XML - TEI per la codifica dei testi

Basato su materiali preparati da Elena Pierazzo
per il corso di Codifica dei testi

Tutto quello che avete già visto su XML rimane valido

- File ben formati (rispettano le regole generali di XML: niente “intrecci”, tag di apertura e tag di chiusura... ne riparleremo presto)
- 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

- Proliferazione degli schemi di codifica
- Scarsa confrontabilità dei dati
- Poca disponibilità di software

In passato: grandi raccolte di testi sono diventate rapidamente inutilizzabili, perché basate su codifiche che non era più possibile supportare (per esempio, MS-Dos)

La TEI: storia e componenti

"The Text Encoding Initiative Consortium is an international organization whose mission is to develop and maintain guidelines for the digital encoding of literary and linguistic texts. The Consortium publishes the Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange: an international and interdisciplinary standard that is widely used by libraries, museums, publishers, and individual scholars to represent all kinds of textual material for online research and teaching."

- Fondata nel 1987 e formata da un gruppo internazionale di studiosi, università e associazioni:
- Association for Computing and the Humanities (ACH)
- Association for Computational Linguistic (ACL)
- Association for Literary and Linguistic Computing (ALLC)

La TEI oggi

Consorzio (fondato nel 2000) ospitato da alcune università:

- Bergen (Norvegia)
- Oxford (Regno Unito)
- Brown University (USA)
- Virginia (USA)

Che cosa produce la TEI?

Non direttamente testi codificati, ma **criteri**:

- 1994: pubblicazione della prima versione completa e stabile delle *Guidelines for Text Encoding and Interchange* (TEI P3) per **SGML**
- 1997: passaggio a **XML**
- 2000: TEI P4 (per SGML/**XML**)
- 2007: TEI P5 per XML (espresse usando XML **schema** invece delle **DTD**)
- Il sito: www.tei-c.org

< Text Encoding Initiative >

TEI: Text Encoding Initiative

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. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. 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.

The TEI Consortium is a non-profit membership organization composed of academic institutions, research projects, and individual scholars from around the world. Members contribute financially to the Consortium and elect representatives to its Council and Board of Directors.

Want to become active in the TEI community? Become a TEI Member, join a Special Interest Group, sign up for the TEI-L mailing list, and come to our

La codifica TEI

- Già prima della scelta di XML: predilezione per una marcatura di tipo dichiarativo-strutturale
- vengono usati anche dei marcatori più specifici o procedurali, utilizzabili quando la scelta della marcatura non è praticabile senza introdurre problemi
- Obiettivo: descrivere i testi in modo soddisfacente per tutti (difficile, ma ci si può provare)

Le Guidelines

- www.tei-c.org/P4X/ (Nota bene: la TEI sta incoraggiando il passaggio a P5)

Ambizioni:

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

Conseguenze

- la scelta di SGML, XML e ISO 646
- la preparazione di un ampio set di **elementi predefiniti** (sezione , paragrafo <p>, verso <l>...), che non devono quindi essere reinventati ogni volta
- la distinzione fra codifica **richiesta**, **raccomandata** e **opzionale**
- la codifica per diverse interpretazioni del testo
- la presenza di codifiche alternative per la stessa caratteristica testuale
- la creazione di un sistema per **estensioni** dello schema definite dall'utente

le *Guidelines* non danno suggerimenti o restrizioni quanto all'importanza relativa delle caratteristiche del testo.

La filosofia delle *Guidelines* è "se vuoi codificare questa caratteristica, fallo in questo modo".

Poche delle indicazioni sono vincolanti a priori.

Strumento fondamentale: la DTD (e oggi lo Schema)

- Luogo in cui vengono inseriti (con un linguaggio particolare) i vincoli che il contenuto del file XML deve rispettare
- Per esempio: un elemento <autore> **deve** contenere un elemento <nome> e uno <cognome>, e così via...
- Una DTD "professionale" può contenere migliaia di vincoli di questo genere
- Oggi si usano i più flessibili Schema; per uniformità con il resto del corso noi continuiamo a usare la vecchia versione

Struttura della DTD TEI

- insiemi di elementi generici che possono comparire in ogni tipo di testo (***core tag sets***)
- insiemi di elementi specifici per vari tipi fondamentali di documenti: testo in prosa, testo in versi, testo drammatico, dizionari o trascrizione di registrazioni verbali (***base tag sets***)
- insiemi di elementi 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, rappresentazione di strutture interpretative profonde, rappresentazione di strutture ipertestuali (***additional tag sets***)
- insiemi di elementi per esigenze di codifica ausiliarie e specializzate, come la documentazione dello schema di codifica, o la dichiarazione di particolari sistemi di scrittura (***auxiliary DTD***).

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

Parametrizzata ≠ compilata

1. Più facile da aggiornare e realizzare
- Più difficile da leggere e comprendere

Voi avete visto finora solo DTD compilate (e saranno le uniche con cui vi sarà chiesto di familiarizzare)

NB: non potrete modificare direttamente la DTD TEI – ci sono alcuni strumenti (complessi) per estenderla in formato compatibile, ma spesso si prende così com'è (sia pure a blocchi)

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'  >
]]>
```

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" >
```

Progetti che usano TEI

- Elenco sul sito TEI
- Biblioteca Italiana
- Laboratori Linguistica
- Ecc...

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 sezioni dell'header

- una descrizione del file, marcata <fileDesc> (OBBLIGATORIO)
- una descrizione della codifica, marcata <encodingDesc>
- un profilo del testo, marcato <profileDesc>
- una cronologia delle revisioni, marcata <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 unitario: *Promessi Sposi*
- Testo composito: *Opere complete* di Giordano Bruno
- Petrarca, *Canzoniere*?

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 qualsiasi materiale prefatorio (intestazione, frontespizio, prefazioni, dedicatorie, 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 qualsiasi tipo di 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 prefatorio 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 prefatorio del testo composito va qui. -->    </front>
   <group>
     <text>
       <front>         <!-- materiale prefatorio 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, di luogo, date
5. Analisi linguistica

Modello di codifica: vista TEI

Es:

```
<!DOCTYPE TEI.2 PUBLIC "-//TEI Consortium//DTD TEI
    P4//EN" " tei2.dtd" [
<!ENTITY % TEI.prose 'INCLUDE'>
<!ENTITY % TEI.linking 'INCLUDE'>
<!ENTITY % TEI.figures 'INCLUDE'>
<!ENTITY % TEI.names.dates 'INCLUDE'>
<!ENTITY % TEI.analysis 'INCLUDE'>
<!ENTITY % TEI.XML 'INCLUDE'>
]>
```

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

Titoli e chiusure

- <head>: titolo
  - Attributi
    - type: tipologia
    - globali
- <trailer>: chiusura (es.: Fine Atto Primo)
  - Attributi
    - 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>
</TEI.2>

Elementi in sintesi

<text>: il testo unitario

<body>: il corpo del testo

- Attributi
  - lang: lingua del testo

<div1>: divisioni di primo livello

- Attributi
  - type: capitolo
  - Id: identificativo

<head>: il titolo

- Attributi
  - type: descrittivo, ordinale

<p>: paragrafi

<num>: numeri