Ιόνιο Πανεπιστήμιο
Τμήμα Αρχειονομίας – Βιβλιοθηκονομίας

Μεταπτυχιακό πρόγραμμα σπουδών
Στην Επιστήμη της Πληροφορίας

Μεταδεδομένα στο ψηφιακό περιβάλλον: TEI (Text Encoding Initiative)

Κατερίνα Τοράκη

Απρίλιος 2004

TEI

- Ξεκίνησε το 1987 με σκοπό την κωδικοποίηση και ανταλλαγή ηλεκτρονικών κειμένων (κύρια στο χώρο των ανθρωπιστικών σπουδών και της γλωσσολογίας). Μπορεί να εφαρμοστεί σε κείμενα ανεξάρτητα φυσικής γλώσσας, χρονικής κάλυψης, μορφής και περιεχομένου, σε συνεχόμενα (κείμενα) ή μη συνεχόμενα υλικά (π.χ. λεξικά).

- Το TEI παρέχει τα μέσα για την αναπαράσταση χαρακτηριστικών του κειμένου που είναι απαραίτητα να προσδιοριστούν προκειμένου να διευκολύνεται η επεξεργασία του από προγράμματα υπολογιστών.

TEI

- Για το λόγο αυτό, χρησιμοποιούνται ετικέτες (tags) που παρεμβάλλονται σε σημεία του κειμένου (στην ηλεκτρονική του αναπαράσταση) και δίνουν τη δυνατότητα της κωδικοποίησης (markup, encoding) συγκεκριμένων χαρακτηριστικών του για περαιτέρω αξιοποίηση (αναζήτηση, ανταλλαγή κλπ).

- Η κωδικοποίηση του κειμένου βασίζεται στη γλώσσα SGML η XML. (Για περισσότερες πληροφορίες και αναφορές, βλέπε : http://www.tei-c.org/Lite/U5-bibapp.html)

TEI

- Δε σχετίζεται με τη μορφοποίηση και τον τρόπο εμφάνισης του κειμένου.

Παράδειγμα:

  - There are very few *risqué* passages in *Paradise Lost*
  - Html: There are very few risqué passages in Paradise Lost
  - TEI: There are very few <foreign>risqué</foreign> passages in <title>Paradise Lost</title>

- Καλύπτει:
  - Δομή (παράγραφοι, σελίδες, διάλογοι, υποσημειώσεις, σύνδεσμοι)
  - Γλωσσική επεξεργασία (διάλεκτοι, ονόματα, προτάσεις, λέξεις, εκφράσεις, στοιχεία μετάφρασης)
  - Μεταδεδομένα (βιβλιογραφικά στοιχεία, εκδοτικό ιστορικό κ.λπ.)
  - Εμφάνιση κειμένων

TEI

Θεωρείται ότι υπάρχει ένας βασικός πυρήνας χαρακτηριστικών του κειμένου που μπορούν να κωδικοποιηθούν σε όλα τα κείμενα, ενώ υπάρχει και ένας αριθμός διαφορετικών στοιχείων που μπορεί επίσης να κωδικοποιηθούν. Κάθε κείμενο μπορεί να έχει μία δομή ανάλογα με τον τύπο του ή/και τον σκοπό που εξυπηρετεί.

Για παράδειγμα, μία έκθεση μπορεί να έχει κεφάλαια, τμήματα, παραγράφους, προτάσεις. Ένα ποίημα έχει στροφές, στίχους, γραμμές. Ένα οποιοδήποτε έντυπο δημοσίευμα έχει τόμους, συλλογές, σελίδες.

Γιατί το TEI;

Ομοιότητες TEI με βιβλιογραφική καταλογογράφηση.

Τότε γιατί άλλο ένα πρότυπο; Για διευκόλυνση παραγωγών πρωτογενών κειμένων που δεν γνωρίζουν κανόνες καταλογογράφησης.

Βασικά στοιχεία

- **<teiHeader>**

  Δεν αποτελεί μέρος του υπο κωδικοποίηση κειμένου, αλλά παρέχει πληροφορίες (μεταδεδομένα) για αυτό.

