XML και TEI: μια σύντομη εισαγωγή στην κωδικοποίηση κειμένων

Dr. Αθανάσιος Ν. Καρασίμος
akarasimos@gmail.com
akarasimos@academyofathens.gr
DARIAH-GR// PARTHENOS

ΔΥΑΣ // δυά(ς)λογοι #03 | Τετάρτη 10-2-2016

<HTML> και <TEI>

Μια μικρή εισαγωγή στις γλώσσες HTML και TEI

HTML: Ορισμός

- Η HTML (Hyper Text Markup Language) είναι μια γλώσσα σήμανσης η οποία παρέχει πληροφορία για τη δομή και την εμφάνιση του περιεχομένου μια ιστοσελίδας.
- Αποτελείται από ένα σύνολο από ετικέτες σήμανσης (markup tags) και η κάθε μία από αυτές περιλαμβάνει έναν αριθμό από γνωρίσματα (attributes).

HTML: Ετικέτες

- Είναι δεσμευμένες λέξεις που περικλείονται στα σύμβολα ‘<’ και ‘>’
- Συνήθως χρησιμοποιούνται σε ζευγάρια και διακρίνονται σε ετικέτα αρχής (start tag) και ετικέτα τέλους (end tag)
- Η ετικέτα τέλους γράφεται όπως η ετικέτα αρχής, αλλά με ένα slash (/) πριν το όνομα της ετικέτας
- Για παράδειγμα:
  - <body> … </body>
  - …
  - <h1> … </h1>

HTML: Δομή εγγράφου

- ```
  <!DOCTYPE html>
  <html>
      <head>
          <title>Page Title</title>
      </head>

      <body>
          <h1>My First Heading</h1>
          My first paragraph.
      </body>
  </html>
  ```
  ```

XML: Ορισμός

- e**X**tensible **M**arkup **L**anguage (XML) αποτελεί μια εξαιρετικά απλή εκδοχή της γλώσσας **S**tandard **G**eneralized **M**arkup **L**anguage (SGML), η οποία αναπτύχθηκε με στόχο να διευκολύνει το χειρισμό, την επεξεργασία, τη διακίνηση και αποθήκευση τεκμηρίων στον Παγκόσμιο Ιστό (web).
- Αποτελεί συνδυασμό SGML και HTML, δηλαδή η ισχύς της SGML με την απλότητα της HTML.
- Επιτρέπει τον ορισμό νέων γλωσσών σημειοθέτησης, με τη βοήθεια δηλώσεων τύπων εγγράφων (Document Type Declarations - DTDs).

XML: Δομή εγγράφου

- <?xml version="1.0" encoding="UTF-8"?>

```
<shiporder orderid="889923” xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:noNamespaceSchemaLocation="shiporder.xsd">
        <orderperson>John Smith</orderperson>
        <shipto>
                <name>Ola Nordmann</name>
                <address>Langgt 23</address>
                <city>4000 Stavanger</city>
                <country>Norway</country>
        </shipto>
        <item>
                <title>Empire Burlesque</title>
                <note>Special Edition</note>
                <quantity>1</quantity>
                <price>10.90</price>
        </item>
        <item>
                <title>Hide your heart</title>
                <quantity>1</quantity>
                <price>9.90</price>
        </item>
</shiporder>
```

XML: Βασικά δομικά στοιχεία

- *Στοιχεία* (*elements*).
  - Οι βασικές δομικές μονάδες της XML.
  - Ετικέτα αρχής και ετικέτα τέλους.
  - Πρέπει να είναι κατάλληλα εμφωλευμένα.
- Τα στοιχεία μπορούν να διαθέτουν *γνωρίσματα* (attributes) τα οποία παρέχουν επιπλέον πληροφορία αναφορικά με τα στοιχεία.
- *Οντότητες*: όπως οι μακροεντολές, αναπαριστούν ένα συχνά εμφανιζόμενο κείμενο.
- *Σχόλια*.
- *Οδηγίες επεξεργασίας* (*processing instructions*): αναπαριστούν οδηγίες για εφαρμογές.
- *Δηλώσεις τύπων εγγράφων* (Document Type Declarations - DTDs).

XML: Ετικέτες και γνωρίσματα

- Ένα στοιχείο της XML είναι δυνατό να διαθέτει ένα σύνολο από *γνωρίσματα* (attributes).
- Τα γνωρίσματα ορίζονται σαν ζεύγη *ονομάτων–τιμών.*
- Τα γνωρίσματα τοποθετούνται στην ετικέτα αρχής του στοιχείο στο οποίο αναφέρονται.
- Για παράδειγμα:
  - <βιβλίο id=“12-155-419”>
    - <τίτλος>Το Σκήπτρο του Φοίνικα</τίτλος>
    - <συγγραφέας>Αθανάσιος Ν. Καρασίμος</συγγραφέας>

    </βιβλίο>

XML: Ορθώς διαμορφωμένο κείμενο

- Σωστή σειρά εμφάνισης ετικετών:

<κείμενο><είδος></είδος><συγγραφέας></συγγραφέας></κείμενο>

- Λανθασμένη σειρά εμφάνισης ετικετών:

<κείμενο><είδος></είδος><συγγραφέας></κείμενο></συγγραφέας>

