Basi di Dati e XML: una prospettiva accademica

Valeria De Antonellis, Università di Brescia

Milano, 8 Novembre 2006

Sommario

- **Il presente**: XML - origini e diffusione
- **Il passato che non è mai passato**: Basi di Dati - risultati e nuove esigenze
- **Il passato che incontra il presente**: Basi di Dati e XML

Sommario

- **Il presente**: XML - origini e diffusione
- **Il passato che non è mai passato**: Basi di Dati - risultati e nuove esigenze
- **Il passato che incontra il presente**: Basi di Dati e XML

- Origini:
  - linguaggio di *mark-up* (W3C 1998)
- Motivazione per l'esplosione:
  - base per il "Web del futuro"
- Il vero successo:
  - sintassi per formati di ogni tipo

Linguaggi di mark-up

- Originariamente, annotazioni manuali su manoscritti o dattiloscritti, rivolte ai compositori tipografici
- Con l'automazione, veri e propri linguaggi che descrivono la formattazione di un documento
  - Ad esempio, molto diffuso nel mondo scientifico, LaTeX
- Con l'avvento del Web, tutti abbiamo conosciuto un linguaggio di mark-up: HTML

Originariamente

1.2 Definizione del modello relazionale

Il modello relazionale dei dati si basa sul concetto di relazione come definita nella teoria degli insiemi.

Si dice che r è una relazione sugli insiemi di valori $D_1, D_2, \ldots, D_n$, non necessariamente distinti, se è un sottoinsieme del prodotto cartesiano $D_1 \times D_2 \ldots \times D_n$. In simboli: $r \subseteq D_1 \times D_2 \ldots \times D_n$. $D_1, D_2, \ldots, D_n$ sono detti domini della relazione; n è detto grado della relazione.

lizzazione fisica. In particolare, partendo dall'obiettivo della indipendenza dei dati, Codd individua nella *tabella* una struttura di dati sufficientemente semplice e omogenea per garantire l'indipendenza; tale struttura ha quale suo fondamento matematico la relazione n-aria.

1.2 Definizione del modello relazionale

Il modello relazionale si basa sul concetto di relazione come definita nella teoria degli insiemi.

Si dice che r è una *relazione* sugli insiemi di valori $D_1, D_2, \ldots, D_n$, non necessariamente distinti, se è un sottoinsieme del prodotto cartesiano $D_1 \times D_2 \ldots \times D_n$. In simboli: $r \subseteq D_1 \times D_2 \ldots \times D_n$. $D_1, D_2, \ldots, D_n$ sono detti *domini* della relazione; n è detto *grado* della relazione.

Una relazione r è, quindi, un insieme di *ennuple* ordinate di valori $\langle d_1, d_2, \ldots, d_n \rangle$ tali che ogni valore d_i appartiene al dominio D_i, per $i = 1, 2, \ldots, n$. Il numero delle ennuple della relazione è detto *cardinalità* della relazione; tale numero è, nella realtà delle applicazioni a basi di dati, un numero finito, poiché ogni applicazione tratta un numero finito di dati; tuttavia in alcune trattazioni teoriche si considerano anche relazioni infinite.

Secondo la definizione data:

1) una relazione è un insieme di ennuple; pertanto, non è definito nessun ordinamento tra le ennuple (in un insieme l'ordine degli elementi è irrilevante) e le ennuple sono tra loro distinte;

2) più precisamente, una relazione è un insieme di ennuple ordinate (cioè l'i-esimo valore di ogni ennupla appartiene all'i-esimo dominio); pertanto l'ordinamento tra i domini di una relazione è, da un punto di vista insiemistico, rilevante.