  – Στοιχεία του teiHeader element:

    fileDesc, profileDesc, encodingDesc, revisionDesc

- **<text> με τα παρακάτω μέρη:**

  – **<front>**

    Περιέχει προκαταρκτικό περιεχόμενο (επικεφαλίδες, σελίδες τίτλων, πρόλογοι κ.λπ.) που βρίσκονται πριν την αρχή του κανονικού κειμένου

  – **<back>**

    Περιέχει παραρτήματα κ.λπ. που ακολουθούν το κυρίως κείμενο

  – **<body>**

    Περιέχει το σώμα ενός μοναδικού κειμένου εκτός του front και back περιεχομένου

TEI - Κωδικοποίηση

- Ο προσδιορισμός του τεκμηρίου που κωδικοποιείται γίνεται με την παράθεση του ορισμού της κατηγορίας του - DTD (Document Type Definition) - που τοποθετείται σε αγκύλες.
- Στο παρακάτω παράδειγμα, ορίζονται τα στοιχεία ενός τεκμηρίου που αφορά ένα ποίημα, τμήμα μίας ανθολογίας, χρησιμοποιώντας διάταξη SGML.

```
<!ELEMENT anthology  - -  (poem+)>
<!ELEMENT poem       - - (title?, stanza+)>
<!ELEMENT title      - - (#PCDATA) >
<!ELEMENT stanza     - - (line+)   >
<!ELEMENT line       - - (#PCDATA) >
```

Κωδικοποίηση απλού κειμένου

```
<TEI.2>
    <teiHeader> [ TEI Header information ]
      </teiHeader>
    <text>
        <front> [ front matter ... ]   </front>
        <body>  [ body of text ... ]   </body>
        <back>  [ back matter ...  ]   </back>
    </text>
  </TEI.2>
```

Κωδικοποίηση σύνθετου κειμένου (composite text)

<TEI.2>
<teiHeader> [περιεχόμενο του στοιχείου teiHeader της συλλογής] </teiHeader>
<text>
<front> [περιεχόμενο του στοιχείου front της συλλογής] </front>
<group>
<text> <--! Εδώ αρχίζει το πρώτο κείμενο -->
<front> [περιεχόμενο του στοιχείου front του πρώτου κειμένου] </front>
<body> [περιεχόμενο του στοιχείου body του πρώτου κειμένου] </body>
<back> περιεχόμενο του στοιχείου back του πρώτου κειμένου] </back>
</text>

Κωδικοποίηση σύνθετου κειμένου (composite text) ...

<text> <--! Εδώ αρχίζει το δεύτερο κείμενο -->

<front> [περιεχόμενο του στοιχείου front του δευτέρου κειμένου] </front>

<body> [περιεχόμενο του στοιχείου body του δευτέρου κειμένου] </body>

<back> [περιεχόμενο του στοιχείου back του δευτέρου κειμένου] </back>

</text>

[σε αυτό το σημείο μπορεί να ακολουθούν και περισσότερα κείμενα]

</group>

<back> [περιεχόμενο του στοιχείου back της συλλογής] </back>

</text>

</TEI.2>

* *Μία κεφαλίδα (header) για όλα τα κείμενα*

Κωδικοποίηση συλλογής κειμένων (TEI corpus)

```
<teiCorpus>
    <teiHeader> [header information for the corpus]</teiHeader>
    <TEI.2>
        <teiHeader>[header information for first text]</teiHeader>
        <text> [first text in corpus] </text>
    </TEI.2>
    <TEI.2>
        <teiHeader>[header information for second text]</teiHeader>
        <text> [second text in corpus] </text>
    </TEI.2>
</teiCorpus>
```

* *Κάθε κείμενο έχει δική του κεφαλίδα (header)*
* *Δεν μπορεί να υπάρξει συλλογή από συλλογές (corpora)*

Κωδικοποίηση κειμένου

Για την κωδικοποίηση ενός κειμένου, χρειάζεται να οριστούν τα παρακάτω:

- **Κατηγορία τεκμηρίου – κωδικοποίηση δομής** (π.χ. πεζό κείμενο, ποίημα, θεατρικό έργο, ομιλία, λεξικό, ορολογία)
- **Πρόσθετα στοιχεία κατηγοριών** (π.χ. σύνδεση υπερκειμένων, σχήματα, πίνακες, γλωσσολογική ανάλυση, μεταγραμματισμός, ονόματα και ημερομηνίες, γραφική παράσταση κλπ)
- **Κωδικοποίηση χαρακτήρων** (πρότυπα ISO, Unicode)

Ακολουθεί η σύνταξη σε γλώσσα κωδικοποίησης SGML ή XML.