```
<κείμενο>
        <είδος>
        </είδος>
        <συγγραφέας>
        </συγγραφέας>
</κείμενο>
```

Text Encoding Initiative

Το πρότυπο κωδικοποίησης κειμένων TEI

Text Encoding Initiative (TEI): Ορισμός

- Το **Text Encoding Initiative** (TEI) είναι ένα πρότυπο για την αναπαράσταση και κωδικοποίηση του γραπτού υλικού σε ψηφιακή μορφή.
- Αποτελεί συνεργατική προσπάθεια μιας κοινότητας ερευνητών, κυρίως από τις Ανθρωπιστικές, Κοινωνικές επιστήμες και τη Γλωσσολογία με το όνομα TEI Consortium. Στόχος τους είναι η ανάπτυξη, δημοσίευση και συντήρηση του προτύπου κωδικοποίησης κειμένου τεκμηριώνοντας τις κατευθυντήριες γραμμές, τη συζήτηση και την ανάπτυξη του προτύπου.
- Ο βασικός οδηγός τους είναι το TEI Guidelines που χρησιμοποιείται ευρέως από μουσεία, βιβλιοθήκες, εκδότες, ερευνητές, ψηφιακά αποθετήρια.

Text Encoding Initiative (TEI): Ιστορικό

- Προηγούμενες απόπειρες
  - COCOA (1960s, 1970s)
  - Oxford Concordance Program - OCP (1980s)
  - Textual Analysis Computing Tools - TACT (1990s)
  - Standard Generalized Markup Language – SGML (9186)
  - eXtensible Markup Language – XML (1998)

Text Encoding Initiative (TEI): Δείγμα XML

- <breakfast_menu>
  - <food>
    - <name>Belgian Waffles</name>
    - <price>$5.95</price>
    - <description>
    - Two of our famous Belgian Waffles with plenty of real maple syrup
    - </description>
    - <calories>650</calories>
  - </food>
  - <food>
    - <name>Strawberry Belgian Waffles</name>
    - <price>$7.95</price>
    - <description>
    - Light Belgian waffles covered with strawberries and whipped cream
    - </description>
    - <calories>900</calories>
  - </food>

Text Encoding Initiative (TEI): Δείγμα XML

- <?xml version="1.0" encoding="UTF-8"?>

<note> → start-tag

<to>**Αθανάσιος**</to>

<from>**Ελένη**</from>

<heading>**Reminder**</heading>

<body>**Μην ξεχάσεις το κείμενο της συνάντησης**
 → line-break

Η παρουσίαση θα γίνει στην Αναγνωστοπούλου.</body>

</note> → end-tag

Text Encoding Initiative (TEI): Συστατικά δομείς

- Processing Instructions
- Elements
- Attributes (optional)
- *Entity References*
- (P)CDATA

