Da Xml a Xml-TEI per la codifica dei testi

Informatica per le scienze umane

28 novembre 2008

Mirko Tavosanis

Basato su materiali preparati da Elena Pierazzo

per il corso di Codifica di testi

Tutto quello che avete già visto su XML rimane valido

Concetti chiave:>

- File ben formati (rispettano le regole generali di XML: niente “intrecci”, tag di apertura e tag di chiusura...)
- File validi (oltre a essere ben formati seguono le regole inserite nella DTD)
- Possibilità di interrogazione con motori di ricerca
- Eccetera...

Anche un'opera letteraria può essere codificata in XML...

<poema>

<canto>Canto I

Nel mezzo del cammin di nostra vita...

</canto>

</poema>

- Ne avete già visto degli esempi
- Alla base: testo inserito **all'interno** di elementi XML

... ma è facile perdere il controllo

<poema>
<canto>Canto I
Nel mezzo del cammin di nostra vita...
</canto>
</poema>

<poema>
<canto><titolo>Canto I</titolo>
<versi>Le donne, i cavallier, l'arme, gli amori</versi>
</canto>
</poema>

Problemi tipici

Anche se si condivide la scelta per Xml, gli schemi di codifica proliferano, e quindi:

- I dati non possono essere confrontati tra di loro
- I software sviluppati non possono essere ceduti ad altri

Esperienze passate: grandi raccolte di testi diventate rapidamente inutilizzabili

Una risposta: la TEI

- La Text Encoding Initiative nasce nel 1986
- I componenti:
  - Association for Computing and the Humanities (ACH)
  - Association for Computational Linguistic (ACL)
  - Association for Literary and Linguistic Computing (ALLC)
  - Progetti di ricerca
  - Studiosi singoli
- Nel 2000 la TEI si è trasformata in consorzio

Che cosa fa la TEI?

- The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a **standard** for the representation of texts in digital form. Its chief deliverable is a set of **Guidelines** which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics.
- In addition to the Guidelines themselves, the Consortium provides a variety of supporting resources, including resources for learning TEI, information on projects using the TEI, TEI-related publications, and software developed for or adapted to the TEI.

Storia

- 1994: pubblicazione della prima versione delle *Guidelines (TEI P3)* per SGML
- 1997: passaggio a **XML** (appena inventato)
- 2002: TEI P4 per SGML/**XML** (con **DTD**)
- 2007: **TEI P5** per XML, la versione oggi di riferimento e consigliata per ogni futuro sviluppo (invece delle DTD, usa XML **schema**)

Le Guidelines sono pubblicate sul sito web della TEI:
http://www.tei-c.org/

The Text Encoding Initiative

TEI: Yesterday's information tomorrow

Home

News

Electronic Text Editing
Electronic Textual Editing is a volume of essays jointly sponsored by the Modern Language Association and the TEI Consortium, and scheduled for publication in paper form in late

The Text Encoding Initiative (TEI) Guidelines are an international and interdisciplinary standard that facilitates libraries, museums, publishers, and individual scholars represent a variety of literary and linguistic texts for online research, teaching, and preservation.

The TEI standard is maintained by a Consortium of leading Institutions and Projects worldwide. Information on projects which use the TEI, who is a member, and how to join, can all be found via the links above. Consortium members contribute to its financial stability and elect members to its Council and Board.

The Guidelines are the chief deliverable of the TEI Consortium, along with a range of tutorials, case studies, presentations, and software developed for or adapted to the TEI. The latest release of the Guidelines under development is P5.

The TEI was originally sponsored by the Association of Computers in the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association of Literary and Linguistic Computing (ALLC). Major support has been received from the U.S. National Endowment for the Humanities (NEH), the European Community, the Mellon Foundation, and the Social Science and Humanities Research Council of Canada.