TEI Header

Η σελίδα τίτλου σε κάθε ηλεκτρονικό κείμενο κωδικοποιείται με την κεφαλίδα TEI (TEI Header), που έχει τα παρακάτω μέρη:

<fileDesc>

βιβλιογραφική περιγραφή του ηλεκτρονικού αρχείου

<encodingDesc>

τεκμηρίωση της σχέσης του ηλεκτρονικού κειμένου με την πηγή ή τις πηγές απ' όπου παράγεται

<profileDesc>

λεπτομερής περιγραφή των μη βιβλιογραφικών δεδομένων (π.χ. γλώσσα, συμμετέχοντες κλπ)

<revisionDesc>

καταγραφή του ιστορικού αναθεωρήσεων του αρχείου

TEI Header - fileDesc

Παράδειγμα:

```
<teiHeader>
<fileDesc>
    <titleStmt> ... </titleStmt>
    <publicationStmt> ... <publicationStmt>
    <sourceDesc> ... <sourceDesc>
</fileDesc>
</teiHeader>
```

TEI Header – encodingDesc

Παράδειγμα:

```
<encodingDesc>
   <projectDesc>Texts collected for use in the Claremont Shakespeare Clinic, June 1990.
   </projectDesc>
</encodingDesc>
```

```
<encodingDesc>
   <samplingDecl>Samples of 2000 words taken from the beginning of the text
   </samplingDecl>
</encodingDesc>
```

TEI Header – encodingDesc …

```
<editorialDecl>
   <p>The part of speech analysis applied throughout
      section 4 was added by hand and has not been
      checked.
   <p>Errors in transcription controlled by using the
         WordPerfect spelling checker.
   <p>All words converted to Modern American spelling
         using Webster's 9th Collegiate dictionary.
   <p>All quotation marks converted to entity
         references &odq; and &cdq;.
</editorialDecl>
```

TEI Header – profileDesc

```
<creation>
    <date value='1992-08'>August 1992</date>
    <name type="place">Taos, New Mexico</name>
</creation>

<langUsage>

<textClass>
<keywords scheme="LCSH">
     <list>
     <item>English literature -- History and criticism --Data processing.</item>
     <item>English language -- Style – Data processing.</item>
     </list>
</keywords>
</textClass>
```

TEI Header – revisionDesc

```
<revisionDesc>
<change><date>6/3/91:</date>
   <respStmt><name>EMB</name><resp>ed.</resp></re
   spStmt>
   <item>File format updated</item>
</change>
<change><date>5/25/90:</date>
   <respSmt><name>EMB</name><resp>ed.</resp>
   <item>Stuart's corrections entered</item>
</change>
</revisionDesc>
```

TEI Header

- Εάν πρόκειται για συλλογή κειμένων με κοινά χαρακτηριστικά, μπορεί να υπάρχει μία κεφαλίδα για το σύνολο και ξεχωριστές για κάθε συστατικό του. Στην περίπτωση αυτή, ορίζεται ο τύπος της επικεφαλίδας ως εξής:

<teiHeader type="corpus">

- Η περιοχή TEI Header περιέχει τα περιγραφικά δεδομένα ενός ηλεκτρονικού κειμένου που η σύνταξή τους μπορεί να ακολουθήσει τους κανόνες καταλογογράφησης AACR2 καθώς και τις οδηγίες των βιβλιογραφικών περιγραφών ISBD.