```
<?xml version="1.0" encoding="UTF-8"?>
<document><!-- paragraphs go here -->
<paragraph number="1">Paragraph one of <title>an XML
example</title>.</paragraph>
<paragraph number="2">Paragraph two of this example.</paragraph>
</document>
```

Text Encoding Initiative (TEI): Ενότητες (modules)

- Ένα σημαντικό μέρος των κανόνων στο TEI Guidelines ισχύουν για την έκφραση των περιγραφικών και διαρθρωτικών μετα-πληροφοριών σχετικά με το κείμενο. Ωστόσο, το TEI ορίζει τις έννοιες που αντιπροσωπεύουν ένα πολύ ευρύτερο φάσμα των φαινομένων του κειμένου, που ανέρχονται στο συνολικό αριθμό των 503 στοιχείων και των 210 χαρακτηριστικών. Αυτά είναι οργανωμένα σε 21 ενότητες, ομαδοποίηση σχετικών στοιχείων και χαρακτηριστικών.

Text Encoding Initiative (TEI): Ενότητες (modules)

- The TEI Infrastructure
- The TEI Header
- Elements Available in All TEI Documents
- Default Text Structure
- Representation of Non-standard Characters and Glyphs
- Verse
- Performance Texts
- Transcriptions of Speech
- Dictionaries
- Manuscript Description
- Representation of Primary Sources
- Critical Apparatus
- Names, Dates, People, and Places
- Tables, Formulæ, and Graphics
- Language Corpora
- Linking, Segmentation, and Alignment
- Simple Analytic Mechanisms
- Feature Structures
- Graphs, Networks, and Trees
- Certainty and Responsibility
- Documentation Elements

Text Encoding Initiative (TEI): Δείγμα TEI

- <TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
  - <teiHeader>
    - <fileDesc>
      - <titleStmt>
        - <title>A TBE customisation</title>
        - <author>The TBE Crew</author>
      - </titleStmt>
      - <publicationStmt>
        - for use by whoever wants it
      - </publicationStmt>
      - <sourceDesc>
        - created on Thursday 24th July 2008 10:20:17 AM by the form at http://www.tei-c.org/Roma/
      - </sourceDesc>
    - </fileDesc>
  - </teiHeader>
  - <text>
    - <front>
      - <divGen type="toc"/>
    - </front>
    - <body>
      - My TEI Customization starts with modules tei, core, header, and textstructure
      - <schemaSpec ident="TBEcustom" docLang="en" xml:lang ="en" prefix=""><moduleRef key="tei"/>
        - <moduleRef key="header"/>
        - <moduleRef key="core"/>
        - <moduleRef key="textstructure"/>
        - </schemaSpec>
    - </body>
  - </text>
- </TEI>

Text Encoding Initiative (TEI): Δείγματα TEI

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>A sample TEI document</title>
         </titleStmt>
         <publicationStmt>
            <publisher> KANTL </publisher>
            <pubPlace>Ghent</pubPlace>
            <date when="2009"/>
         </publicationStmt>
         <sourceDesc>
            <p>No source, born digital</p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <p>This is a sample paragraph, illustrating a <name
         type="organisation">TEI</name> document.</p>
      </body>
   </text>
</TEI>
```

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <text>
      <body>
         <p>This is a sample paragraph, illustrating a
         <gi>orgName</gi>TEI<gi>orgName</gi> document.</p>
      </body>
   </text>
</TEI>
```

- Το TEI *απαιτεί* το <teiHeader> να είναι πάντα παρόν σε κάθε έγγραφο και ότι προηγείται του <text>.

Text Encoding Initiative (TEI): Δομή εγγράφου

- <TEI></TEI>
  - Το tag που ορίζει την αρχή και το τέλος ενός TEI εγγράφου.
- <teiHeader></teiHeader>
  - Εσωκλείει τις βασικές πληροφορίες και τα μεταδεδομένα του εγγράφου.
- <text></text>
  - Αυτό καθ’ εαυτό το κείμενο.
- <!--...-->
  - Οποιαδήποτε σχόλια στο XML που δεν θα διαβαστούν.

