Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2013/2014

Corso di Laurea Magistrale in Editoria e Giornalismo

XML per l'editoria elettronica

materiale in parte prodotto dalla Dott.ssa Barbara Oliboni

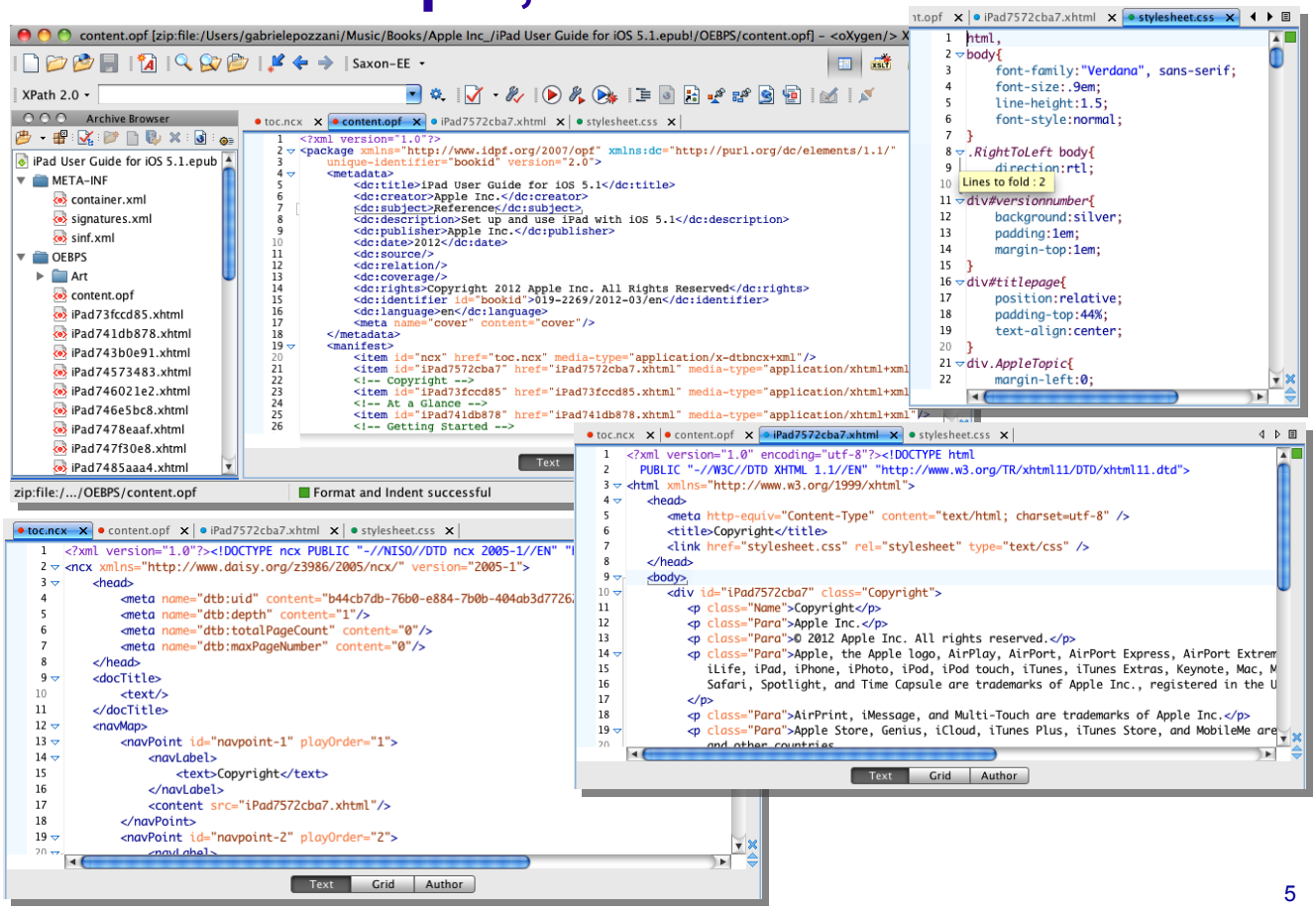
XML come formato documentale

- 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

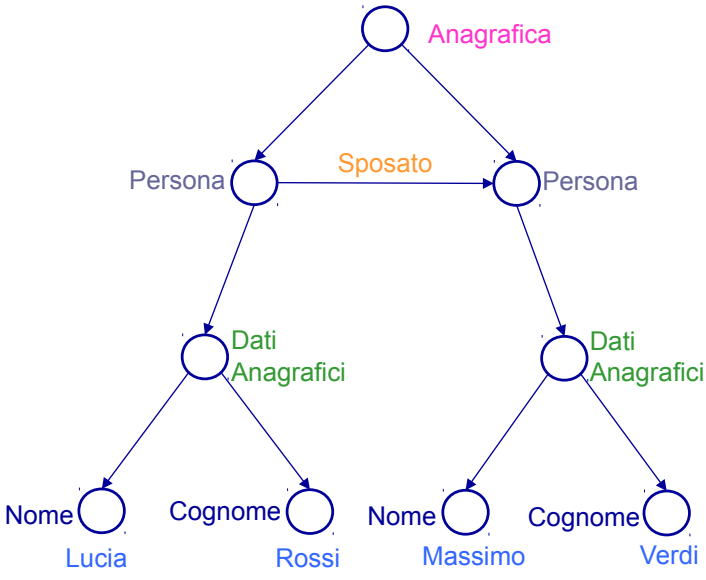
- XML soddisfa le esigenze di:
  - Programmatori (o utenti esperti) addestrati a lavorare con le strutture rigide tipiche delle applicazioni orientate ai dati.
  - Scrittori ed editori (o utenti normali) che preferiscono la forma libera di un libro o di un articolo.
- XML soddisfa le esigenze di entrambe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

Documenti XML: struttura

- Alberi
  - Senza particolari restrizioni su:
    - modo in cui i nodi sono ordinati
    - come i nodi sono connessi fra loro
    - verso quali nodi può connettersi ogni nodo
- Grafi
  - Orientati
    - Etichettati su nodi
    - Etichettati su nodi e archi

Grafo XML: esempio

```
<Anagrafica>
 <Persona id="obj1">
  <Sposato idref="obj2">
  <DatiAnagrafici>
       <Nome> Lucia </Nome>
       <Cognome> Rossi </Cognome>
 </DatiAnagrafici>
 </Persona>
 <Persona id= "obj2">
  <DatiAnagrafici>
       <Nome> Massimo </Nome>
       <Cognome> Verdi </Cognome>
 </DatiAnagrafici>
 </Persona>
</Anagrafica>
```



Grafo con etichette su nodi e archi