Esempio 1.1

Rappresentiamo mediante una relazione l'informazione relativa al calendario delle lezioni in un corso di laurea universitario. Chiamiamo calendario tale relazione:

calendario $\subseteq$ Docente $\times$ Corso $\times$ Data $\times$ Data

calendario = {⟨Rossi, Logica, 100183, 300583⟩
⟨Costa, Algebra, 200283, 300683⟩
⟨Vela, Fisica, 150183, 300783⟩}

Il dominio Data ha, nella relazione, due ruoli distinti, a indicare, rispettivamente, la data di inizio e la data di fine di un corso. E', quindi, importante tener conto dell'ordinamento dei domini per interpretare correttamente la relazione.

che a ogni occorrei
terizzato da un no
corrispondenza do
buti $\mathcal{A}$ associa il c
Si noti che, con
un insieme di copp

Esempio 1.2
Consideriamo l'
buti:

Docente c
Corso c
Inizio c
Fine c

Notiamo, nell'
minio siano coinci
La ennupla ⟨Ro

{(Docente,

In questo mod
trascurando l'ordi
Abbiamo visto
un frammento di
definite. Una rel
acquisita a un ce
esempio, non tut
siano noti, è rag
non specificati.
giungendo il cos
indica con il sir
saranno ampliam

1.3 Definizio

Un modello d
di interesse per u

1) *livello estensi*
loro definite;

Con l'automazione

```
\documentclass{article}
 \setlength {\topmargin}{-23mm}
 \newtheorem{domanda}{Domanda}
\begin{document}
 \begin{center}
  {\large\bf Tecnologia delle basi di dati (ex Basi di dati, primo modulo)\\
  25 settembre 2006}
 \end{center}

 \begin{domanda}\rm (25\%)
  Considerare le seguenti richieste ricevute da un gestore del controllo
  di concorrenza:

  \begin{center}
    \(r_3(x),r_2(x),r_4(y),w_2(x),c_2,r_6(y),r_1(x),c_1,w_3(x),c_3,
    w_4(y),c_4,w_7(x),c_7,w_6(y),c_6,r_5(x),c_5\)
  \end{center}

  \noindent Indicare possibili effetti del controllo della concorrenza (indicare
  cioè quali operazioni vengono eseguite
  e in quale ordine) prodotti da controllori dei due tipi principali:
  \begin{enumerate}
    \item basato su 2PL
    \item basato su timestamp
\end{enumerate}
\end{domanda}
```

Tecnologia delle basi di dati (ex Basi di dati, primo modulo)
25 settembre 2006

Tempo a disposizione: due ore. **Nota**: è richiesta una "bella copia" comprensibile e ordinata.

Domanda 1 (25%) Considerare le seguenti richieste ricevute da un gestore del controllo di concorrenza (assumendo che si tratti delle prime richieste ricevute dopo l'avvio del sistema e indicando con c_i il commit della transazione i, che permette il rilascio dei lock da essa acquisiti):

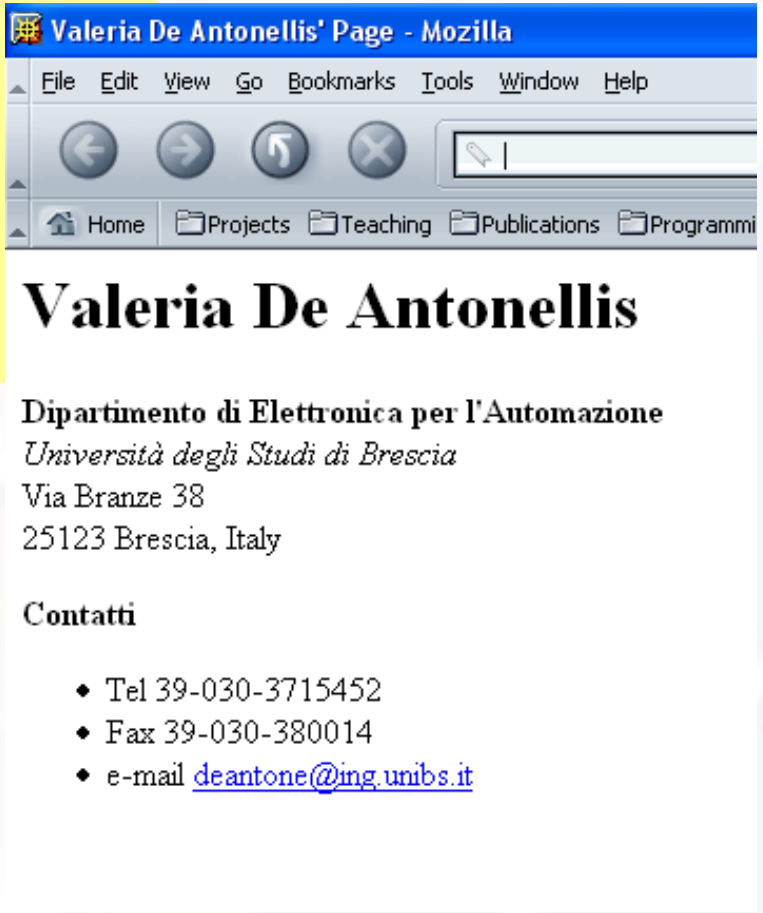
$$r_3(x), r_2(x), r_4(y), w_2(x), c_2, r_6(y), r_1(x), c_1, w_3(x), c_3, w_4(y), c_4, w_7(x), c_7, w_6(y), c_6, r_5(x), c_5$$