```
<TEI>
   <teiHeader>
      <!--...-->
   </teiHeader>
   <text>
      <!--...-->
   </text>
</TEI>
```

Text Encoding Initiative (TEI): TEI Header

- Η κεφαλίδα <**teiHeader**> είναι υποχρεωτική και περιλαμβάνει περιγραφικές μετα-πληροφορίες σχετικά με το έγγραφο. Το <**teiHeader**> περιέχει τουλάχιστον μια περιγραφή του ηλεκτρονικού αρχείου μέσα στο <**fileDesc**>. Το συγκεκριμένο στοιχείο αποτελείται από τρία υποχρεωτικά συστατικά:
  - η δήλωση του τίτλου <**titleStmt**>, παρέχοντας πληροφορίες για τον τίτλο <**title**>, τον συγγραφέα <**author**> και άλλους υπεύθυνους για το ηλεκτρονικό κείμενο.
  - η δήλωση δημοσίευσης <**publicationStmt**>, παρέχοντας λεπτομέρειες σχετικά με τη δημοσίευση του ηλεκτρονικού κειμένου σε ένα δομημένο τρόπο ή πεζά μέσα σε μια παράγραφο (<**p**>).
  - η περιγραφή της πηγής (<**sourceDesc**>), όπου εντάσσονται βιβλιογραφικά στοιχεία σχετικά με την πηγή του υλικού (αν υπάρχει) με δομημένο τρόπο ή σε μια παράγραφο γενικά (<**p**>).

Text Encoding Initiative (TEI): TEI Header

```
<teiHeader>
   <fileDesc>
      <titleStmt>
         <title>The Strange Adventures of Dr. Burt Diddledygook: a machine-readable transcription</title>
         <respStmt>
            <resp>editor</resp>
            <name xml:id="EV">Edward Vanhoutte</name>
         </respStmt>
      </titleStmt>
      <publicationStmt>
         <p>Not for distribution.</p>
      </publicationStmt>
      <sourceDesc>
         <p>Transcribed from the diaries of the late Dr. Roy Offire.</p>
      </sourceDesc>
   </fileDesc>
</teiHeader>
```

Text Encoding Initiative (TEI): Text

- Το πραγματικό κείμενο <**text**> περιέχει ένα ενιαίο κείμενο οποιουδήποτε είδους. Αυτό συνήθως περιέχει το ίδιο το κείμενο και ίσως άλλες κωδικοποιήσεις. Ένα κείμενο <**text**> περιέχει τουλάχιστον ένα σώμα κειμένου <**body**>. Το σώμα περιέχει δομές χαμηλότερου επιπέδου κείμενο, όπως παραγράφων <**p**>, ή διαφορετικές δομές ανάλογα με το είδος κειμένου: γραμμές για την ποίηση, διαλόγους για το δράμα.

```
<text>
   <body>
      <p>For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW). It
      was a sunny day in late September 1960 bang on noontime and Dr Burt was looking forward to a stroll in the park instead. He hoped his fellow members of theRAW
      weren't even going to notice his absence.</p>
   </body>
</text>
```

Text Encoding Initiative (TEI): Text > Front

- Δίπλα στο <**body**>, ένα κείμενο μπορεί να περιέχει προαιρετικά στην αρχή του κάτι που να κωδικοποιείται με το <**front**>, όπως π.χ. τίτλος σελίδων, κεφαλίδες, πρόλογοι, ή αφιερώσεις.
- Μία λίστα από χαρακτηριστικά
  - **preface**: a foreword or preface addressed to the reader
  - **ack**: a formal declaration of acknowledgement by the author
  - **dedication**: a formal offering or dedication of a text by the author
  - **abstract**: a summary of the content of a text as continuous prose
  - **contents**: a table of contents. A **<list>** element should be used to mark its structure
  - **frontispiece**: a pictorial frontispiece, possibly including some text

Text Encoding Initiative (TEI): Text > Front

```
<front>
   <div type="dedication">
      <p>In memory of Lisa Wheeman.</p>
   </div>
   <div type="contents">
      <head>Table of Contents</head>
      <list>
         <item>I. The Decision</item>
         <item>II. The Fuss</item>
         <item>III. The Celebration</item>
      </list>
   </div>
</front>
```

Text Encoding Initiative (TEI): Text > Back

- Όλα τα πίσω κομμάτια ενός κειμένου μπορούν να ομαδοποιηθούν στο <**back**>. Όπως συμβαίνει και με το <**front**>, είτε με αρίθμηση ή μη αριθμημένα τμήματα <**div**> με ένα χαρακτηριστικό από την ακόλουθη λίστα:
  - **appendix**: an appended self-contained section of a work, often providing additional information or text
  - **glossary**: contains a list **<list>**of terms and their explanations
  - **notes**: a section in which textual or other kinds of notes are gathered together
  - **bibliogr**: contains a list of bibliographical citations **<listBibl>**
  - **index**: any form of index to the work
  - **colophon**: a statement appearing at the end of a book describing the conditions of its physical production