La codifica TEI

Marcatura di tipo dichiarativo-strutturale e non procedurale:

- Non si dice “Da qui comincia il corsivo”
- Si dice invece: “**Questo** è un capitolo” (e si racchiude il capitolo dentro un elemento Xml)
- (diversi formati, per esempio Rtf, usano descrizioni procedurali)
- Viene data molta importanza agli elementi di struttura: capitoli, sezioni, paragrafi e così via

Le Guidelines

- www.tei-c.org/P4X/ (Nota bene: noi ci baseremo ancora sulla P4)

Ambizioni:

- fornire un formato standard per l'interscambio di informazioni
- fornire una guida per la codifica in questo formato
- supportare la codifica di molte caratteristiche di ogni genere di testo
- essere indipendente dalle applicazioni

Conseguenze

- Alla base, la TEI si fonda su XML e Unicode
- La TEI si basa poi su moltissimi **elementi predefiniti** (la "sezione" del documento è un , il paragrafo è <p>, il verso è <l>...)
- Gli elementi quindi **non** possono essere inventati liberamente (anche se esiste un sistema per **estendere** la dotazione predefinita)
- I livelli di codifica sono molto diversi: c'è un nucleo di base **obbligatorio** a cui possono essere facoltativamente aggiunte molte cose...
- ... in vista di scopi diversi (linguistico, filologico, letterario...)

In altre parole

- Le *Guidelines* **non** dicono che per esempio devono essere codificati gli aspetti linguistici di un testo, o quelli di impaginazione, eccetera.
- Le *Guidelines* dicono invece: “se vuoi codificare per esempio gli aspetti linguistici, segui questo standard”
- L’importante è che sia rispettato il livello di codifica di base (per fortuna molto ridotto)

Strumento fondamentale: la DTD

- Nella DTD, come avete visto, vengono inserite le regole che il documento deve rispettare
- Per esempio: nella DTD si stabilisce che un elemento <autore> **deve** contenere un elemento <nome> e uno <cognome>, e così via...
- In questo esempio, se l'elemento <autore> del documento non contiene <nome> e <cognome>, il documento non è valido
- Una DTD professionale può contenere migliaia di vincoli di questo genere

Composizione della DTD TEI

- Esiste un nucleo base della DTD
- Il nucleo base include gli elementi generici che possono comparire in ogni tipo di testo
- Alla base **possono** essere aggiunti blocchi supplementari di DTD per vari tipi fondamentali di documenti: testo in prosa, testo in versi, testo teatrale...
- Esistono anche pezzi di DTD per la rappresentazioni di caratteristiche evidenziate da particolari prospettive analitiche ed applicazioni specializzate: codifica di fonti primarie (manoscritti) e di apparati di varianti, codifica di strutture morfosintattiche, eccetera
- Altri blocchi possono essere creati per rispondere a esigenze specifiche degli utenti

La DTD TEI è modulare e parametrizzata = raggruppa gli elementi, gli attributi e persino i *content model* (o porzioni degli stessi), in classi

Rispetto alle normali DTD, una DTD parametrizzata è...

- Più facile da creare e gestire
- Più difficile da leggere

Voi avete visto finora solo DTD compilate (e saranno le uniche richieste)

Importante: anche per questo motivo, non potrete modificare direttamente la DTD TEI!

L’elemento <p>, compilato

```
<!ELEMENT p
      (#PCDATA | ident | code | kw | abbr | address | date | name
      | num | rs | time | add | corr | del | orig | reg | sic
      | unclear | formula | emph | foreign | gloss | hi | mentioned
      | soCalled | term | title | ptr | ref | xptr | xref | s
      | seg | gi | eg | bibl | biblFull | figure | cit | q | label
      | list | listBibl | note | stage | table | text | anchor
      | gap | index | interp | interpGrp | lb | milestone | pb)* >

<!ATTLIST p
      corresp IDREFS #IMPLIED
      next IDREF #IMPLIED
      prev IDREF #IMPLIED
      ana IDREFS #IMPLIED
      id ID #IMPLIED
      n CDATA #IMPLIED
      lang IDREF #IMPLIED
      rend CDATA #IMPLIED
      TEIform CDATA "p" >
```