Indicare possibili effetti del controllo della concorrenza (indicare cioè quali operazioni vengono eseguite e in quale ordine) prodotti da controllori dei due tipi principali:

1. basato su 2PL; in questo caso supporre che: (a) quando una transazione viene bloccata a causa della mancata concessione di un lock, le sue richieste "rinviate" arrivino poi una dopo l'altra, quando il lock viene concesso; (b) che lo stallo venga immediatamente rilevato e che venga risolto uccidendo la transazione che ha formulato l'ultima delle richieste che hanno causato lo stallo; (c) ogni transazione uccisa per risolvere lo stallo venga riavviata subito e sia in grado di richiedere immediatamente le azioni svolte in precedenza (dopo però le

Con l'avvento del Web

```
<html>
 <head>
  <title>Valeria De Antonellis' Page</title>
 </head>
 <body>
  <h1>Valeria De Antonellis</h1>
  <p>
   <b>Dipartimento di Elettronica per l'Automazione</b><br />
   <i>Universit&agrave; degli Studi di Brescia</i><br />
   Via Branze 38 <br />
   25123 Brescia, Italy
  </p>
  <p>
   <b>Contatti</b><br />
   <ul>
     <li>Tel 39-030-3715452</li>
     <li>Fax 39-030-380014</li>
     <li> e-mail <a href="mailto:deantone@ing.unibs.it">
              deantone@ing.unibs.it</a></li>
   </ul>
  </p>
 </body>
</html>
```



Limite di HTML

- Nato per descrivere le modalità di presentazione e non le caratteristiche del contenuto informativo
- Il Web basato su HTML è un'idea geniale, ma soprattutto per un uso prevalentemente individuale
- Un uso condiviso richiede la separazione fra contenuto e presentazione

```
<html>
 <head>
  <title>Valeria De Antonellis' Page</title>
 </head>
 <body>
  <h1>Valeria De Antonellis</h1>
  <p>
   <b>Dipartimento di Elettronica per l'Automazione</b><br />
   <i>Universit&agrave; degli Studi di Brescia</i><br />
   Via Branze 38 <br />
   25123 Brescia, Italy
  </p>
  <p>
   <b>Contatti</b><br />
   <ul>
     <li>Tel 39-030-3715452</li>
     <li>Fax 39-030-380014</li>
     <li> e-mail <a href="mailto:deantone@ing.unibs.it">
              deantone@ing.unibs.it</a></li>
   </ul>
  </p>
 </body>
</html>
```

- **Design Principles of the Web:** The Web is an application built on top of the Internet and, as such, has inherited its fundamental design principles

  - *Interoperability*
  - *Evolution: The Web must be able to accommodate future technologies.*
  - *Decentralization*