```
<back>
   <div type="colophon">
      <p>Typeset in Haselfoot 37 and Henry 8. Printed and bound by Whistleshout, South Africa.</p>
   </div>
</back>
```

Text Encoding Initiative (TEI): Text (πλήρες δείγμα)

```
<text>
   <front>
      <div type="dedication">
         <p>In memory of Lisa Wheeman.</p>
      </div>
      <div type="contents">
         <head>Table of Contents</head>
         <list>
            <item>I. The Decision</item>
            <item>II. The Fuss</item>
            <item>III. The Celebration</item>
         </list>
      </div>
   </front>
   <body>
      <p>For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW). It
      was a sunny day in late September 1960 bang on noontime and Dr Burt was looking forward to a stroll in the park instead. He hoped his fellow members of the RAW
      weren't even going to notice his absence.</p>
   </body>
   <back>
      <div type="colophon">
         <p>Typeset in Haselfoot 37 and Henry 8. Printed and bound by Whistleshout, South Africa.</p>
      </div>
   </back>
</text>
```

Text Encoding Initiative (TEI): Υπόδειγμα

```
<TEI>
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>
               <!--Title-->
            </title>
         </titleStmt>
         <publicationStmt>
            <p>
               <!--Publication Information-->
            </p>
         </publicationStmt>
         <sourceDesc>
            <p>
               <!--Information about the source-->
            </p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <!--Some structural division, paragraph, line group, speech, ...-->
      </body>
   </text>
</TEI>
```

Text Encoding Initiative (TEI): Πλήρες δείγμα

```
<TEI>
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>The Strange Adventures of Dr. Burt Diddledygook: a machine-readable transcription</title>
            <respStmt>
               <resp>editor</resp>
               <name xml:id="EV">Edward Vanhoutte</name>
            </respStmt>
         </titleStmt>
         <publicationStmt>
            <p>Not for distribution.</p>
         </publicationStmt>
         <sourceDesc>
            <p>Transcribed from the diaries of the late Dr. Roy Offire.</p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <front>
         <div type="dedication">
            <p>In memory of Lisa Wheeman.</p>
         </div>
         <div type="contents">
            <head>Table of Contents</head>
            <list>
               <item>I. The Decision</item>
               <item>II. The Fuss</item>
               <item>III. The Celebration</item>
            </list>
         </div>
      </front>
      <body>
         <p>For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaan (RAW).
         It was a sunny day in late September 1960 bang on noontime and Dr Burt was looking forward to a stroll in the park instead. He hoped his fellow members of
         the RAW weren't even going to notice his absence.</p>
      </body>
      <back>
         <div type="colophon">
            <p>Typeset in Haselfoot 37 and Henry 8. Printed and bound by Whistleshout, South Africa.</p>
         </div>
      </back>
   </text>
</TEI>
```

Text Encoding Initiative (TEI): Title page

```
<front>
   <titlePage>
      <docAuthor>Roy Offire</docAuthor>
      <docTitle>
         <titlePart type="main">The Strange Adventures of Dr. Burt Diddledygook</titlePart>
         <titlePart type="sub">Wanderings in the life of a buoyant academic</titlePart>
      </docTitle>
      <byline>Transcribed from the diaries.</byline>
      <docEdition>First Edition</docEdition>
      <docImprint><pubPlace>Kirkcaldy</pubPlace>, <publisher>Bucket Books</publisher>,
         <docDate>1972</docDate>
      </docImprint>
   </titlePage>
</front>
```

Text Encoding Initiative (TEI): Διάφορα δείγματα

```
<p>For the first time in twenty-five years, <choice>
   <abbr type="title">Dr</abbr>
   <expan>Doctor</expan>
</choice> Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa ( <choice>
   <abbr type="acronym">RAW</abbr>
   <expan>Royal Academy of Whoopledywhaa</expan>
</choice>). </p>
```

```
<p>It was titled 'While thou art here', by Sir Edmund <choice>
   <corr>Peckwood</corr>
   <sic>Petwood</sic>
</choice></p>
```

