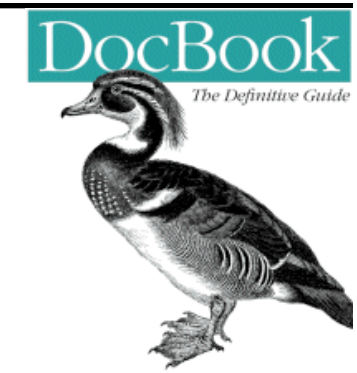
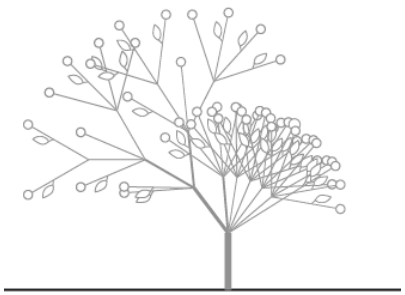


Scrivere in XML-Docbook

Breve corso introduttivo

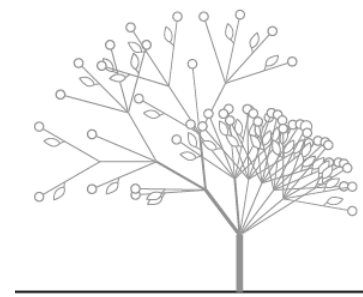
Francesca Di Donato
didonato(at)sp.unipi.it
http://www.sp.unipi.it/index.php?page=/hp/didonato



http://bfp.sp.unipi.it/~didonato/ovre/xmldocbook/

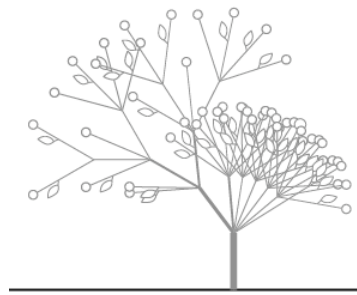
Di che cosa parleremo

- Che cosa sono i linguaggi e i formati di codifica?
- Web semantico e XML
- Introduzione a XML
- XML Docbook



Materiali didattici

- Le slides (che sono la "scaletta" di ciò di cui parliamo)
- Bibliolinkografia divisa per argomenti



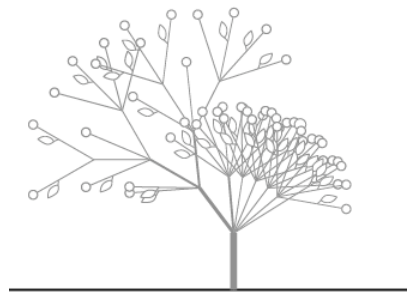
http://bfp.sp.unipi.it/~didonato/ovre/xmldocbook/

1. LINGUAGGI E FORMATI DI CODIFICA

Definizioni

- La codifica: "In informatica, rappresentare dati secondo un **sistema simbolico**, che consenta di **esprimerli in forma convenzionale**". Definizione del Dizionario della lingua italiana di Devoto e Oli.

- Linguaggi e formati



Codifica

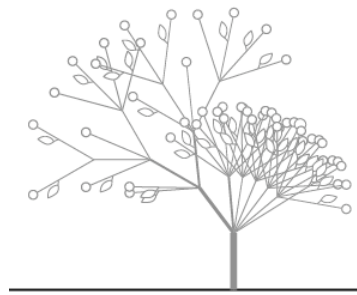
"In informatica, rappresentare dati secondo un sistema simbolico, che consenta di esprimerli in forma convenzionale".

Sistema simbolico: **linguaggio**, che ha una grammatica e una sintassi.

Forma convenzionale: schema strutturato di dati (**formato**).

N.B.

- Il modello di codifica deve essere isomorfo
- Il modello di codifica presuppone la scelta delle caratteristiche da codificare



Linguaggi di mark up

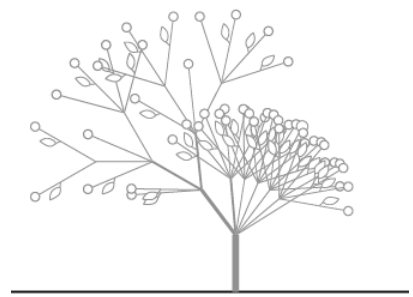
Markup: termine inglese per "caratterizzazione editoriale" (esplicita la formattazione dei documenti)

Tagging: termine inglese per "annotazione editoriale".

