Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2015/2016

Corso di Laurea Magistrale in Editoria e Giornalismo

XML per l'editoria elettronica

materiale in parte prodotto dalla Dott.ssa Barbara Oliboni

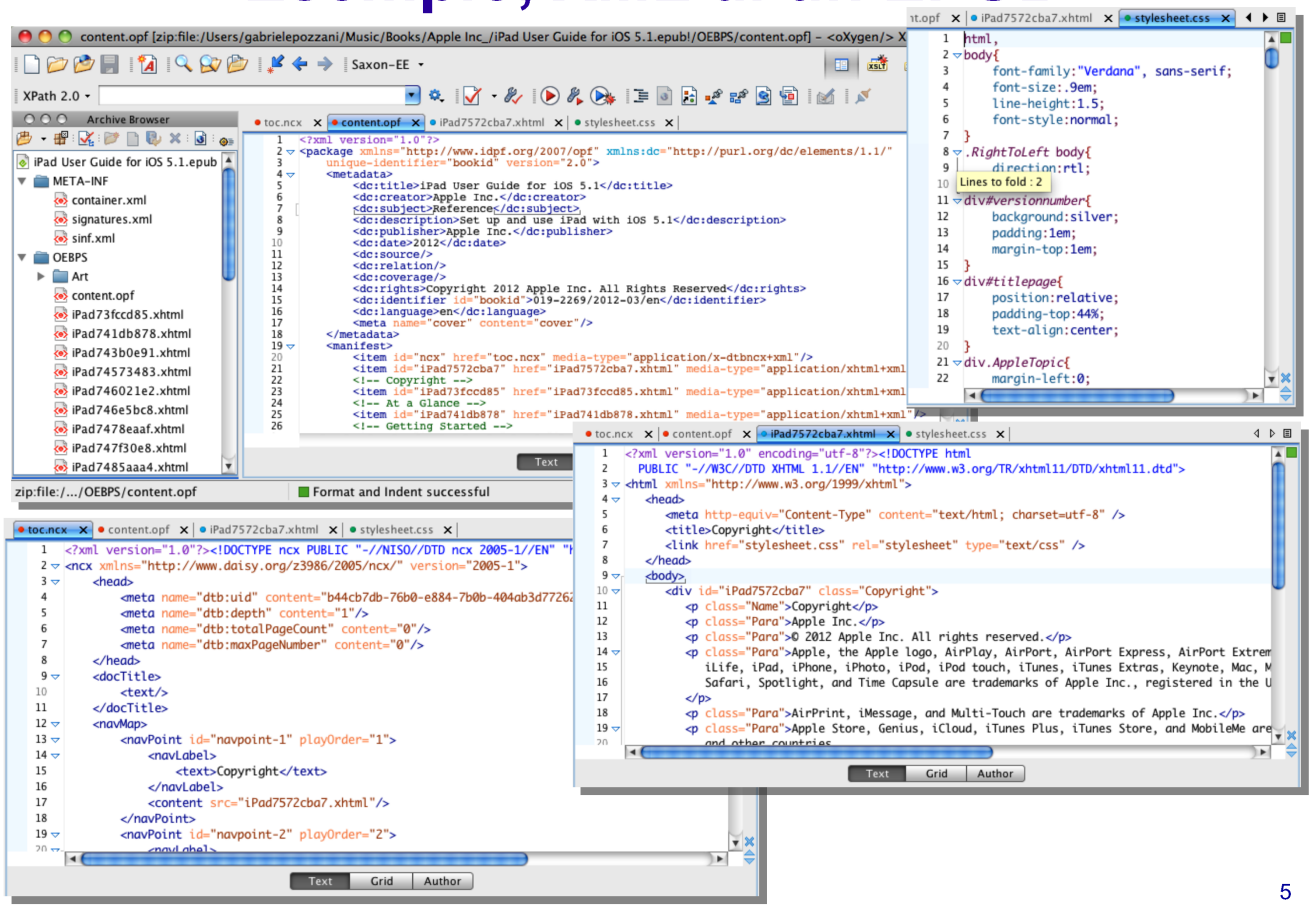
XML come formato documentale

- Utilizzare XML per:
  - Pagine Web
  - Libri
  - Articoli
  - Manuali di riferimento/istruzione
  - Corsi
  - Libri di testo

