Εισαγωγή στη γλώσσα XML

Μανόλης Γεργατσούλης

Χρήστος Παπαθεοδώρου

Ομάδα Βάσεων Δεδομένων και Πληροφοριακών Συστημάτων, Τμήμα Αρχειονομίας – Βιβλιοθηκονομίας

Ιόνιο Πανεπιστήμιο

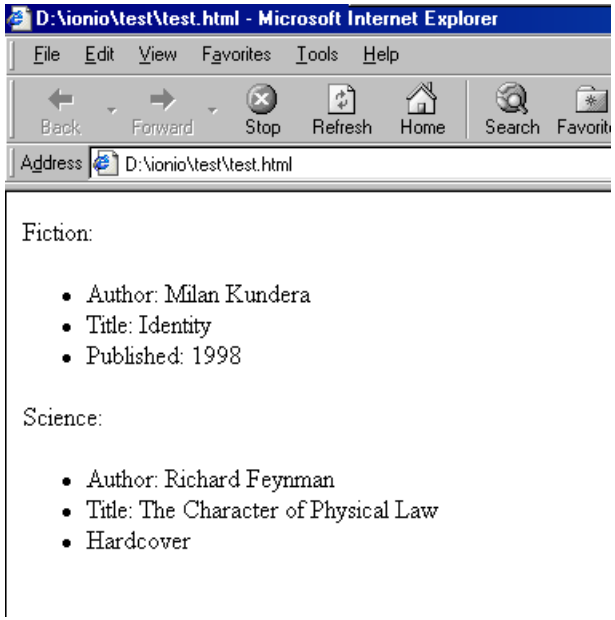
DBIS
database & information systems group
ionian university

HTML

- Απλή *γλώσσα σημειοθέτησης* (*markup language*)
- Το κείμενο εμπλουτίζεται με “εντολές” της γλώσσας οι οποίες ονομάζονται *ετικέτες* (tags), οι οποίες συνήθως αποτελούνται από μια *ετικέτα αρχής* (*start tag*) και μια *ετικέτα τέλους* (*end tag*).
- Με την HTML περιγράφουμε πως θέλουμε να παρουσιάζεται η πληροφορία ενός κειμένου.

Παράδειγμα HTML: Λίστα Βιβλίων

```
<HTML>
<BODY>
Fiction:
<UL><LI>Author: Milan Kundera</LI>
      <LI>Title: Identity</LI>
      <LI>Published: 1998</LI>
</UL>
Science:
<UL><LI>Author: Richard Feynman</LI>
      <LI>Title: The Character of Physical Law</LI>
      <LI>Hardcover</LI>
</UL>
</BODY>
</HTML>
```