➡ Il Mark up nei sistemi **WYSIWYG** (What You See Is What You Get):

- Formattazione incorporata
- Codifica invisibile all'utente
- Documenti difficilmente gestibili

➡ I linguaggi di **mark up dichiarativi**: SGML

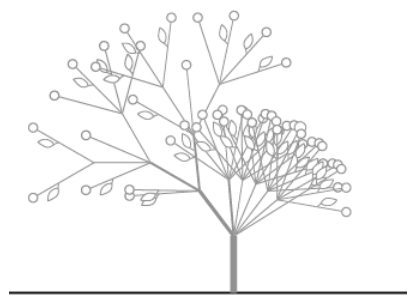


SGML

- SGML (Standard Generalized Markup Language) è il padre degli attuali linguaggi di Mark up
- I "dialetti" SGML:

➡ HTML (Hyper Text Mark up Language), inventato da Tim Berners-Lee nel 1989.

➡ XML (eXtensible Mark up Language), raccomandazione del W3C dal 1999.



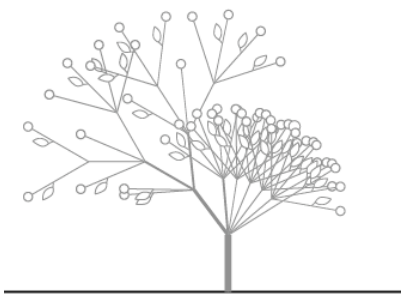
Interoperabilità e portabilità

Interoperabilità: la capacità di sistemi elettronici, informatici e telematici di scambiare i dati con altri sistemi, utilizzando formati e protocolli comuni.

Portabilità: la capacità di un software di essere adattata al fine di funzionare su un sistema diverso da quello in cui/per cui è stato scritto.

I **problemi** dei documenti digitali:

- Legati alla disponibilità di dispositivi hardware e software
- Elevata obsolescenza
- Difficile portabilità su piattaforme diverse
- Difficile condivisione dei dati e dei risultati



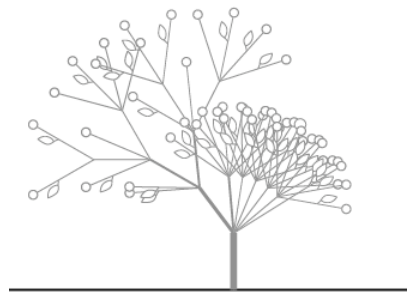
Standard e portabilità

Standard: formali **e** di fatto

Le **risposte:** standard portabili

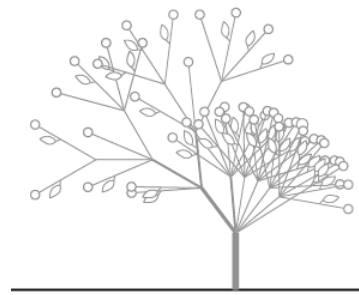
- Indipendenza dall'hardware
- Indipendenza dal software
- Indipendenza dal sistema di codifica dei caratteri
- Indipendenza logica dalle tipologie di elaborazione

ESEMPIO: Standard di **character encoding**:
ASCII (ISO 646), Latin-1 (Latin-1), Unicode (ISO 10646)