Documenti XML orientati alla narrazione: struttura (1)

- Il documento ha una radice che potrebbe essere assimilata al documento stesso
  - Se il documento fosse un libro → **Radice = book**
  - Se il documento fosse un articolo → **Radice = articolo**

Documenti XML orientati alla narrazione: struttura (2)

- Generalmente l'elemento radice (i.e., il documento XML) contiene anche gli elementi che forniscono meta-informazioni:
  - Titolo opera
  - Nome autore
  - Data di stesura del documento
  - Data ultima modifica
  - …

Documenti XML orientati alla narrazione: struttura (3)

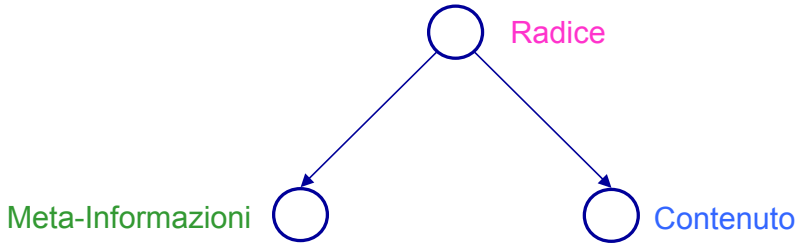
- I documenti di grosse dimensioni sono solitamente suddivisi in sezioni
  - Capitoli di un libro
  - Sezioni di un articolo
  - Citazioni di un documento legale
- Tipi differenti di sezioni
  - Una per l'indice capitoli
  - Una per l'indice termini
  - Una per ognuno dei capitoli di un libro

Documenti XML orientati alla narrazione: struttura (4)

- Le sezioni di un documento possono essere suddivise in sottosezioni
- Le sottosezioni possono essere ulteriormente suddivise
- Ogni sezione o sottosezione generalmente ha:
  - Un titolo
  - Elementi e attributi che forniscono meta-informazioni sulla (sotto)sezione stessa

Documenti XML orientati alla narrazione: struttura (5)

- Meta-informazioni in un elemento figlio della radice.
- Contenuto del documento in un altro elemento figlio della radice.
- Stessa struttura di HTML.