```
<figure n="2">
   <figure n="2a">
      <graphic url="wtatcoverfront.jpg"/>
      <head>Front</head>
   </figure>
   <figure n="2b">
      <graphic url="wtatcoverback.jpg"/>
      <head>Back</head>
   </figure>
   <head>Figure 2:</head>
   <p>Front and back cover of the first print edition of "While thou art here" by Sir Edmund Peckwood from the rare books collection of the National Library of
   Whoopledywhaa.</p>
</figure>
```

TEI: Κλάσεις γνωρισμάτων και κοινά γνωρίσματα

- Κλάσεις γνωρισμάτων (attribute classes): στοιχεία που “μοιράζονται” κοινά γνωρίσματα, όπου τα ονόματα αυτών των κλάσεων έχουν το πρόθεμα “att”.

- Κοινά γνωρίσματα: παρέχει έναν αριθμό ή μια άλλη ετικέτα αναφοράς (label) σε ένα στοιχείο, η οποία δεν είναι υποχρεωτικά μοναδική μέσα στο TEI τεκμήριο.
- Για παράδειγμα xml:lang, υποδηλώνει τη γλώσσα του περιεχομένου του στοιχείου το οποίο συνοδεύει. Η ένδειξη κάθε γλώσσας είναι ένα σύνολο χαρακτήρων που καθορίζεται από το πρότυπο BCP 47.

XML και Document Type Definitions

- **Πλεονέκτημα της XML**: επιτρέπει να ορίσουμε και να χρησιμοποιήσουμε στοιχεία, γνωρίσματα και οντότητες της αρεσκείας μας.
- Είναι χρήσιμο να τίθενται κοινά αποδεκτοί κανόνες που προδιαγράφουν συγκεκριμένο λεξιλόγιο από επιτρεπτά ονόματα στοιχείων και γνωρισμάτων, και θέτουν περιορισμούς ως προς την πολλαπλότητα εμφάνισης των στοιχείων, την μεταξύ τους σειρά κ.λ.π.
- Για την επιβολή τέτοιων περιορισμών απαιτείται ένας τρόπος να περιγραφούν αυτοί. Αυτό μπορεί να γίνει με τη βοήθεια *Δηλώσεων Τύπου Τεκμηρίων* (DTD).

XML και Document Type Definitions

- *Δηλώσεις τύπου τεκμηρίων:* σύνολα κανόνων που ορίζουν τα στοιχεία, τα γνωρίσματα και τις οντότητες που επιτρέπεται να εμφανίζονται στα XML έγγραφα.
- Το περιεχόμενο ενός DTD παρέχει μετα πληροφορία στα *προγράμματα συντακτικής ανάλυσης* (parsers) των XML τεκμηρίων. Η πληροφορία αφορά τους περιορισμούς σύνταξης που πρέπει να πληρούν τα τεκμήρια ώστε να θεωρούνται *έγκυρα* ως προς το συγκεκριμένο DTD.
- *Έγκυρο* (valid) XML τεκμήριο: αν συνοδεύεται από ένα DTD και είναι δομημένο σύμφωνα με τους κανόνες που ορίζει το DTD.
- Ένα DTD λειτουργεί ως *γραμματική* grammar) για μια κατηγορία XML τεκμηρίων, αφού παρέχει ένα λεξιλόγιο αποδεκτά ονόματα στοιχείων και γνωρισμάτων καθώς και σύνολο από κανόνες που διέπουν τη σειρά εμφάνισης, το πλήθος των εμφανίσεων κ.λ.π. των στοιχείων σε ένα XML τεκμήριο προκειμένου αυτό να θεωρείται έγκυρο.

Γιατί ΤΕΙ;

Βιβλιογραφία

Ενδεικτικά βιβλία για Υπολογιστική Γλωσσολογία, Ανάκτηση Πληροφορίας και Ανάλυση Κειμένου

Βιβλιογραφία

- Berry D. (2012). Understanding Digital Humanities. Palgrave McMillan.
- Clark A., Fox C. & Lapin S. (2013). *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell.
- Ingersoll G., Morton T.S. & Farris S. (2013). *Taming Text: How to find, organize and manipulate it*. Manning Publications.
- Jurafsky M. & Martin J. (2006). *Speech and Language Processing*. Prentice Hall.
- Manning C.D., Raghavan P. & Schutze H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- TEI_by_example (http://www.teibyexample.org/)
- TEI Consortium (eds.). TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium: Oxford, Providence, Charlottesville, Nancy. http://www.tei-c.org/Guidelines/P5/.