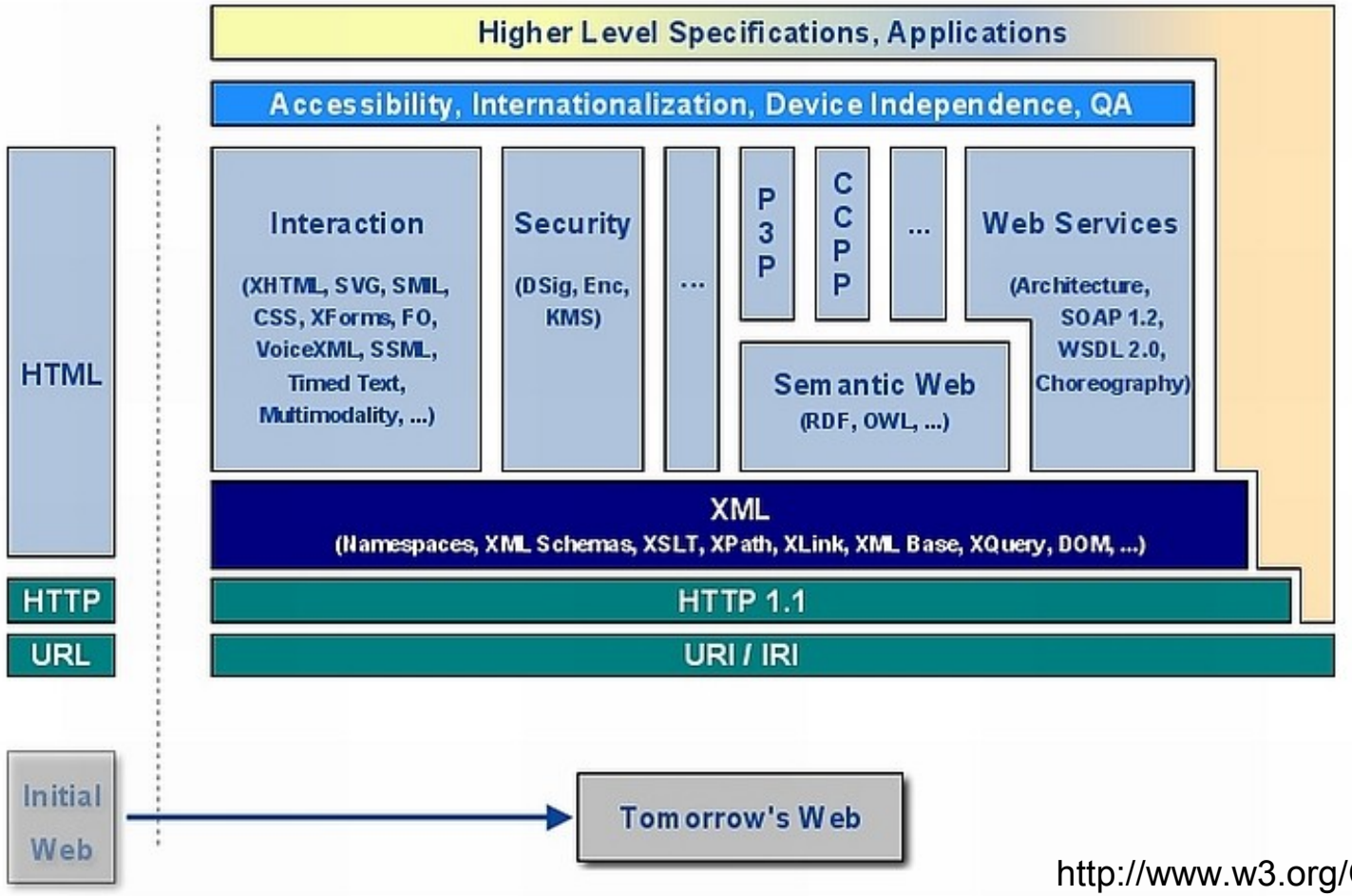
http://www.w3.org/Consortium/

- Nato per descrivere il contenuto informativo

```
<html>
 <head>
  <title>Valeria De Antonellis' Page</title>
 </head>
 <body>
  <h1>Valeria De Antonellis</h1>
  <p>
   <b>Dipartimento di Elettronica per l'Automazione</b><br />
   <i>Universit&agrave; degli Studi di Brescia</i><br />
   Via Branze 38 <br />
   25123 Brescia, Italy
  </p>
  <p>
   <b>Contatti</b><br />
   <ul>
     <li>Tel 39-030-3715452</li>
     <li>Fax 39-030-380014</li>
     <li> e-mail <a href="mailto:deantone@ing.unibs.it">
              deantone@ing.unibs.it</a></li>
   </ul>
  </p>
 </body>
</html>
```

```
<persona>
 <nome>Valeria</nome>
 <cognome>De Antonellis</cognome>
 <affiliazione>
  <divisione>
      Dipartimento di Elettronica per l'Automazione
  </divisione>
  <ente>Università degli Studi di Brescia</ente>
  <indirizzo>
    <via>Via Branze</via>
    <numero>38</numero>
    <cap>25123</cap>
    <città>Brescia</città>
  </indirizzo>
 </affiliazione>
 <telefono>39-030-3715452</telefono>
 <fax>39-030-380014</fax>
 <email>deantone@ing.unibs.it</email>
</persona>
```

XML nel Web del futuro (o presente)



http://www.w3.org/Consortium/

Diffusione di XML (non solo nel Web)

- Formato universale in tutti i settori dell'IT
  - (XML e DTD, XMLSchema, XPath, XQuery, SOAP,..)
- Alcuni esempi:
  - formato di scambio tra applicazioni
  - formato di rappresentazione dati (standard in molti domini applicativi)
  - formato di rappresentazione documenti (word, spread-sheet, etc.)

Sommario

- **Il presente**: XML - origini e diffusione
- **Il passato che non è mai passato**: Basi di Dati - risultati e nuove esigenze
- **Il passato che incontra il presente**: Basi di Dati e XML

- **Raccolta di dati:**
  - **grandi quantità, strutturati e persistenti**
  - **raggruppati in insiemi omogenei in relazione tra loro**
  - **organizzati con la minima ridondanza**
  - **per essere condivisi da applicazioni diverse**