Documenti XML orientati alla narrazione: struttura (6)

- I documenti orientati alla narrazione non sono composti da una sequenza di caratteri qualsiasi, ma da parole
- Contengono un testo concepito per essere letto da persone
- Sono caratterizzati da un flusso narrativo

Documenti XML orientati alla narrazione: struttura (7)

- Il testo vero e proprio potrebbe essere suddiviso in:
  - Paragrafi
  - Altri elementi di tipo blocco:
    - Titoli
    - Figure
    - Note laterali
    - Note a piè di pagina
- I DTD di documenti generici (DocBook) non sono in grado di specificare precisamente cosa conterranno tali elementi

Documenti XML orientati alla narrazione: struttura (8)

- I paragrafi e gli altri elementi di tipo blocco contengono principalmente una sequenza di parole, ovvero del testo
- Alcuni di essi possono contenere marcatori all'interno del testo
  - I marcatori permettono di definire la semantica di alcune parti del testo
  - e.g., marcatore per indicare che una certa stringa rappresenta una data, un indirizzo, un personaggio
- La maggior parte del testo non viene caratterizzata da marcatori

Riassumendo, 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>
…
```

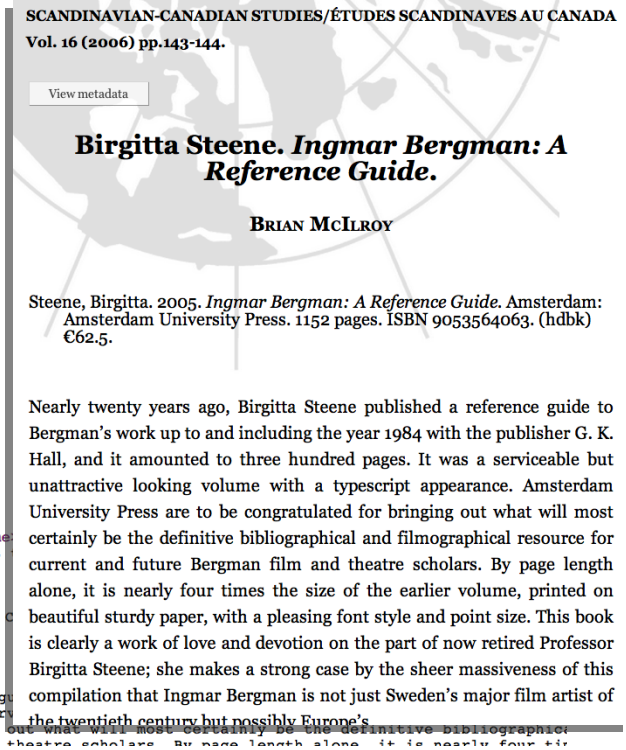

TEI example: Plato, Parmenides (II)

```
…
<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 serv
          Amsterdam University Press are to be congratulated for bringing
          filmographical resource for current and future Bergman film and
```



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>
```