Στοιχεία του <body>

1. Βασικά δομικά στοιχεία: **div head**
2. Στοιχεία παραγράφων **p cit q l lg sp**
3. Λίστες, πίνακες και σχήματα: **list item table row cell figure figDesc**
4. Στοιχεία φράσεων: **date emph foreign hi name num soCalled term title**
5. Στοιχεία χωρισμού σελίδων και γραμμών: **milestone pb lb**
6. Στοιχεία για σύνδεση στοιχείων: **ref rs ptr xref xptr**
7. Βιβλιογραφικά στοιχεία: **bibl author editor publisher respStmt resp pubPlace**

Κανόνες δόμησης

- Το **body** ενός κειμένου χωρίζεται από elements (μέρη, κεφάλαια, τμήματα κλπ)
- Τα ** ** elements χωρίζονται σε tags, όπως:

  <p> (παράγραφος

  <q> (εδάφιο με εισαγωγικά)

  <l> (γραμμή)

  <lg> (ομάδα γραμμών, π.χ. stanza, refrain)

  <sp> (λόγος, παράσταση, με το γνώρισμα who)

  <speaker> (ομιλητής)

  <stage> (σε παράσταση, με γνωρίσματα entrance, exit κλπ)
- Τα **<p>** και παρόμοια στοιχεία περιλαμβάνουν κείμενο (#PCDATA), το οποίο κωδικοποιείται από στοιχεία φράσεων

Στοιχεία δομής

- ** **: υποδιαίρεση μέχρι 7 επίπεδα, εφαρμόζεται και στα front, back. Γνωρίσματα (attributes):
  - **type**: 'Book', 'Chapter', 'Part', poem, song, sonnet, speech κ.λπ.
  - **id**: μοναδικός κωδικός υποδιαίρεσης
  - **n**: όνομα ή αριθμός υποδιαίρεσης
- **head**: ο τίτλος της υποδιαίρεσης, <!ELEMENT head #PCDATA>
- **<p>**: σημείο παραγράφου <!ELEMENT p #PCDATA>

```
<div1 id="UGT1" n="Winter" type="Part">
<div2 id="UGT11" n="1" type="Chapter">
<head>Mellstock-Lane</head>
<p>I fully appreciate Gen. Pope's splendid… </p>
```

Στοιχεία δομής - DTD

```
<!ELEMENT anthology   (poem+)>
<!ELEMENT poem        (title?, stanza+)>
<!ELEMENT title       (#PCDATA) >
<!ELEMENT stanza      (line+) >
<!ELEMENT line        (#PCDATA) >
```

προσδιοριστής στοιχείου

μοντέλο περιεχομένου

δείκτης συχνότητας

σειρά των στοιχείων

(parsed character data)

+ : >=1

? : <=1

** : 0 ή >=1*

, : κατά σειράν

| : *εναλλακτικά*

```
<!ELEMENT poem        (title?, (stanza+ | couplet+ | line+) )>
```

Γνωρίσματα

Στο TEI-DTD εφαρμόζονται τα ακόλουθα γνωρίσματα στα στοιχεία:

```
<!ATTLIST element
    id ID #IMPLIED
    n CDATA #IMPLIED
    lang IDREF #IMPLIED (γλώσσα)
    rend CDATA #IMPLIED (τυπογραφική αναπαράσταση)
    status  CDATA #REQUIRED >
```

ID: μοναδικό αναγνωριστικό

CDATA: οποιοιδήποτε χαρακτήρες

IDREF: μοναδικό αναγνωριστικό σε άλλο στοιχείο

#REQUIRED: απαιτείται τιμή

#IMPLIED: δεν απαιτείται τιμή

π.χ.

```
<q lang="FR" rend="italics">
<!ATTLIST poem  id   ID #IMPLIED
                status  (draft l revised l published) #REQUIRED >
```

Παράδειγμα

```
<div1 type ="Act" n="I">
<head>ACT I</head>
    <div2 type ="Scene" n="1">
    <head>SCENE I</head>
    <stage rend="italic"> Enter Barnardo and Francisco, two Sentinels, at
    several doors</stage>
    <sp><speaker>Barn</speaker>
    <l part="Y">Who's there?</l></sp>
    <sp><speaker>Fran</speaker>
    <lg type="stanza" part="I">
    <l>But why drives on that ship so fast</l>
    <l>Withouten wave or wind?</l>
    </lg> </sp>
    <sp><speaker>Barn</speaker><l part="i">Long live the
    King!</l></sp>
    <sp><speaker>Fran</speaker><l part="m">Barnardo?</l></sp>
    <sp><speaker>Barn</speaker><l part="f">He.</l></sp>
    <p> I went to the store to buy<list><item>bread,</item>
    <item>milk,</item> <item>and bananas</item></p>
```