XML come formato di salvataggio

- I principali formati di salvataggio sono basati su XML:
  - Office Open XML (OOXML, OpenXML): docx, pptx, xlsx
  - Open Document Format (ODF): odt, odp, ods
  - EPUB

Esempio, XML di un EPUB



XML per diversi utenti

- XML soddisfa le esigenze di:
  - Programmatori (o utenti esperti) addestrati a lavorare con le strutture rigide tipiche delle applicazioni orientate ai dati.
  - Scrittori ed editori (o utenti normali) che preferiscono la forma libera di un libro o di un articolo.
- XML soddisfa le esigenze di entrambe le comunità in maniera equa e soddisfacente.

Usi dei tag XML

- L'arricchimento di un documento con una struttura tramite l'uso di XML porta a:
  - Rappresentazione sia del contenuto che di metadati
  - Strutturazione del contenuto tramite più suddivisioni
    - Tra loro indipendenti
    - Con semantiche diverse
    - Con i propri metadati
  - Arricchimento semantico delle parti “interessanti” del contenuto tramite appositi marcatori che ne definiscono il significato del contenuto

TEI (Text Encoding Initiative)

http://www.tei-c.org/

- Applicazione per la marcatura della letteratura classica
- Primo esempio di DTD orientato a documenti di tipo narrativo
- Include elementi per:
  - Le strutture letterarie più comuni (capitoli, scena, stanza, …)
  - La tipografia
  - Le strutture grammaticali

TEI example: Plato, Parmenides (I)

From http://www.perseus.tufts.edu/

```
…
<revisionDesc>
  <change>
    <date>July,1992</date>
      <respStmt>
        <name>WPM</name>
        <resp>(n/a)</resp>
      </respStmt>
      <item>
        Tagged in conformance with Prose.e dtd.
      </item>
  </change>
</revisionDesc>
…
```

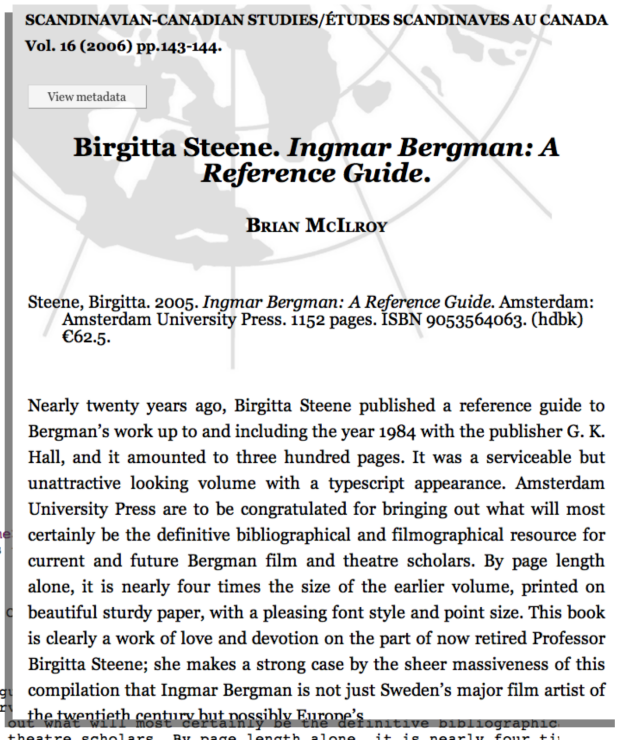

TEI example: Plato, Parmenides (II)

```
…
<text><group>
<text n="Parm.">
<body>
  <head>Parmenides</head>
  <castList>
    <castItem type="role">
      <role>Cephalus</role>
    </castItem>
    <castItem type="role">
      <role>Antiphon</role>
    </castItem>
    <castItem type="role">
      <role>Aristoteles</role>
    </castItem>
  </castList>
…
```