  - **in modo controllato**

  - **PERSISTENZA**
  - **CONDIVISIONE**
  - **AFFIDABILITA' (DBMS: efficienza, efficacia, privatezza, integrità, recovery)**

Basi di dati: la formazione nel tempo

- P. Atzeni, V. De Antonellis
  La teoria relazionale dei dati (Relational Database Theory)
  Boringhieri, **1985** (Benjamin Cummings 1993)

- A. Albano, V. De Antonellis, A. Di Leva
  Computer-aided database design: the DATAID project
  North-Holland, **1985**

  (.....)

- P. Atzeni, S. Ceri, P. Fraternali, S. Paraboschi, R. Torlone
  Basi di dati: architetture e linee di evoluzione
  McGraw-Hill Italia, **2003**

Basi di dati: cambiano le prospettive

- Ogni organizzazione ha **una** base di dati condivisa, che organizza tutti i dati di interesse in forma integrata e non ridondante
- Ogni organizzazione ha di solito **più** basi di dati **distribuite**, **autonome, eterogenee** che devono essere integrate o interoperare (e.g. isole legacy)
- Diverse organizzazioni con proprie basi di dati autonome possono voler cooperare e scambiare dati (e.g. internetworked enterprise, B2B)

A ten-year goal for database research

- The "Asilomar report"
  (Bernstein et al. Sigmod Record 1999
  www.acm.org/sigmod):

  - ***The information utility:***
    make it easy for everyone to store, organize, access, and analyze the majority of human information online

- La maggior parte delle informazioni di interesse non sono nelle basi di dati!
  - **Pagine Web, siti Web**
  - **Banche di dati multimediali**
  - **Librerie di documenti**

- Le basi di dati sono nate per le "applicazioni gestionali", con
  - persistenza, condivisione, affidabilità
  - dati a struttura semplice, con dati di tipo numerico/simbolico
  - transazioni concorrenti di breve durata (OLTP)
  - interrogazioni complesse, espresse mediante linguaggi dichiarativi e con accesso di tipo “associativo”
- Molti dati e informazioni non hanno queste caratteristiche