Στοιχεία Φράσεων

- **<emph>** έμφαση φράσης για γλωσσικό ή ρητορικό σκοπό
- **<foreign>** φράση ή λέξη που ανήκει σε άλλη γλώσσα από το τριγύρω κείμενο
- **<term>** τεχνικός όρος
- **<title>** τίτλος με γνωρίσματα:
  - **level** *m* βιβλία, συλλογές, έργα ενός τόμου ή πολύτομα, *s* σειρές, *j* περιοδικό, *u* μη δημοσιευμένο υλικό, *a* αναλυτικός τίτλος που ανήκει σε κάποιο άλλο τεκμήριο (άρθρο, ποίημα κ.λπ.
  - **type** abbreviated, main, subordinate (υπότιτλοι και τίτλοι μερών), parallel (παράλληλοι).

Στοιχεία Φράσεων ...

<name>, <date>, <time>, <num>

- <name type="person"> Walter de la Mare</name> was born at <name type="place">Charlton</name>, in <name type="county">Kent</name>, in <date value="1873-02-21">21 Feb 1873</date>
- <l>specially when it's nine below zero</l> <l>and <time value="15:00">three o'clock in the afternoon</time></l>
- <num value="33">xxxiii</num> <num type="cardinal" value="21">twenty-one</num> <num type="percentage" value="10">ten percent</num> <num type="percentage" value="10">10%</num> <num type="ordinal" value="5">5th</num>

Γραμμές και σελίδες

- **Γραμμές**

  <lb n="25"/> Fie, that you'll say so! He plays o' th'
  <lb n="26"/> viol-de-gamboys, and speaks three or four languages
  <lb n="27"/> word for word without book, and hath all the good
  <lb n="28"/> gifts of nature.

- **Σελίδες**

  I wrote to Moor House and to Cambridge immediately, to say what I had done: fully explaining also why I had thus acted. Diana and <pb ed="ED1" n="475"/> Mary approved the step unreservedly. Diana announced that she would <pb ed="ED2" n="485"/>just give me time to get over the honeymoon, and then she would come and see me.

Αναφορές

Για αναφορές από ένα σημείο σε άλλο σημείο του ίδιου κειμένου, χρησιμοποιούνται τα στοιχεία:

- **ptr**
- **ref** (συνήθως όταν ακολουθεί και άλλο κείμενο)

Για βιβλιογραφικές αναφορές, χρησιμοποιείται το στοιχείο **bibl**

Παραδείγματα:

- See especially <ref target="SEC12">section 12 on page 34</ref>
- See especially <ptr target="SEC12"/>

```
<bibl>
    <title>The Riverside Shakespeare</title>
    <author>Shakespeare,William</author>
    <publisher>Boston: Houghton Mifflin,1974</publisher>
    <biblScope>5. 88ff</biblScope>
</bibl>
```

Ένα ολοκληρωμένο παράδειγμα

```
<?xml version="1.0"?>
<!DOCTYPE TEI.2 SYSTEM "teixbaby.dtd">
<TEI.2>
<teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Hamlet, Prince of Denmark: an electronic edition</title>
                <author>Shakespeare,William</author>
            </titleStmt>
            <publicationStmt>
              <publisher>Houghton Mifflin</publisher>
              <pubPlace>Boston MA</pubPlace>
              <date>1997</date>
        </publicationStmt>
        <sourceDesc>
        <bibl>
                <title>The Riverside Shakespeare</title>
                <author>Shakespeare,William</author>
                <publisher>Boston: Houghton Mifflin,1974</publisher>
        </bibl>
        </sourceDesc>
        </fileDesc>
    </teiHeader>
```