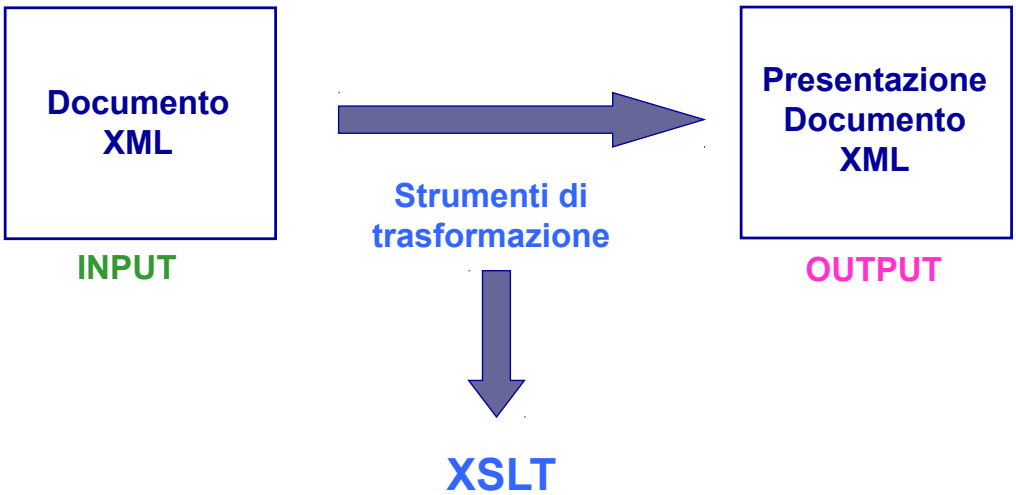

DocBook example (II)

```
    <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)



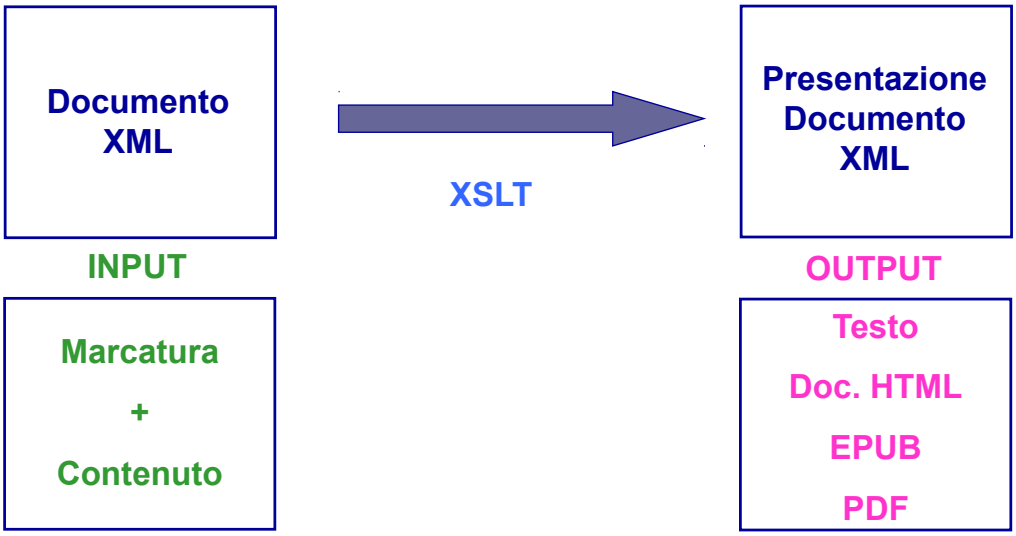
XSLT (eXtensible Stylesheet Language Transformation)

- Un documento XSLT contiene un elenco di modelli
- Ogni modello contiene un pattern che specifica quali elementi e quali nodi vengono messi in corrispondenza
- XSLT definisce dei fogli di stile

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'archivistica

- Oltre per la pubblicazione di singoli documenti, XML è utilizzato anche per la descrizione dei contenuti di archivi, collezioni, musei, biblioteche
  - Encoded Archival Description (EAD) `http://www.loc.gov/ead/`
- EAD
  - Standard XML
  - Sviluppato dall'Università della California a Berkeley e dalla Biblioteca del Congresso USA
  - Include linee guida e un DTD per la definizione delle informazioni/elementi utilizzabili per la descrizione dei documenti in un archivio

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”

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

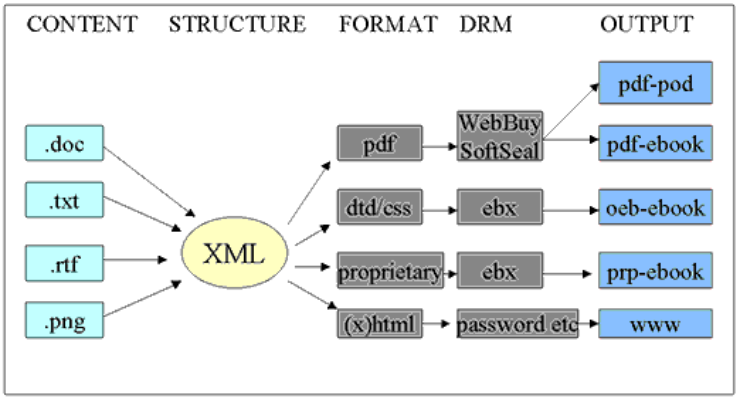


Figure: From raw content and XML to distributed publications

- 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, …

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 del distacco tra contenuto e sua presentazione
  - La comprensione che è il costrutto semantico d 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

Dibattito

- Il dibattito è “vecchio” ed è tuttora in corso, ce se ne può fare un'idea:
  - Hillesund vs. Walsh
  - Post e commenti su TeleRead.com
    - [1], [2], [3]