L’elemento <p>: parametrizzato

```
<!ENTITY % p 'INCLUDE' >
<!ENTITY % n.p "p">
<![ %p; [
<!ELEMENT %n.p; %om.RO;  %paraContent;>
<!ATTLIST %n.p;
      %a.global;
      TEIform CDATA 'p'  >
]]>
```

TEI Lite

- www.tei-c.org/Lite/
- In italiano: www.tei-c.org/Lite/teiu5_it.htm
- Una “vista” adatta a tutti i gusti (più o meno…)
- Un sottoinsieme ragionato della DTD estesa
- Adatta per le esigenze poste da progetti di codifica di corpus testuali e dalle creazioni di vasti archivi documentali
- È meno adeguata per la codifica di testi a fini di ricerca specifica

I metadati: il teiHeader

Le informazioni *sul* testo vengono inserite in una sezione introduttiva del testo stesso e contengono informazioni su:

- il tipo di testo codificato
- la fonte
- il tipo di codifica adottato
- il responsabile della codifica
- le successive revisioni del testo.

4 parti

- una descrizione del **file**, contenuta in <fileDesc> (OBBLIGATORIO)
- una descrizione della **codifica**, contenuta in <encodingDesc>
- un **profilo** del testo, contenuto in <profileDesc>
- una cronologia delle **revisioni**, contenuta in <revisionDesc>

Il minimo…

```
<teiHeader>
  <fileDesc>
   <titleStmt>
    <title>Il Canzoniere di Petrarca: versione elettronica
    </title>
   </titleStmt>
   <publicationStmt>
    <publisher>Universit&agrave; degli Studi di Pisa</publisher>
   </publicationStmt>
   <sourceDesc>
<p>Testo esemplato sull'edizione critica curata da G. Contini</p>
   </sourceDesc>
  </fileDesc>
 </teiHeader>
```

Testi unitari e testi compositi

- Testo sicuramente unitario: *Promessi sposi* (più o meno)
- Testo sicuramente composito: *Tutte le poesie* di Giovanni Pascoli (raccolte diverse...)
- E i *Rerum Vulgarium Fragmenta*? Dipende...

Struttura testi TEI

- **<text>** contiene un singolo testo di qualsiasi tipo, unitario o composito; per esempio una poesia, un testo drammatico, una raccolta di saggi, un romanzo, un dizionario, un corpus.
- **<front>** contiene il materiale introduttivo (intestazione, frontespizio, prefazione, dedica, ecc.) che si incontra prima dell'inizio del testo vero e proprio.
- **<body>** contiene il corpo di un singolo testo unitario, escluso qualsiasi materiale preliminare o finale.
- **<back>** contiene appendici, indici, ecc. che seguono la parte principale del testo.
- **<group>** contiene il corpo di un testo composito, raggruppando sequenze di testi distinti, che sono considerabili in ogni caso come legati fra di loro (ad esempio la raccolta delle opere di un autore, una sequenza di saggi, ecc). L'elemento <group> deve contenere almeno un elemento <text>, a sua volta contenente almeno l'elemento <body> ed eventualmente gli elementi <front> e <back>.