TEI example (2): la rivista Scandinavian-Canadian Studies

Dall'XML vengono ottenuti in automatico la corrispondente pagina Web e il PDF

```
▼<TEI.2 id="mcilroy_1_16">
 ▶<teiHeader>...</teiHeader>
 ▼<text>
  ▼<front>
   ▼<docTitle n="Ingmar Bergman: A Reference Guide">
    ▼<titlePart type="Main">
       <name reg="Steene, Birgitta">Birgitta Steene</name>
       .
       <title level="m">Ingmar Bergman: A Reference Guide</title>
       .
     </titlePart>
    ▼<titlePart type="ReviewedBook">
     ▼<listBibl>
      ▼<biblStruct>
       ▼<monogr>
        ▼<author>
           <name reg="Steene, Birgitta">Birgitta Steene</name>
         </author>
         <title level="m">Ingmar Bergman: A Reference Guide</title>
        ▼<imprint>
           <publisher>Amsterdam University Press</publisher>
           <pubPlace>Amsterdam</pubPlace>
           <date value="2005">2005</date>
           <biblScope type="pages">1152 pages</biblScope>
         </imprint>
       </monogr>
       <note>ISBN 9053564063. (hdbk) €62.5.</note>
      </biblStruct>
     </listBibl>
    </titlePart>
   </docTitle>
  ▼<docAuthor>
     <name key="mcilroy_brian" reg="McIlroy, Brian">Brian McIlroy</name
     teaches film studies at the University of British Columbia. He is
     <title level="m">World Cinema 2: Sweden</title>
     (London: Flicks Books, 1986).
   </docAuthor>
   <titlePart type="short_affil">Dept. of Film, University of British C
  </front>
 ▼<body>
  ▼<div0>
   ▼<p>
     Nearly twenty years ago, Birgitta Steene published a reference gu
     G. K. Hall, and it amounted to three hundred pages. It was a ser
     Amsterdam University Press are to be congratulated for bringing
     filmographical resource for current and future Bergman film and theatre scholars. By page length alone, it is nearly four ti
```



SCANDINAVIAN-CANADIAN STUDIES/ÉTUDES SCANDINAVES AU CANADA
Vol. 16 (2006) pp.143-144.

View metadata

Birgitta Steene. *Ingmar Bergman: A Reference Guide.*

BRIAN MCILROY

Steene, Birgitta. 2005. *Ingmar Bergman: A Reference Guide.* Amsterdam: Amsterdam University Press. 1152 pages. ISBN 9053564063. (hdbk) €62.5.

Nearly twenty years ago, Birgitta Steene published a reference guide to Bergman's work up to and including the year 1984 with the publisher G. K. Hall, and it amounted to three hundred pages. It was a serviceable but unattractive looking volume with a typescript appearance. Amsterdam University Press are to be congratulated for bringing out what will most certainly be the definitive bibliographical and filmographical resource for current and future Bergman film and theatre scholars. By page length alone, it is nearly four times the size of the earlier volume, printed on beautiful sturdy paper, with a pleasing font style and point size. This book is clearly a work of love and devotion on the part of now retired Professor Birgitta Steene; she makes a strong case by the sheer massiveness of this compilation that Ingmar Bergman is not just Sweden's major film artist of the twentieth century but possibly Europe's

DocBook

http://www.oasis-open.org/docbook/

- Applicazione per documenti nuovi (non vecchi)
- Formato per la creazione del testo (non di un prodotto finito e pronto per essere presentato al pubblico)
- Utilizzato nella documentazione relativa al campo informatico
- Sintassi molto semplice
- Modulare e quindi utilizzabile solo in parte (porzioni e strutture che servono)