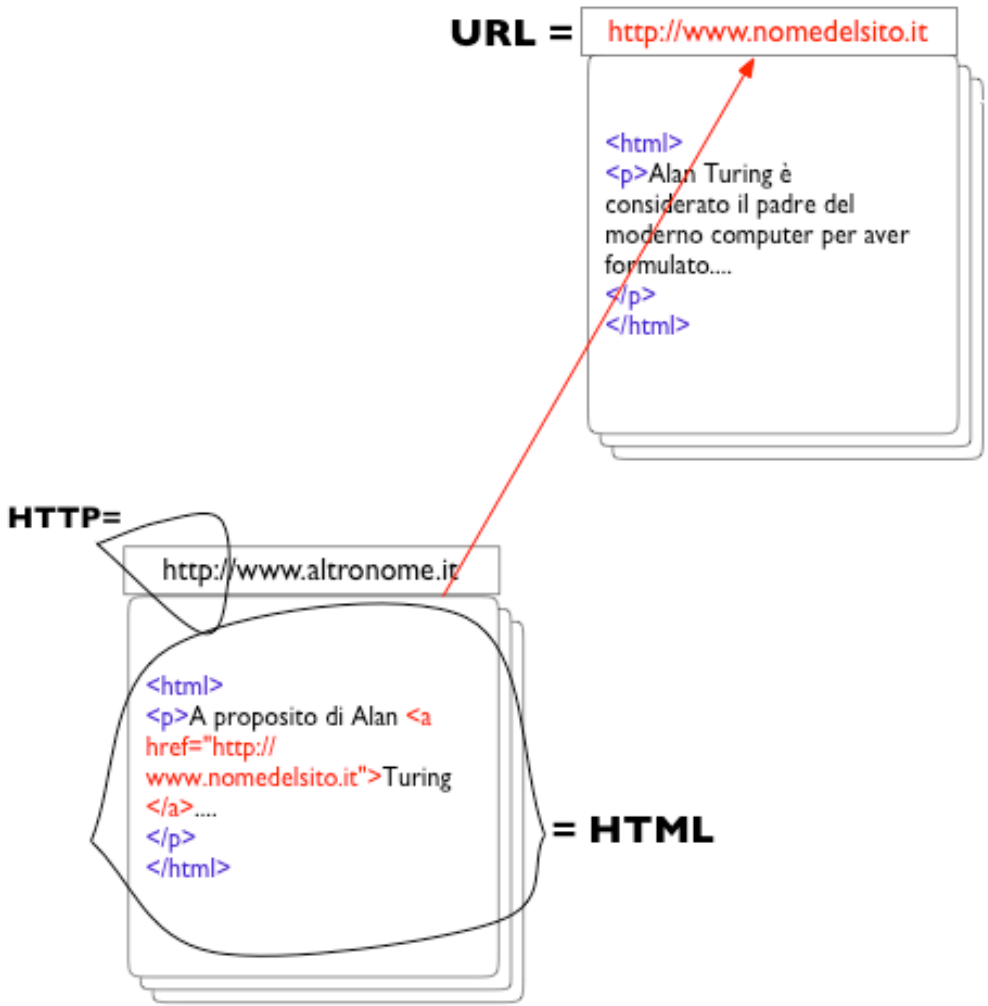
2. WEB SEMANTICO E XML

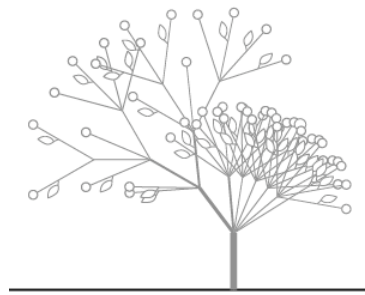
- World Wide Web
- HTML
- Semantic Web: introduzione



Premesse:

l'architettura del Web vista dal browser





Premesse:

HTML, il formato di mark up del Web

HyperText Markup Language

E' un dialetto di SGML creato da Tim Berners-Lee per la pubblicazione di pagine Web dotate di **link ipertestuali**; a differenza di SGML,

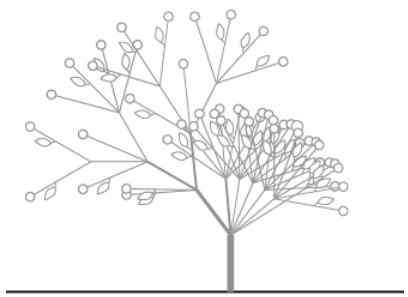
HTML

non è personalizzabile (gli elementi sono contenuti in tag di apertura e di chiusura, come nell'esempio a destra i caratteri in blu).

E' un linguaggio *interpretato*, *standard* e *portabile*.

La sua facilità di utilizzo ha senz'altro favorito il diffondersi del Web.

```
<html>
<head>
<title>Titolo del documento, che si vede
nel browser</title>
</head>
<body>

<h1>Titolo del documento</h1>

<p>Testo contenuto in un paragrafo che qui

<br />

va a capo</p>

</body>
</html>
```



Premesse:
I problemi di HTML

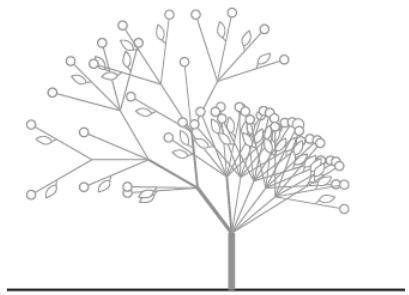
Un documento HTML è un file di testo che contiene:

Testo + Markup

I **prqblemi** di HTML

- Confonde l'**aspetto** di un documento e la **semantica** del testo
- Comprende la possibilità di **overlapping**

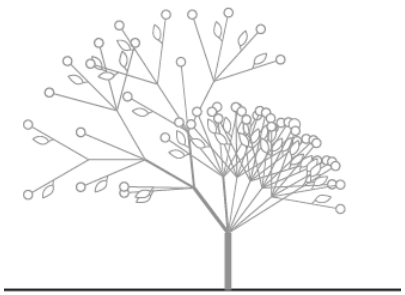
HTML **non permette** lo **scambio di informazione strutturata tra le macchine**



Il Web semantico

Storia: Orientamento adottato da poco più di cinque anni dal W3C.

Filosofia: Lo spazio dell'informazione del Web è stato progettato con l'obiettivo che fosse utile non solo per la comunicazione tra umani, ma perché vi partecipassero anche le macchine. Il "Web Semantico", al fine di rendere i dati comprensibili da appositi software, **accentua la separazione tra "data"** (utilizzabili ed elaborabili dalle macchine) **e "documents"** (leggibili dagli umani); per farlo, il W3C ha dato vita a **nuovi formati di codifica**, tali da poter esprimere le informazioni in forma processabile dalla macchina.



Il Web semantico:

Metadati

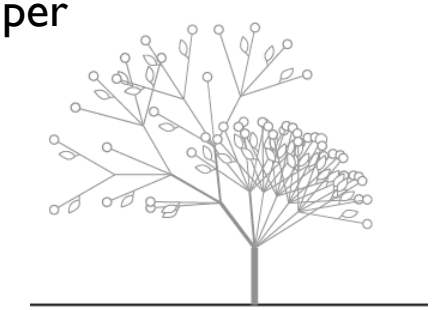
I metadati, "dati suoi dati", sono informazioni che arricchiscono le pagine Web, così che software appositamente creati possano farne uso. Nuove informazioni, strutturate tramite metadati, esprimono il contesto dei dati stessi – sono, cioè, informazioni aggiuntive sull'informazione. Sono metadati, ad esempio, l'autore e il titolo di un articolo.

Il Web sarà semantico soltanto quando diventerà un sistema:

globalmente inclusivo: **se chiunque potrà creare metadati su qualunque pagina.**

collaborativo: **i metadati devono essere condivisi da tutti gli utenti, e non utilizzati soltanto dai grandi connettori della rete (come Google).**

interoperabile: **dovranno essere definiti e usati standard per l'interscambio dei metadati.**