Struttura testo unitario

```
<TEI.2>
<teiHeader> <!-- metadati -->
</teiHeader>
<text>
<front>      <!-- il materiale iniziale va qui -->
</front>
<body>       <!-- il corpo del testo va qui -->
</body>
<back>       <!-- il materiale finale va qui -->
</back>
</text>
</TEI.2>
```

Struttura testo composito

```
<TEI.2>
<teiHeader> <!-- metadati --> </teiHeader>
<text>
   <front>     <!-- il materiale iniziale del testo composito va qui. -->    </front>
   <group>
     <text>
       <front>         <!-- materiale iniziale del primo testo unitario-->    </front>
       <body>          <!-- corpo del testo del primo testo unitario  -->        </body>
       <back>          <!-- materiale finale del primo testo unitario -->        </back>
     </text>
     <text>
       <body>          <!-- corpo del testo del secondo testo unitario  -->      </body>
     </text>
   </group>
   <back>     <!-- materiale finale del testo composito  -->    </back>
  </text>
</TEI.2>
```

Modello di codifica

1. Testo in prosa
2. Possibilità di collegare parti diverse
3. Inserimento di immagini
4. Marcatura di nomi di persona e di luogo, date
5. Analisi linguistica

Elementi per la segmentazione del testo

1. Paragrafi

<p>

- Attributi
  - globali

2. Divisioni strutturali

 <div0> <div1> <div2> <div3> <div4> <div5> <div6> <div7>

- Attributi
  - type: tipologia
  - globali

Attributi globali: tra i più diffusi...

- id = identificativo
- n = per numerazione e altro

Codifica del *Milione* di Marco Polo

```
<?xml version="1.0"?>
<!DOCTYPE TEI.2 PUBLIC "-//TEI//DTD TEI Lite XML ver. 1//EN"
"/tei-emacs/xml/dtds/tei/teixlite.dtd" []>
<TEI.2>
<teiHeader>
  <fileDesc>
   <titleStmt>
    <title>Il Milione: versione elettronica
    </title>
   </titleStmt>
   <publicationStmt>
    <publisher>Universit&agrave; degli Studi di Pisa</publisher>
   </publicationStmt>
   <sourceDesc>
<p>Testo esemplato sull'edizione critica</p>
   </sourceDesc>
  </fileDesc>
<profileDesc>
<langUsage>
<language id='ita'>Italiano</language>
<!-- eventuali altre lingue -->
</langUsage>
</profileDesc>
 </teiHeader>
```

<text>
<body lang="ita">
<div1 id="cap1" type="capitolo">
<head type="ordinale">1</head>
Signori imperadori, re e duci e·ttutte altre genti che volete sapere le diverse generazioni delle genti e·lle diversità delle regioni del mondo, leggete questo libro dove le troverrete tutte le grandissime maraviglie e gran diversitadi delle genti d'Erminia, di Persia e di Tarteria, d'India e di molte altre province. E questo vi conterà il libro ordinatamente siccome messere Marco Polo, savio e·nnobile cittadino di Vinegia, le conta in questo libro e egli medesimo le vide. Ma ancora v'à di quelle cose le quali elli non vide, ma udille da persone degne di fede, e però le cose vedute dirà di veduta e·ll'altre per udita, acciò che 'l nostro libro sia veritieri e sanza niuna menzogna.
...
E·ssì·vvi dico ched egli dimorò in que' paesi bene trentasei anni; lo quale poi, stando nella prigione di Genova, fece mettere inn—iscritto tutte queste cose a messere Rustico da·pPisa, lo quale era preso in quelle medesime carcere ne gli anni di Cristo <num>1298</num>.
</div1>
<div1 id="cap2" type="capitolo">
<head type="ordinale">2</head>
<head type="descrittivo">Lor partita di Gostantinopoli.</head>
Egli è vero che al tempo che Baldovino era imperadore di Gostantinopoli — ciò fu ne gli anni di Cristo <num>1250</num> —, messere Niccolaio Polo, lo quale fu padre di messere Marco, e messere Matteo Polo suo fratello, questi due fratelli erano nella città di Gostantinopoli venuti da·vVinegia con mercatantia, li quali erano nobili e savi sanza fallo. Dissono fra·lloro e ordinorono di volere passare lo Gran Mare per guadagnare, e andarono comperando molte gioie per portare, e partironsi in su una nave di Gostantinopoli e andarono in Soldania.
</div1>
</body>
</text>