Esercitazioni: XML e TEI

Mirko Tavosanis

A.a. 2007-2008

Informazione e Scienze Umane

- Informazione strutturata
  - dizionari
  - thesauri
  - …
- Informazione non strutturata (debolmente strutturata)
  - documenti testuali
    - libri, giornali, pagine web, conversazioni, e-mail, chat, ecc.
- Meta-informazione
  - repertori bibliografici
  - cataloghi on-line

Obiettivo dell'esercitazione

- Richiamare i concetti di codifica dell’informazione in XML
- Primi esempi di codifica in XML-TEI

Importante!

- A differenza di altri casi, in XML-TEI buona parte del lavoro di analisi è già fatta
- Invece che analizzare in modo del tutto libero, occorre aderire a uno standard…
- … da interpretare con intelligenza…
- … e con strumenti adeguati

Primo passo: la codifica XML

- Analizzare informazioni (per esempio, testi)
- Vedere in che modo strutturarle (da un minimo a un massimo)
- Codificarle in modo compatibile con XML, cioè:
  - elementi
  - attributi
  - entità
  - DTD o Schema (per la validazione)

Testo non digitale

Oggi a Bruxelles vertice della Ue sull'Iraq. Blair scrive agli altri per chiedere una posizione comune

La Nato in difesa della Turchia

Accordo in extremis sull'aiuto all'alleato in prima linea

DAL NOSTRO CORRISPONDENTE
FRANCO PAPITTO

BRUXELLES — Tutti bloccati per una giornata intera dal piccolo Belgio. Sembrava disponibile la Francia e taceva la Germania: a quel punto Lord Robertson, il Segretario generale della Nato, tentava il grande colpo per chiudere finalmente, con una decisione sul sostegno alla Turchia, la settimana più lacerante mai vissu-

Alle 21 il Belgio aveva strappato quasi tutto ma continuava a ritenere «non soddisfacente» la formulazione del legame fra le decisioni dell'Onu e quelle della Nato.

In questo clima di gravi tensioni si riuniscono stamane i ministri degli Esteri dell'Ue nel tentativo di concordare una posizione comune che dovrebbe essere avallata stasera dai capi di governo. Blair ha scritto nei giorni scorsi a tutti i partner europei per chieder loro di non escludere, nel testo che approveranno stasera, un intervento militare contro Saddam. E' una richiesta che metterà sicuramente in grande imbarazzo Schroeder che ha sinora ha detto «no» all'apartecipazione del suo paese a qualsiasi intervento, deciso o meno dell'Onu. Aznar, il premier spagnolo, ha chiamato al telefono Chirac che gli avrebbe promesso di fare «tutto il possibile» perchè oggi si raggiunga un accordo. Lo stesso Aznar ha scritto al greco Costas Simitis, presidente di turno dell'Ue, per chiedergli che la dichiarazione europea di stasera solleciti una «rigida applicazione» della risoluzione 1441 dell'Onu da parte di Saddam.

Bruxelles ha presentato tre emendamenti sulla solidarietà al governo di Ankara

Il segretario generale della Nato George Robertson

linguaggio
paragrafi
data
didascalia
punteggiatura
titolo
sottotitolo
argomento
autore
caratteri

metatesto

Testo e metatesto machine readable

Oggi a Bruxelles vertice della Ue sull'IRAQ. Blair scrive agli altri per chiedere una posizione comune
<title> La NATO in difesa della Turchia </title>
<subtitle> Accordo in extremis sull'aiuto all' alleato in prima linea </subtitle>
DAL NOSTRO CORRISPONDENTE <author> FRANCO PAPITTO </author>
 BRUXELLES — Tutti bloccati per una giornata intera dal piccolo Belgio.
 Sembrava disponibile la Francia e taceva la Germania; a quel punto Lord Robertson, il Segretario generale della Nato, tentava il grande colpo per chiudere finalmente, con una decisione sul sostegno alla Turchia

titolo
sottotitolo
linguaggio
punteggiatura
caratteri
paragrafi
argomento
data
autore
didascalia

metatesto

Per esempio:

<titolo>Oggi grande svendita</titolo>

<testo>Oggi è prevista una grande svendita di surgelati e prodotti vari al mercato di Piazza delle Vettovaglie</testo>

Requisiti per i file XML

- Ben formati

= 1. C'è un elemento radice che contiene tutto il testo
2. Non ci sono sovrapposizioni tra elementi
3. Ogni elemento ha il tag di apertura e quello di chiusura

(si controlla anche con Internet Explorer)

- Validi

= conformi a una DTD o a un XML Schema

(si controlla con un parser)

☞ Un testo può essere ben formato e al tempo stesso non valido!

- Corretti dal punto di vista semantico

(lo può fare solo un essere umano)

Come si generano i file Xml?