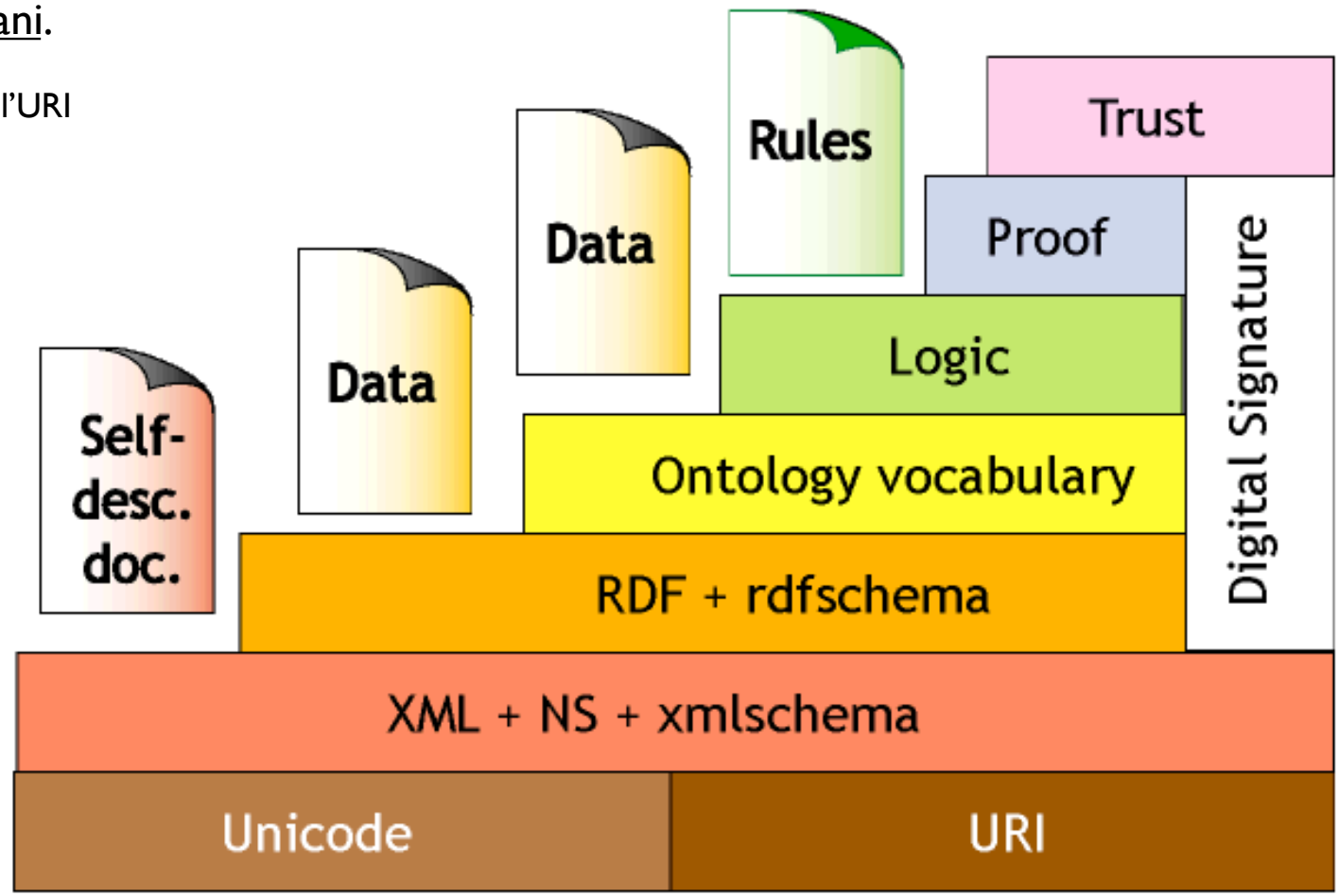


Il Web semantico:
Architettura

Tim Berners-Lee ha disegnato l'architettura del web semantico come una piramide di sette piani.

Alla sua **base (marrone)** si trovano l'URI (già a fondamento del WWW) e l'internazionalizzazione, cioè le problematiche legate all'encoding dei caratteri (l'ultimo standard di codifica adottato è Unicode, che comprende più di 65.000 caratteri).

Al **piano rosso** XML e il suo schema. Lo schema è la grammatica di riferimento del linguaggio, che ne definisce gli elementi e il content model.



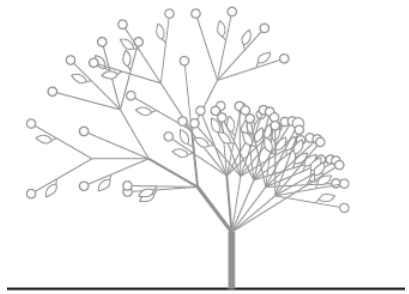
Che cos'è XML?

L'e**X**tensible **M**arkup **L**anguage è un linguaggio di marcatura (mark-up) che permette di strutturare semanticamente un documento senza definire come debba apparire. XML è una semplificazione di SGML e sta sostituendo HTML come formato per la ***produzione*** di pagine destinate al web.

Scrivere in XML presenta diversi **vantaggi:**

- Separa la grafica dalla struttura, e permette di pubblicare un documento in diversi formati (in particolare in pdf e in html) e con diversi layout grafici, a partire da un unico file.

- Permette di strutturare un documento e le sue sottoparti; nella pratica, ciò significa che diviene possibile individuare e linkare (citare) singole parti di documenti, e far crescere il fattore di impatto delle pubblicazioni.



Le applicazioni di XML

- E' nato come dialetto SGML per il Web
- E' il linguaggio dei data base
- E' il linguaggio dell'informatica umanistica (nata negli anni Settanta).

Il nostro uso:

➡ SCRIVERE i testi che produciamo IN XML

(Dobbiamo SCEGLIERE un buon **editor XML**)