Aree applicative con altre caratteristiche

- Sistemi multimediali
  - sistemi informativi con documenti, immagini, grafici
  - sistemi di supporto alle decisioni
  - sistemi di gestione ambientale e territoriale

- Sistemi di engineering
  - CAD/CAM (Computer-Aided Design/Manufacturing)
  - CIM (Computer Integrated Manufacturing)
  - CASE (Computer-Aided Software Engineering)

Esempi di "nuove" applicazioni

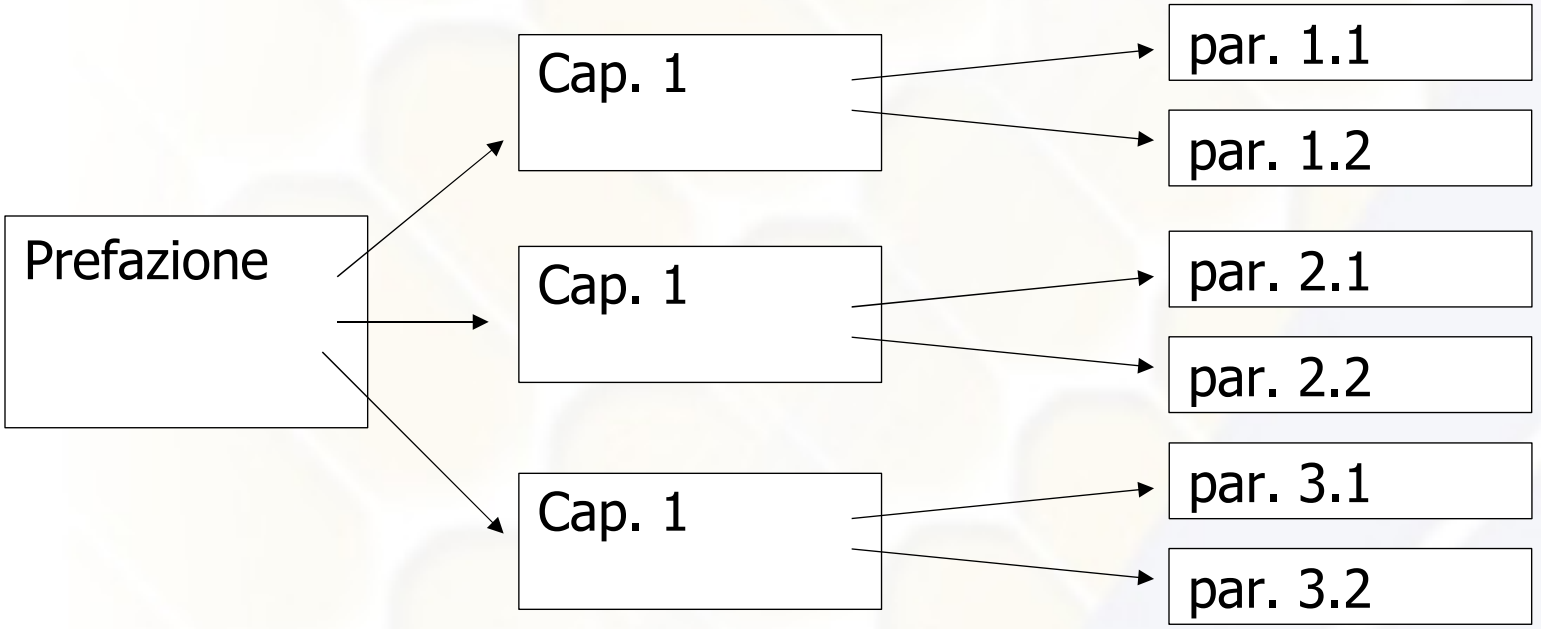
- archivio turistico di fotografie, con didascalie, coordinate geografiche ed esigenze di interrogazioni complesse:
  - "trova le foto con un tramonto scattate a Portofino o dintorni"
- archivio sinistri di una compagnia assicurativa (con foto, grafici, luogo) finalizzato alla ricerca delle frodi
- archivio progettuale con immagini, moduli, versioni temporali finalizzato alla gestione evolutiva di progetti