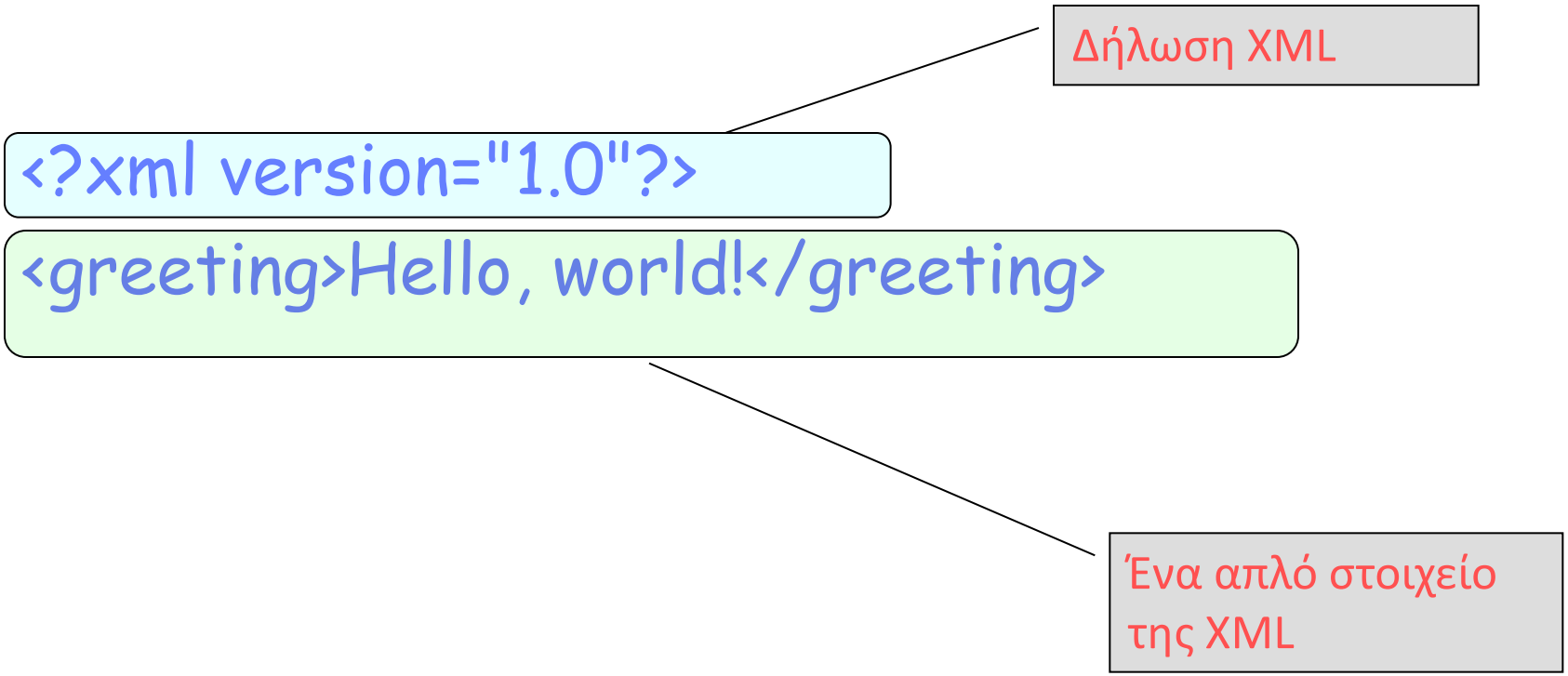


Πέρα από την HTML: XML

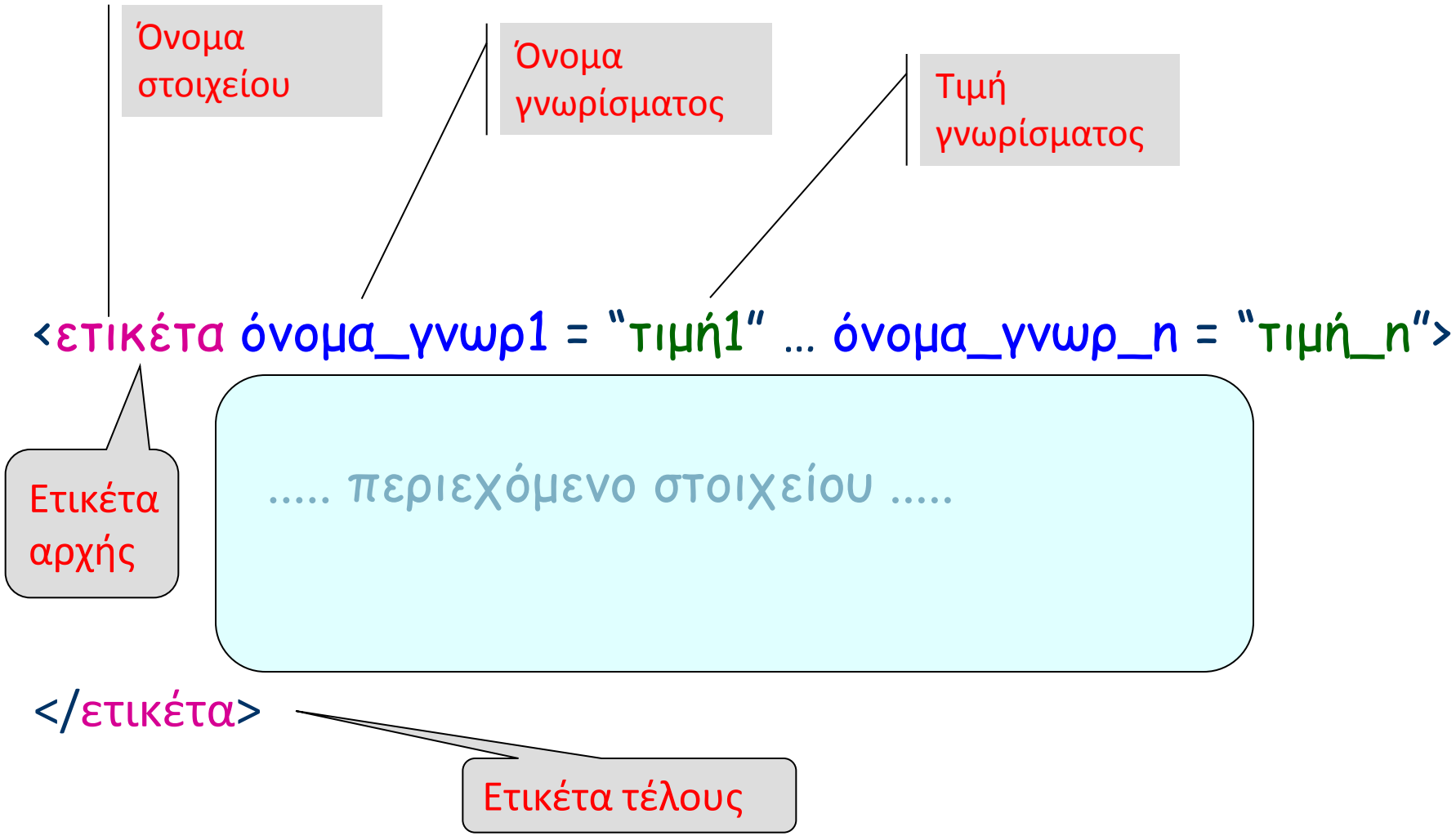
- *e**X**tensible **M**arkup **L**anguage* (XML) αποτελεί μια εξαιρετικά απλή διάλεκτο της γλώσσας *S**tandard G**eneralized M**arkup L**anguage* (SGML), η οποία αναπτύχθηκε με στόχο να διευκολύνει το χειρισμό, επεξεργασία, διακίνηση και αποθήκευση τεκμηρίων στον *Παγκόσμιο Ιστό* (web).
- Συνδυασμός SGML και HTML: 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).
- Τεχνικά εγχειρίδια:
  - «Extensible Markup Language (XML) 1.0 (Second Edition)» βρίσκεται στη διεύθυνση: *http://www.w3.org/TR/REC-xml*

Πως ξεκινά ένα XML τεκμήριο

Ένα απλό XML τεκμήριο:



XML:Στοιχεία και γνωρίσματα



Παράδειγμα XML τεκμηρίου

Βιβλιογραφία

- S. Abiteboul, P. Buneman, D. Suciu "*Data on the Web: From Relations to Semistructured Data and XML*" Morgan Kaufmann Publishers, 2000.
- Norman Walsh "*A Guide to XML*" World Wide Web Journal, Vol. 2, Issue 4, 1997, pages 97-107.

```
<bibliography>
    <book>
        <author>S. Abiteboul</author>
        <author>P. Buneman</author>
        <author>D. Suciu</author>
        <title>Data on the Web: From Relations to Semistructured Data and XML</title>
        <publisher>Morgan Kaufmann Publishers</publisher>
        <year>2000</year>
    </book>
    <article>
        <author>Norman Walsh</author>
        <title>A Guide to XML</title>
        <journal>World Wide Web Journal</journal>
        <volume>2</volume>
        <issue>4</issue>
        <year>1997</year>
        <pages>97-107</pages>
    </article>
</bibliography>
```

XML: Βασικά Δομικά Στοιχεία

- *Στοιχεία* (*elements*).
  - Οι βασικές δομικές μονάδες της XML.
  - *Ετικέτα αρχής*, *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).