DocBook example (I)

```
<!DOCTYPE article PUBLIC "-//OASIS//DTD DocBook V4.1//EN">
<article>
  <articleinfo>
    <title>An Example Article</title>
    <author>
      <firstname>Your first name</firstname>
      <surname>Your surname</surname>
      <affiliation>
        <address>
          <email>foo@example.com</email>
        </address>
      </affiliation>
    </author>
    <copyright>
      <year>2000</year>
      <holder>Copyright string here</holder>
    </copyright>
```

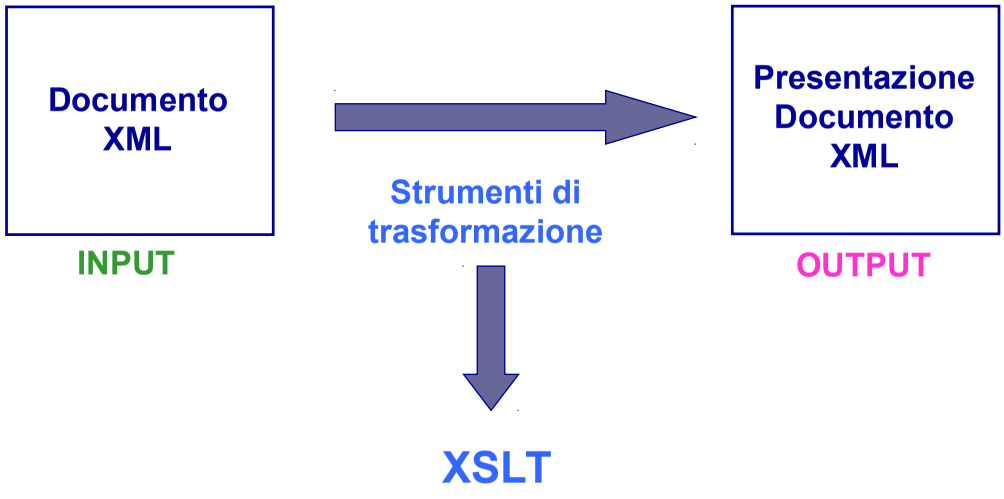

DocBook example (II)

```
    <abstract>
      <para>If your article has an abstract
            then it should go here.</para>
    </abstract>
  </articleinfo>
  <sect1>
    <title>My First Section</title>
    <para>This is the first section
          in my article.</para>
    <sect2>
      <title>My First Sub-Section</title>
      <para>This is the first sub-section
            in my article.</para>
    </sect2>
  </sect1>
</article>
```

Trasformazione e presentazione (1)

- La marcatura di un documento XML descrive la struttura del documento e non la sua presentazione.
  - Specifica l'organizzazione.
  - Non specifica come deve apparire.
- Un documento XML può essere letto nel suo formato nativo (marcatori + testo).
- Normalmente viene tradotto in un formato differente adatto alla presentazione.
- Per XML il formato di INPUT non deve necessariamente corrispondere al formato di OUTPUT.
  - Il formato di INPUT serve per agevolare chi scrive.
  - Il formato di OUTPUT serve per agevolare chi legge.

Trasformazione e presentazione (2)