- Oltre a persistenza, condivisione e affidabilità
  - dati a struttura complessa
    - dati non-numerici — immagini, dati spaziali, sequenze temporali, ...
    - tipi pre-definiti e tipi definiti dall'utente (e riutilizzati)
    - relazioni esplicite ("semantiche") tra i dati (riferimenti), aggregazioni complesse
  - operazioni complesse
    - specifiche per i diversi tipi di dato — es. multimedia
    - associate anche ai tipi definiti dall'utente

Due requisiti fondamentali

- Integrazione e cooperazione tra sistemi distribuiti, autonomi ed eterogenei (interoperabilità)
- Gestione di dati "non tradizionali"

- Alle esigenze di interoperabilità
  - formato di interscambio
- Alle esigenze non tradizionali:
  - strutture complesse e nidificate (ipertesto)



Sommario

- **Il presente**: XML - origini e diffusione
- **Il passato che non è mai passato**: Basi di Dati - risultati e nuove esigenze
- **Il passato che incontra il presente**: Basi di Dati e XML

Una considerazione di fondo

- Basi di dati (relazionali):
  - Strutturate
  - Interrogabili con SQL
  - …
- Web
  - Poca struttura
  - Navigabile con un browser

VORREMMO

“navigare le basi di dati” e “interrogare il Web”

- Basi di dati (relazionali):
  - Strutturate
  - Normalizzate
  - Interrogabili con SQL
- Documenti XML
  - Semistrutturati
  - Gerarchici e non normalizzati
  - Navigabili con browser, interrogabili con XQuery

```
<libro editore="McGraw-Hill" pubblicazione="01/07/2002" formato="paperback">
  <titolo>Basi di dati: modelli e linguaggi di interrogazione</titolo>
  <autore> Paolo Atzeni</autore>
  <autore> Stefano Ceri </autore>
  <autore> Stefano Paraboschi</autore>
  <autore> Riccardo Torlone </autore>
  <parte numero="1">
    <capitolo> Il modello relazionale </capitolo>
    <capitolo> Algebra e calcolo relazionale </capitolo>
    <capitolo> SQL </capitolo>
    <capitolo> SQL nei linguaggi di programmazione</capitolo>
  </parte>
  <parte numero="2">
    <capitolo> Metodologie e modelli per il progetto </capitolo>
    <capitolo> La progettazione concettuale </capitolo>
    <capitolo> La progettazione logica </capitolo>
```

```
for $x IN document("libro.xml")//autore

return $x
```

Basi di dati e XML, vantaggi

- Poter cogliere il meglio dei due mondi:
  - Gestione contestuale di dati tradizionali e dati non gestiti da basi di dati relazionali
  - Supporto all'integrazione (XML è il formato standard di scambio dati)

Basi di dati relazionali e XML, difficoltà

- Come memorizzare documenti XML in basi di dati relazionali?
- Approcci estremi:
  - frammentare e normalizzare, perdendo la visione di insieme (e rischiando la degenerazione delle prestazioni per ricostruire)
  - memorizzare interi documenti come valori di singoli campi, perdendo la flessibilità
- Servono approcci mirati e integrati

- Reale integrazione, che cerchi di conciliare, senza compromessi, le varie esigenze e tecnologie, garantendo efficienza, efficacia e affidabilità

- Gruppi in tutto il mondo, università e laboratori industriali
- In Italia, nel mondo accademico (www.sebd.org)
- Temi principali:
  - Memorizzazione e indicizzazione di documenti XML
  - Esecuzione delle interrogazioni su dati XML
  - Integrazione (stretta o lasca, anche P2P) di informazioni
  - Trasformazione e Mapping semantico di rappresentazioni
  - Ricerca semantica di informazioni (e servizi)

Valeria De Antonellis

Università di Brescia

8 Novembre 2006

Basi di dati e XML: una prospettiva accademica