Απλά και Σύνθετα Στοιχεία

- Ένα απλό στοιχείο (έχει για περιεχόμενο απλό κείμενο):

```
<φοιτητής> Νίκος  Νικολάου </φοιτητής>
```

- Ένα σύνθετο στοιχείο (περιλαμβάνει άλλα στοιχεία):

```
<φοιτητής>
   <όνομα> Νίκος </όνομα>
   <επώνυμο> Νικολάου </επώνυμο>
</φοιτητής>
```

Περιεχόμενο στοιχείων

Σύνθετα Στοιχεία με Ανάμικτο Περιεχόμενο

- Στοιχείο με ανάμικτο περιεχόμενο:

<φοιτητής>
Το όνομα του φοιτητή είναι <όνομα>Νίκος</όνομα>
ενώ το επώνυμο του είναι
<επώνυμο>Νικολάου</επώνυμο>
</φοιτητής>

Ανάμικτο περιεχόμενο

Καλά Διαμορφωμένο XML Τεκμήριο

- Για να είναι ένα XML τεκμήριο καλά διαμορφωμένο (well-formed) πρέπει να υπακούει στους κανόνες σύνταξης της XML:
  - Οι ετικέτες του τεκμηρίου πρέπει να είναι ισορροπημένες: σε κάθε ετικέτα αρχής πρέπει να αντιστοιχεί μια ετικέτα τέλους η οποία να βρίσκεται μετά την ετικέτα αρχής μέσα στο τεκμήριο.
  - Αν μια ετικέτα αρχής E1 εμφανίζεται νωρίτερα από μια ετικέτα αρχής E2, τότε η ετικέτα τέλους που αντιστοιχεί στην E1 εμφανίζεται αργότερα από την ετικέτα τέλους που αντιστοιχεί στην E2. Επομένως, οι ετικέτες τέλους πρέπει να εμφανίζονται με την ανάστροφη σειρά από αυτήν που εμφανίζονται οι αντίστοιχες ετικέτες αρχής.

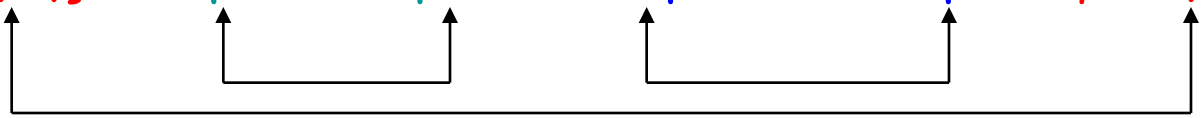
Καλά Διαμορφωμένο XML Τεκμήριο

- Καλά Διαμορφωμένο XML Τεκμήριο:

```
<φοιτητής>
    <όνομα> Νίκος </όνομα>
    <επώνυμο> Νικολάου </επώνυμο>
</φοιτητής>
```

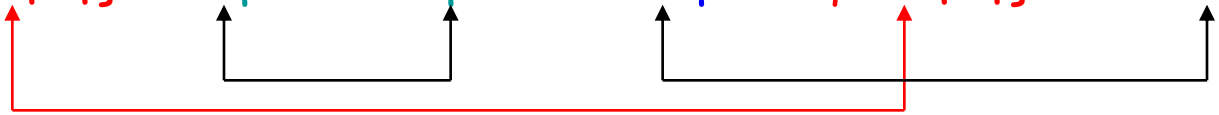
- Σειρά εμφάνισης ετικετών:

<φοιτητής><όνομα></όνομα><επώνυμο></επώνυμο></φοιτητής>



- Λανθασμένη σειρά εμφάνισης ετικετών:

<φοιτητής><όνομα></όνομα><επώνυμο></φοιτητής></επώνυμο>



Ένα μεγαλύτερο παράδειγμα XML τεκμηρίου

- Αναπαράσταση λίστας φοιτητών του ΤΑΒ σε μορφή XML τεκμηρίου:

```
<ΤΑΒ>
    <φοιτητής>
        <όνομα> Νίκος </όνομα>
        <επώνυμο> Νικολάου </επώνυμο>
    </φοιτητής>
    <φοιτητής>
        <όνομα> Πέτρος </όνομα>
        <επώνυμο> Πέτρου </επώνυμο>
    </φοιτητής>
    <φοιτητής>
        <όνομα> Μίνα </όνομα>
        <επώνυμο> Μίνου </επώνυμο>
    </φοιτητής>
    …
</ΤΑΒ>
```