- Evitate i programmi tipo Word
- Usate editor “solo testo”: il semplice blocco note va bene per cominciare – la prossima settimana, Emacs
- Cercate di capire che cosa si trova all’interno del file: solo elementi e caratteri, o qualcos’altro?
- Windows riconosce i file Xml in base all’estensione (.xml) che però può essere non visibile (attivare la visualizzazione con: Opzioni cartella – Visualizzazione – Nascondi le estensioni per i tipi di file conosciuti deselezionato)
- Per l’editing avanzato esiste un’infinità di soluzioni... noi ne vedremo qui una (se non ci sono problemi)

TEI Emacs

- Emacs è un programma di scrittura per utenti avanzati
- È relativamente complesso da imparare
- La versione che useremo (se ne trovano altre versioni sul sito TEI: www.tei-c.org) è predisposta per il lavoro con XML-TEI:
  - Contiene le DTD
  - Include un parser (programma per verificare la validità del file)
  - Inserendo </ chiude automaticamente il tag aperto

Importante!

- Molti comandi *non* funzionano secondo lo standard Windows
- Per esempio, non esiste il “nuovo file”: si crea un nuovo file usando il comando “apri” e scrivendo il nome di un file ancora non esistente
- Nel dubbio, usate i menu in alto

File ben formato

1. Creare con il blocco note il file manzoni.xml sul desktop
2. Inserire all'interno del file questo testo:

<Libro>

<Titolo>I promessi sposi</Titolo>

<Autore><Nome>Alessandro</Nome><Cognome>Manzoni</Cognome></Autore>

<Commento>Romanzo storico</Commento>

</Libro>

3. Salvare
4. Aprire con Internet Explorer

Requisiti per i file XML

- Ben formati

= 1. C'è un elemento radice che contiene tutto il testo
2. Non ci sono sovrapposizioni tra elementi
3. Ogni elemento ha il tag di apertura e quello di chiusura

(si controlla anche con Internet Explorer)

- Validi

= conformi a una DTD o a un XML Schema

(si controlla con un parser)

☞ Un testo può essere ben formato e al tempo stesso non valido!

- Corretti dal punto di vista semantico

(lo può fare solo un essere umano)

File valido (1)

1. Creare con il blocco note il file studente.xml
2. Inserire la DTD:

```
<!DOCTYPE Studente [
<!ELEMENT Studente (Matricola, Nome, DataNascita)>
<!ELEMENT DataNascita (Giorno, Mese, Anno)>
<!ELEMENT Nome (#PCDATA)>
<!ELEMENT Giorno (#PCDATA)>
<!ELEMENT Mese (#PCDATA)>
<!ELEMENT Anno (#PCDATA)>
]>
```

File valido (2)

3. Inserire di seguito il contenuto del file:

<Studente>
<Nome>Francesco Rossi</Nome>
<DataNascita>
<Giorno>10</Giorno>
<Mese>11</Mese>
<Anno>1980</Anno>
</DataNascita>
</Studente>

Inserire...

- Paola Bianco
- Nata a Pisa il 14 settembre 1986
- Numero di matricola 12121

Poi, provare a danneggiare il testo

Strumenti: XML-TEI

- Un vocabolario e una DTD per XML o SGML, che fissano i nomi degli elementi e il modo in cui possono comparire all'interno di un testo
- XML-TEI è uno standard per la codifica di testi
- Permette di validare un testo (il che consente controlli sulla coerenza formale, scambio di dati, ecc.)
- Il sistema è relativamente complesso
- Se ne può usare una versione di base (Core tagset) e una vista semplificata (TEI Lite)

DTD

```
<!DOCTYPE TEI.2 PUBLIC "-//TEI Consortium//DTD TEI P4//EN"
"e:/www.tei-c.org/Software/tei-emacs/sgml/dtds/tei/tei2.dtd" [
<!ENTITY % TEI.prose 'INCLUDE'>
<!ENTITY % TEI.linking 'INCLUDE'>
<!ENTITY % TEI.figures 'INCLUDE'>
<!ENTITY % TEI.analysis 'INCLUDE'>
<!ENTITY % TEI.XML 'INCLUDE'>
<!ENTITY % ISOlat1 SYSTEM "e:/www.tei-c.org/Software/tei-
    emacs/xml/dtds/tei/iso-lat1.ent">
%ISOlat1;
<!ENTITY % ISOlat2 SYSTEM "e:/www.tei-c.org/Software/tei-
    emacs/xml/dtds/tei/iso-lat2.ent">
%ISOlat2;
<!ENTITY % ISOnum  SYSTEM "e:/www.tei-c.org/Software/tei-
    emacs/xml/dtds/tei/iso-num.ent">
%ISOnum;
<!ENTITY % ISOpub  SYSTEM "e:/www.tei-c.org/Software/tei-
    emacs/xml/dtds/tei/iso-pub.ent">
%ISOpub;
]>
```

Struttura di base

```
<TEI.2>
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title></title>
   </titleStmt>
   <publicationStmt>
    <!-- one of (publisher distributor authority pubPlace
     address idno availability date p) -->
   </publicationStmt>
   <sourceDesc>
    <!-- one of (recordingStmt scriptStmt listBibl biblStruct
     biblFull bibl p) -->
   </sourceDesc>
  </fileDesc>
 </teiHeader>
 <text>
  <!-- one of (body group) -->
 </text>
</TEI.2>
```