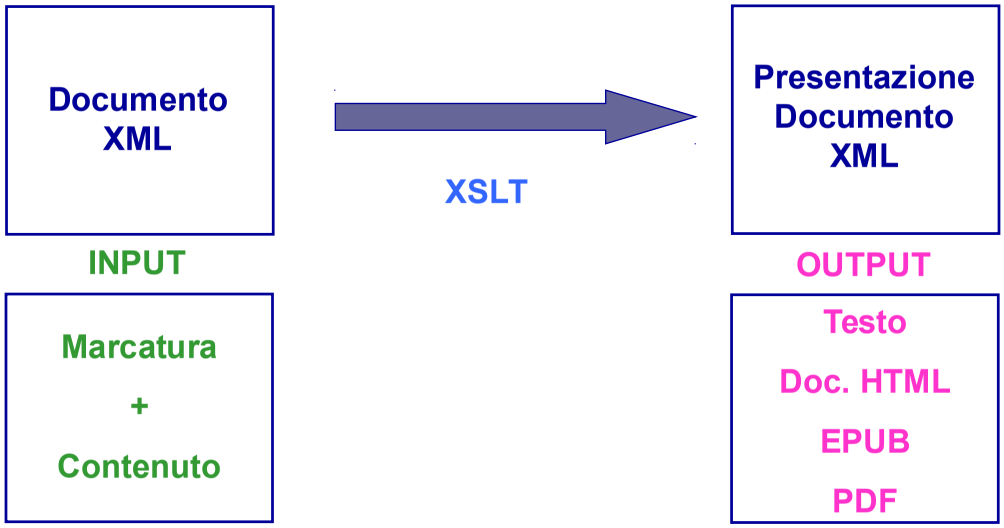
XSLT (eXtensible Stylesheet Language Transformation)

- L'XLST descrive il processo secondo cui un documento XML viene trasformato per produrre il documento finale, pronto per la presentazione
- Un documento XSL contiene un elenco di modelli
  - Definisce dei fogli di stile
- Ogni stile verrà applicato a determinati elementi in modo da renderli “presentabili”

XSLT: funzionamento (1)

- Legge documento in INPUT
- Quando incontra qualcosa (del documento in INPUT) che corrisponde a un modello del foglio di stile
  - Produce in OUTPUT il modello con il relativo contenuto
- Il modello rappresenta il modo in cui il contenuto deve venir presentato

XSLT: funzionamento (2)



XML per l'editoria

- Come si situano tutte queste tecnologie nell'ambito dell'editoria?
- Perché gli editori dovrebbero usare/aver-bisogno di XML?
- Vediamo alcune considerazioni
  - effettuate sia da informatici che da editori

  che possano aiutarci a rispondere a queste domande

XML per l'editoria elettronica

- Come si può facilmente comprendere l'XML, in quanto tecnologia informatica, mostra i suoi maggiori vantaggi nell'ambito dell'editoria elettronica
  - L'editoria elettronica è ancora in uno stadio di comprensione e sviluppo
    - Gli editori non hanno ancora ben compreso come sfruttare le nuove possibilità
      - Quale modello di business?
      - Quali sono le nuove funzionalità?
  - L'uso dell'XML è ancor più limitato

Utilizzi di XML per l'editoria

- La formattazione di documenti/libri/articoli in XML apre due principali possibilità:

1) Arricchimento del testo con metadati e annotazioni semantiche
   - Permettono una maggiore efficacia e ricchezza nella ricerca di informazioni
   - SRI per dati semistrutturati
2) Processo di lavorazione “singolo input, output multipli”
   - Flusso di lavoro XML

Singolo input, output multipli (I)

- XML è un formato
  - Neutrale
  - Intermediario
  - Di interscambio
- Esso può essere usato come punto di partenza per un processo di produzione editoriale basato sul riuso
  - Uno stesso doc XML può essere usato per produrre, automaticamente, diversi formati/tipi di pubblicazione

Singolo input, output multipli (II)

- Uno stesso XML può essere usato per produrre PDF, vari ebook, pagine web, …
  - Con o senza DRM

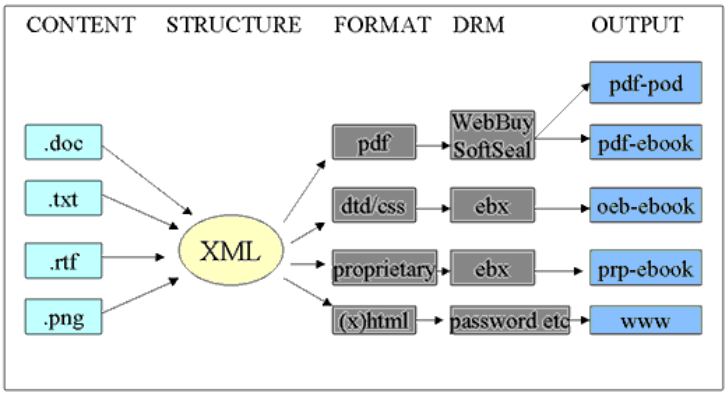


Figure: From raw content and XML to distributed publications

- Per le vecchie produzioni, l'XML può essere ottenuto dai classici formati doc, txt, …
- Per le nuove produzioni, l'XML può essere prodotto direttamente per poi ottenere, se necessario anche documenti in formato doc, …