Κενά στοιχεία στην XML

- Η σύνταξη της XML επιτρέπει *κενά στοιχεία* (empty elements) δηλαδή στοιχεία με κενό περιεχόμενο.
- Τα κενά στοιχεία έχουν τη μορφή:

  `<ετικέτα></ετικέτα>`
- Παρέχεται και η ακόλουθη συντομογραφία για τη σύνταξη κενών στοιχείων:

  `<ετικέτα/>`

Γνωρίσματα στην XML

- 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.
- Στο παρακάτω παράδειγμα το γνώρισμα με όνομα ΑΜ χρησιμοποιείται για να αποτυπωθεί ο αριθμός μητρώου του φοιτητή:

Γνώρισμα

```
<φοιτητής ΑΜ = "12345">
    <όνομα> Νίκος </όνομα>
    <επώνυμο> Νικολάου </επώνυμο>
</φοιτητής>
```

Γνωρίσματα στην XML (συνέχεια)

- Οι τιμές των γνωρισμάτων περικλείονται ανάμεσα σε απλά ή διπλά εισαγωγικά.
- Ένα στοιχείο είναι δυνατόν να διαθέτει περισσότερα του ενός γνωρίσματα.

```
<book isbn="1-55860-622-X"  language="English">
    <title> Data on the Web </title>
    <price currency = "USD"> 100 </price>
</book>
```

- Ενώ ένα στοιχείο μπορεί να διαθέτει υποστοιχεία με το ίδιο όνομα, δεν επιτρέπεται σε περισσότερα του ενός γνωρίσματα του να έχουν το ίδιο όνομα.
- Η σειρά με εμφάνισης των γνωρισμάτων ενός στοιχείου δεν παίζει κανένα ρόλο σε αντίθεση με τη σειρά εμφάνισης των στοιχείων που είναι σημαντική.

Σχόλια

- Τα *σχόλια* (comments) επιτρέπονται οπουδήποτε εκτός από το εσωτερικό των ετικετών.
- Ένα σχόλιο ξεκινά με το <!-- και τελειώνει με το -->.
  - Παράδειγμα:

```
<!-- Αυτό είναι ένα σχόλιο -->
```

- Τα σχόλια τοποθετούνται για να κάνουν το τεκμήριο ευανάγνωστο από τον άνθρωπο.

Οδηγίες Επεξεργασίας

- Οι *οδηγίες επεξεργασίας* (Processing Instructions) PI επιτρέπουν σε ένα XML τεκμήριο να περιέχει οδηγίες που απευθύνονται σε προγράμματα εφαρμογών.
- Μια οδηγία επεξεργασίας περιλαμβάνει το όνομα μιας εφαρμογής στην οποία απευθύνεται, ακολουθούμενο από πληροφορίες (οδηγίες επεξεργασίες, παραμέτρους κ.λ.π.) οι οποίες θέλουμε να περάσουν στην εφαρμογή.
- Παράδειγμα: Η παρακάτω οδηγία επεξεργασίας απευθύνεται στην εφαρμογή xml-stylesheet:

```
<?xml-stylesheet href="book.css" type="text/css"?>
```

Οντότητες και αναφορές οντοτήτων

- Ορισμένοι χαρακτήρες έχουν ειδική σημασία στην XML.
  - Ο χαρακτήρας < υποδηλώνει την έναρξη μιας ετικέτας ενώ ο χαρακτήρας > υποδηλώνει το τέλος της ετικέτας.
  - Οι χαρακτήρες &, ' και ", έχουν ειδική σημασία στην XML.
- Η απευθείας χρησιμοποίηση τέτοιων συμβόλων στο κείμενο που αποτελεί το περιεχόμενο ενός στοιχείου οδηγεί σε συντακτικά λανθασμένα XML τεκμήρια.
  - Παράδειγμα 1: Το παρακάτω στοιχείο είναι συντακτικά λανθασμένο:

    <στοιχείο> Το σύμβολο < δεν μπορεί να εμφανίζεται έτσι </στοιχείο>

    αφού το < στο περιεχόμενο του θα εκληφθεί ως έναρξη ετικέτας.
- Η XML παρέχει ενσωματωμένες οντότητες οι οποίες ονομάζονται *εσωτερικές οντότητες* (internal entities) για την αναπαράσταση τέτοιων συμβόλων σε ένα XML τεκμήριο αποφεύγοντας τα προβλήματα σύνταξης.

