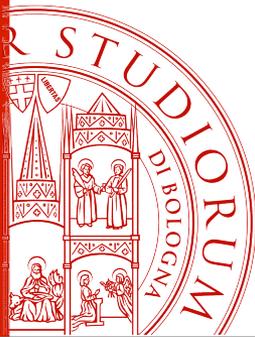
- “A **web of data** that can be processed directly and indirectly by **machines**”

Tim Berners-Lee



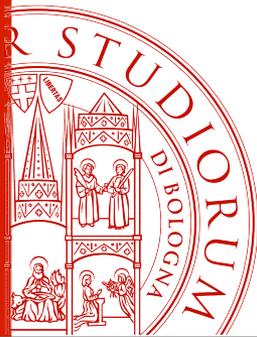
Web of Documents and Data





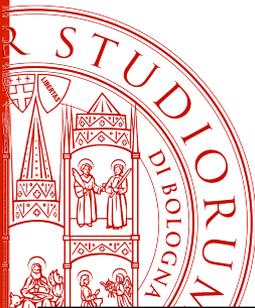
Web 1.0 vs. Semantic Web

	Web	Semantic Web
Oggetti	Contenuto semanticamente non strutturato	Affermazioni e statement
Agenti	Persone	Agenti software
Link	Link ipertestuali tra parti di documenti	Link tipati tra risorse
Vocabolario	Linguaggi di markup per strutturare documenti e layout	Linguaggi per esprimere semantica e logica



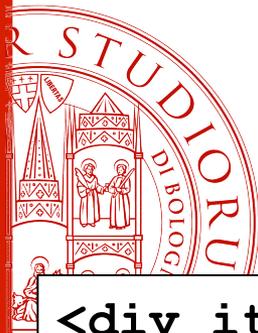
Un esempio (semplice): microdata

- Una soluzione per **iniettare informazioni semantiche** nelle pagine Web
 - Alternativa (e più semplice di) RDFa, il linguaggio chiave del Semantic Web
- Definisce un set di attributi di HTML5 progettati per esprimere proprietà di **item**
- **Fa ormai parte delle specifiche di HTML**
- Esistono diversi vocabolari, adottati dalla comunità, per descrivere *Persone*, *Eventi*, *Organizzazioni*, *Prodotti*, etc.
 - sfruttati dai motori di ricerca per migliorare accuratezza e qualità dei risultati



Esempio

```
<section itemscope itemtype="http://schema.org/Person">
  <p>Hello, my name is
    <span itemprop="name">Angelo Di Iorio</span>, I am
  a <span itemprop="jobTitle">researcher</span> at
  the <span itemprop="affiliation">University of
    Bologna</span>.
  </p>
  <section itemprop="address" itemscope
    itemtype="http://schema.org/PostalAddress">
    My office is in:
    <span itemprop="streetAddress">Via Ranzani 14</span>
    <span itemprop="addressLocality">Bologna</span>,
    <span itemprop="addressCountry">Italy
  </span>.
  </section>
</section>
```



Greys Anatomy in microdata

```
<div itemscope="itemscope"
      itemtype="http://schema.org/TVSeries">
  <span itemprop="name">Greys Anatomy</span> is a medical
  drama television series created by
  <div itemscope="itemscope"
        itemtype="http://schema.org/Person">
    <span itemprop="author">Shonda Rimes</
  span>
  </div>
  Starring:
  <div itemprop="actor" itemscope="itemscope"
        itemtype="http://schema.org/Person">
    <span itemprop="name">Justin Chambers</span>
  </div> ...
  <div itemprop="season" itemscope="itemscope"
        itemtype="http://schema.org/TVSeason">
    <span itemprop="name">Season 1</span> -
    <meta itemprop="numberOfEpisodes" content="14">
```