XML Workflow

- Più in generale l'XML è alla base dell'XML workflow, o flusso di lavoro XML
  - In contrapposizione alla lavorazione “tradizionale”

Flusso di lavoro tradizionale

- Con la lavorazione tradizionale
  - i contenuti vengono importanti nel programma di impaginazione
  - l'impaginatore deve applicare manualmente uno ad uno gli stili appropriati a ciascun elemento del testo (titoli di capitoli, titoli di sezioni, sottotitoli, ecc...)
- Svantaggi:
  - richiede un certo tempo ed è noioso e ripetitivo
  - è possibile commettere errori
  - Non c'è alcun modo di riutilizzare questo lavoro per un libro che abbia delle caratteristiche simili

Flusso di lavoro XML (I)

- Il flusso di lavoro viene separato in due processi paralleli: il lavoro sui contenuti e il lavoro sulla presentazione
  - Da un lato l'editor organizza i contenuti in modo tale da consentire al wordprocessor di applicare dei tag XML a ciascun blocco di testo
    - Un tag XML, in questo caso, non è altro che un'etichetta che descriva il tipo di blocco di testo
    - Può essere fatto già con MS Office e LibreOffice

Flusso di lavoro XML (II)

  - L'impaginatore preparara la presentazione, cioè crea gli stili appropriati per ciascun tipo di elemento strutturale del testo (titoli di capitolo, sottotitoli, box, immagini, …)
    - Usando XSL
    - Per ciascun “formato” di libro che si vorrà ottenere

Flusso di lavoro XML (III)

- Quando l'editor e l'impaginatore hanno concluso il loro lavoro
  - Si importa l'XML dei contenuti nel programma di impaginazione
  - Si chiede al programma di impaginazione di applicare i corretti stili a ciascun blocco di testo

Flusso di lavoro XML (IV)

- Vantaggi:
  - L'applicazione degli stili è automatica, dunque si risparmia tempo
  - Non possono esservi errori nell'applicazione di stili.
  - Il template di stili creato dall'impaginatore è riutilizzabile per qualsiasi altro libro che abbia delle caratteristiche analoghe.
    - Dunque è possibile ad esempio produrre il secondo volume di una collana di ricette lavorando solo alla strutturazione dei contenuti e passare direttamente all'importazione dei contenuti nel programma di impaginazione.
  - L'impaginazione ottenuta con questo processo è molto più idonea alla generazione di ebook

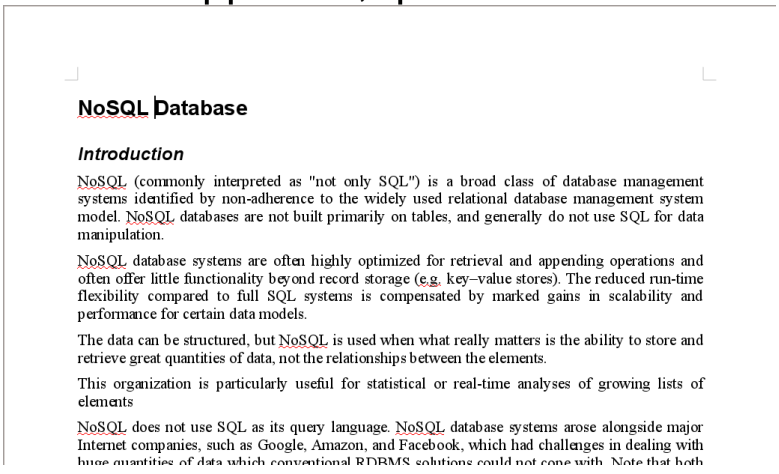
Flusso di lavoro XML (V)