Οντότητες και αναφορές οντοτήτων (συνέχεια)

- Η τοποθέτηση τέτοιων συμβόλων σε ένα XML τεκμήριο γίνεται μέσω αναφορών στις αντίστοιχες οντότητες. Μια *αναφορά οντότητας* (entity reference) ξεκινά με το σύμβολο &, ακολουθείται από το *όνομα της οντότητας*, και τελειώνει με το σύμβολο ;.
  - Παράδειγμα 2: το < αποτελεί αναφορά στην οντότητα με όνομα lt που αναπαριστά το <, ενώ με τα >, &, ', ", αναφερόμαστε στις οντότητες που αντιστοιχούν στα >, &, ' και " αντίστοιχα.
  - Παράδειγμα 3: Το στοιχείο στο Παράδειγμα 1 πρέπει να γραφτεί σαν:

```
<στοιχείο>Το σύμβολο &lt; δεν μπορεί να εμφανίζεται έτσι</στοιχείο>
```

Οντότητες και αναφορές οντοτήτων (συνέχεια)

- Οντότητες XML μπορούν επίσης να χρησιμοποιηθούν για να αναφερθούμε σε κείμενο που επαναλαμβάνεται συχνά. Στην περίπτωση αυτή οι οντότητες παίζουν το ρόλο συντομογραφιών.
- Επίσης, οντότητες μπορούν να χρησιμοποιηθούν για να ενσωματώσουμε το περιεχόμενο εξωτερικών αρχείων.
- Οι οντότητες της κατηγορίας αυτής ονομάζονται και *εξωτερικές οντότητες* (external entities), και θα πρέπει να δηλωθούν από το χρήστη στο DTD, με τον τρόπο που θα δούμε στην αντίστοιχη ενότητα.
- Κάθε οντότητα θα πρέπει να έχει ένα μοναδικό όνομα.
- Οι οντότητες της XML μοιάζουν με τις μακροεντολές των γλωσσών προγραμματισμού.

Αναφορές χαρακτήρων

- Οι *αναφορές χαρακτήρων* (character references) έχουν παρόμοια μορφή με τις αναφορές οντότητας.
- Χρησιμοποιούνται για την εισαγωγή οποιουδήποτε χαρακτήρα του συνόλου ISO/IEC 10646 σε ένα XML τεκμήριο. Αυτό γίνεται περικλείοντας το κωδικό του χαρακτήρα ανάμεσα σε & και ;.
- Αν η αναφορά χαρακτήρα ξεκινά με &#x τότε τα ψηφία που ακολουθούν μέχρι το σύμβολο τερματισμού ; παρέχουν τη δεκαεξαδική αναπαράσταση του χαρακτήρα στο ISO/IEC 10646.
  - Παράδειγμα: ℞
- Αν όμως ξεκινά απλά με το &# τότε τα ψηφία που ακολουθούν μέχρι το σύμβολο τερματισμού ; παρέχουν τη δεκαδική αναπαράσταση του χαρακτήρα στο ISO/IEC 10646.
  - Παράδειγμα: ℞
- Μέσω των αναφορών χαρακτήρων μπορούμε να εισάγουμε χαρακτήρες οι οποίοι δεν είναι προσπελάσιμοι από τις διαθέσιμες συσκευές εισόδου.

XML: Ταυτότητες Αντικειμένων (Oids) και Αναφορές

```
<person id="o555"> <name> John </name>
                  <children idref="o123"/>
</person>
<person id="o456"> <name> Mary </name>
                   <children idref="o123"/>
</person>
<person id="o123" mother="o456" father="o555"
   >
   <name>Jim</name>
</person>
```

Τα oids και οι αναφορές στην XML είναι απλά σύνταξη

XML: Document Type Definitions (DTDs)