Ένα ολοκληρωμένο παράδειγμα...

```
<text>
<front><div type="castlist">
        <list><item id="Oph">OPHELIA, daughter to Polonius</item>
            <item id="King">CLAUDIUS, King of Denmark</item>
            <item id="Queen">GERTRUDE, Queen of Denmark</item>
        </list>
    </div></front>
<body><div type="act" n="4"><div n="4.5" type="scene">
<stage><hi rend="i">Enter</hi>KING.</stage>
    <sp who="Queen"><speaker> <hi rend="i">Queen.</hi></speaker>
    <l n="37" part="Y"> Alas, look here, my lord.</l>
    </sp>
<sp who="Oph"><speaker>
    <hi rend="i">Oph.</hi></speaker><stage> <hi rend="i">Song.</hi></stage>
    <lg part="M" type="song"><l n="38"> "Larded all with sweet flowers,</l>
    <l n="39"> Which bewept to the ground did not go</l>
    <l n="40"> With true-love showers."</l></lg>
    </sp>
<sp who="King"><speaker> <hi rend="i">King.</hi></speaker>
    <l n="41" part="Y"> How do you, pretty lady?</l>
    </sp>
<sp who="Oph"><speaker> <hi rend="i">Oph.</hi></speaker>
    <p><lb n="42"/> Well, <rs key="God">God</rs> dild you! They say the owl was a
    <lb n="43"/> baker's daughter. Lord, we know what we are, but
    <lb n="44"/> know not what we may be. <rs key="God">God</rs> be at your table!</p>
    </sp>
```

Ένα ολοκληρωμένο παράδειγμα…

```
<sp who="King"><speaker> <hi rend="i">King.</hi></speaker>
    <l n="45" part="Y"> Conceit upon her father. </l>
    </sp>
<sp who="Oph"><speaker><hi rend="i">Oph.</hi></speaker>
    <p><lb n="46"/> Pray let's have no words of this, but when
    <lb n="47"/> they ask you what it means, say you this:</p>
    <stage> <hi rend="i">Song.</hi></stage>
    <lg part="M" type="song">
    <l n="48">"To-morrow is <rs key="StValentine">Saint Valentine's</rs> day,</l>
    <l n="49">All in the morning betime,</l>
    <l n="50"> And I a maid at your window, </l>
    <l n="51"> To be your <rs key="StValentine">Valentine</rs>.</l>
    <l n="52"> "Then up he rose and donn'd his clo'es,</l>
    <l n="53"> And dupp'd the chamber-door,</l>
    <l n="54"> Let in the maid, that out a maid</l>
    <l n="55"> Never departed more."</l>
    </lg>
    </sp>
</div> </div></body>
</text>
</TEI.2>
```

Projects using the TEI

English (including Old/Middle English)

- *African American Women Writers of the 19th Century*
- *American Memory from the Library of Congress*
- *American Radicalism*
- *American Sunday School Books Digitization Project*
- *Boccaccio's Decameron*
- *British National Corpus*
- *British Women Romantic Poets, 1789-1832*
- *The Canterbury Tales Project*
- *CELT Project: The Corpus of Electronic Texts*
- *Chronicon*
- *CJK-English Dictionary*
- *Cursus Project*
- *The Digital Dictionary of Buddhism*
- *A Digitized Library of Southern Literature*
- .....

Projects using the TEI

French (including dialects)

- *CELT Project: The Corpus of Electronic Texts*
- *The Charrette Project*
- *Early Canada Online*
- *Early Modern French Women Writers*
- *Electronic Text Archive Leiden*
- *Emblem Project Utrecht*
- *Kolb-Proust Archive for Research*
- *Multilingual (Basque/Spanish/French/English) Dictionaries*
- *Multilingual Text Tools and Corpora (MULTEXT)*
- *The Oslo Multilingual Corpus*
- *ProQuest*
- *Silfide*
- *Voltaire Foundation*
- ……

Projects using the TEI

Italian (including dialects)

- *Boccaccio's Decameron*
- *Emblem Project Utrecht*
- *Multilingual Text Tools and Corpora (MULTEXT)*
- *Thesaurus Musicarum Italicarum (TMI)*
- *The World of Dante*
- .....

Greek

- *Aphrodisias Project*
- *Electronic New Testament Manuscript Project*
- *The EpiDoc Collaborative*
- *Leiden Armenian Database*
- *Perseus Project*
- .....

http://www.tei-c.org/Applications/index-lang.html

http://www.tei-c.org/Lite/teiu5_en.pdf

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

http://www.ionio.gr/~toraki/metadata_tei_met_2004.ppt