- I classici wordprocessor (MS Office e LibreOffice) sfruttano già l'XML
  - I file .docx, .xlsx, .odt, … non sono altro che uno ZIP di file XML
  - Entrambi permettono di esportare in XML
    - LibreOffice permette di esportare in formato XML valido per DocBook
- Vi sono poi editor appositi per XML
- I principali sw di impaginazione basati su XML permettono di importare direttamente i file docx, odt, …

Flusso di lavoro XML, esempio (I)

1)Preparo i contenuti in LibreOffice

- Applicando gli stili che diventeranno poi tag XML
  - Non guardo a come esse appaiono, quello sarà compito dell'XSL



NoSQL Database

Introduction

NoSQL (commonly interpreted as "not only SQL") is a broad class of database management systems identified by non-adherence to the widely used relational database management system model. NoSQL databases are not built primarily on tables, and generally do not use SQL for data manipulation.

NoSQL database systems are often highly optimized for retrieval and appending operations and often offer little functionality beyond record storage (e.g. key–value stores). The reduced run-time flexibility compared to full SQL systems is compensated by marked gains in scalability and performance for certain data models.

The data can be structured, but NoSQL is used when what really matters is the ability to store and retrieve great quantities of data, not the relationships between the elements.

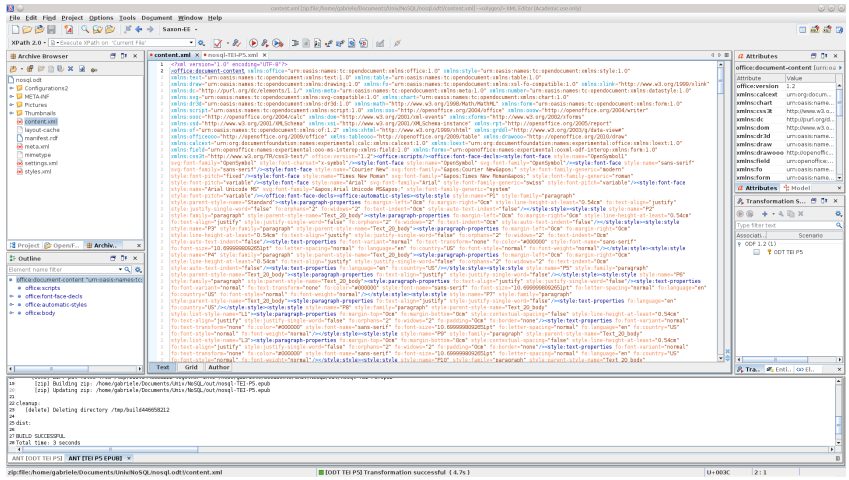
This organization is particularly useful for statistical or real-time analyses of growing lists of elements

NoSQL does not use SQL as its query language. NoSQL database systems arose alongside major Internet companies, such as Google, Amazon, and Facebook, which had challenges in dealing with huge quantities of data which conventional RDBMS solutions could not cope with. Note that both

Flusso di lavoro XML, esempio (II)