- **Πλεονέκτημα της XML**: επιτρέπει να ορίσουμε και να χρησιμοποιήσουμε στοιχεία, γνωρίσματα και οντότητες της αρεσκείας μας.
- Ένα έγγραφο XML είναι *καλά διαμορφωμένο* (*well-formed*) αν:
  - Το έγγραφο ξεκινά με ένα δηλωτικό XML.
  - Διαθέτει στοιχείο ρίζα που περιέχει όλα τα υπόλοιπα στοιχεία.
  - Όλα τα στοιχεία του είναι κατάλληλα εμφωλευμένα.
- 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.
- Κάθε κοινότητα χρηστών μπορεί να προδιαγράψει τη δική της XML διάλεκτο με βάση τις ανάγκες των μελών της.
- Για την επιβολή τέτοιων περιορισμών απαιτείται ένας τρόπος να περιγραφούν αυτοί. Αυτό μπορεί να γίνει με τη βοήθεια *Δηλώσεων Τύπου Τεκμηρίων* (Document Type Definitions) (DTD).
- *Δηλώσεις τύπου τεκμηρίων:* σύνολα κανόνων που ορίζουν τα στοιχεία, τα γνωρίσματα και τις οντότητες που επιτρέπεται να εμφανίζονται στα XML έγγραφα.

XML: Document Type Definitions (DTDs) (συνέχεια)

- 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.
- Το DTD από την οπτική γωνία των βάσεων δεδομένων μπορεί να εκληφθεί σαν *σχήμα* (schema) για τα δεδομένα που αναπαριστά το XML τεκμήριο, με μια σημασία παρόμοια με αυτή των σχεσιακών βάσεων δεδομένων.
- Παρόλα αυτά ένα XML τεκμήριο δεν υποχρεούται να περιλαμβάνει (ή να συνδέεται) με κάποιο DTD.

Παράδειγμα DTD

- Παράδειγμα XML τεκμηρίου που κωδικοποιεί στοιχεία φοιτητών του TAB:

```
<TAB>
    <φοιτητής>
      <όνομα> Νίκος </όνομα>
      <επώνυμο> Νικολάου </επώνυμο>
    </φοιτητής>
    <φοιτητής> ... </φοιτητής>
    ...
</TAB>
```

- Ένα DTD για το πιο πάνω τεκμήριο:

```
<!DOCTYPE TAB [
  <!ELEMENT TAB (φοιτητής*)>
  <!ELEMENT φοιτητής (όνομα, επώνυμο)>

  <!ELEMENT όνομα (#PCDATA)>
  <!ELEMENT επώνυμο (#PCDATA)>

]>
```

-Το κεντρικό στοιχείο είναι το TAB...

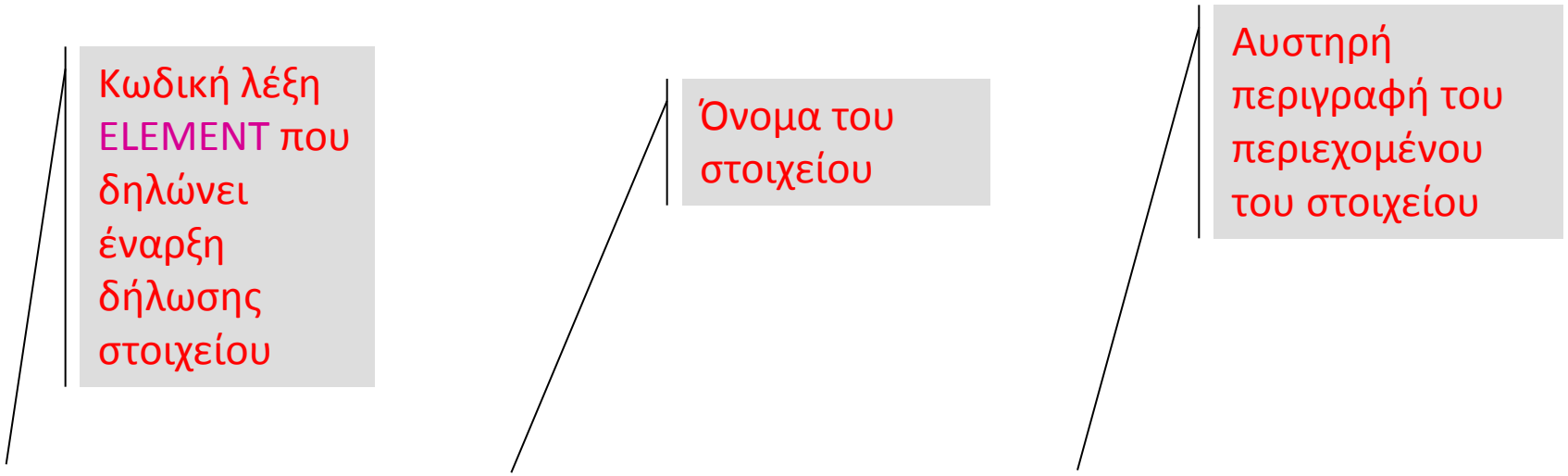
-Αποτελείται από στοιχεία φοιτητής...