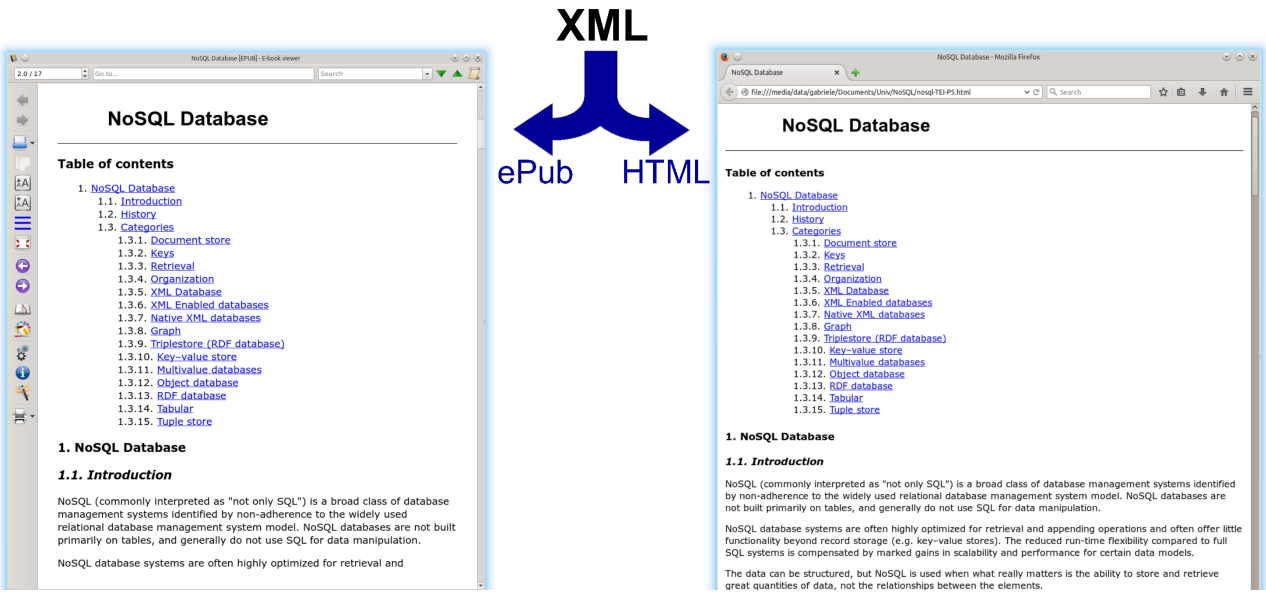
2)Apro l'ODT con l'impaginatore (e.g., <oXygen/>)

- Questo dà accesso direttamente all'XML contenuto nell'odt



Flusso di lavoro XML, esempio (III)

- L'impaginatore contiene già alcuni XSLT di base che possono essere applicati e/o modificati



Obiezioni

- Diversi editori obiettano che
  - il riuso non è così importante
  - I maggiori costi (economici, temporali, …) nell'uso dell'XML non sono quindi giustificati
- Queste obiezioni nascono da considerazioni principalmente editoriali “di vecchio stampo”
- Ovviamente non sempre è necessario il riuso
  - E con esso le tecnologie che lo supportano
  - Dipende dal contesto lavorativo e dal contenuto
- Il fatto che le tecnologie esistano, non significa che vadano sempre usate, ma innanzitutto vanno comprese

Ribattere alle obiezioni

- Certamente l'e-publishing è un ambito in cui il riuso è fondamentale
  - E le tecnologie che lo supportano possono
    - Semplificare
    - Velocizzare
    - Automatizzare

    il processo lavorativo
- Il passaggio ad un processo di produzione editoriale basato sull'XML richiede
  - La formazione di editori ed autori
  - La comprensione della separazione tra contenuto e sua presentazione
  - La comprensione che è il costrutto semantico ad essere importante, non la presentazione finale
- Con lo sviluppo dell'e-publishing ci si può aspettare che riuso e XML possano crescere di importanza e in necessità

<oXygen/>

http://en.wikipedia.org/wiki/Comparison_of_XML_editors

- Editor XML
  - Permette l'editing di documenti XML sia in formato testuale che “grafico” (WYSIWYM)
  - Permette la validazione del documento rispetto al suo DTD
  - Supporto particolare all'editing di documenti TEI, DocBook, e altri principali formati XML
    - Inclusi i formati MS Office, LibreOffice e EPUB
  - Trasformazione dei doc XML tramite XSLT
    - Include già alcuni XSLT per la trasformazione in EPUB, PDF, (X)HTML
  - Creazione ed editing degli stili XSLT