-Το στοιχείο φοιτητής περιλαμβάνει τα στοιχεία όνομα και επώνυμο...

-Το όνομα περιλαμβάνει χαρακτήρες

-Το επώνυμο περιλαμβάνει χαρακτήρες....

DTD:Δηλώσεις Τύπου Στοιχείων



DTD:Δηλώσεις Τύπου Στοιχείων: Παράδειγμα

- Με την έκφραση:

<!ELEMENT s (a, b?,c*)>

δηλώνεται ότι: κάθε στοιχείο με ετικέτα **s** που εμφανίζεται σε ένα έγκυρο XML τεκμήριο, περιλαμβάνει ένα ακριβώς στοιχείο με ετικέτα a ακολουθούμενο προαιρετικά από ένα το πολύ στοιχείο με ετικέτα b, και στη συνέχεια από οσοδήποτε μεγάλο πλήθος (μπορεί και μηδέν) στοιχείων με ετικέτα c.

Δηλώσεις Τύπου Στοιχείων (συνέχεια)

- Για να δηλώσουμε ότι το περιεχόμενο ενός στοιχείου είναι ακολουθία χαρακτήρων χρησιμοποιούμε δηλώσεις της μορφής:

  <!ELEMENT *όνομα_στοιχείου* (#PCDATA)>

- Η παράσταση *τύπος_στοιχείου* είναι επίσης δυνατό να πάρει μια από τις τιμές EMPTY και ANY που σημαίνουν το *κενό στοιχείο*, και το *στοιχείο με οποιοδήποτε περιεχόμενο* αντίστοιχα.
- Αποδεκτές είναι επίσης τιμές που αποτελούν ανάμιξη #PCDATA και ονομάτων στοιχείων.
- **Προσοχή:** ένα στοιχείο δεν επιτρέπεται να δηλώνεται περισσότερο από μια φορά σε ένα DTD.

DTD:Δηλώσεις Λίστας Γνωρισμάτων

Κωδική λέξη ATTLIST που δηλώνει έναρξη δήλωσης λίστας γνωρισμάτων

Όνομα του στοιχείου

Δηλώσεις γνωρισμάτων

<!ATTLIST όνομα_στοιχείου λίστα_δηλώσεων_γνωρισμάτων>

Τριάδες της μορφής:
όνομα_γνωρίσματος τύπος_γνωρίσματος προκαθορισμός_τιμής

Δηλώσεις Λίστας Γνωρισμάτων: Παράδειγμα

<!ATTLIST φοιτητής ΑΜ CDATA #REQUIRED
ΑΔΤ CDATA #IMPLIED >

- Το στοιχείο **φοιτητής** έχει δύο γνωρίσματα με ονόματα **ΑΜ** και **ΑΔΤ**.
- Και τα δύο γνωρίσματα είναι του τύπου **CDATA**.
- Η παρουσία του γνωρίσματος **ΑΜ** είναι υποχρεωτική σε κάθε εμφάνιση του στοιχείου φοιτητής (λόγω του **#REQUIRED**).
- Η παρουσία του γνωρίσματος **ΑΔΤ** δεν είναι υποχρεωτική (λόγω του **#IMPLIED**).

Δηλώσεις Λίστας Γνωρισμάτων: Παράδειγμα

- Στη δήλωση λίστας γνωρισμάτων που ακολουθεί:

  <!ATTLIST book color (red|green|blue) "blue">

  ορίζεται ότι:
  - Το στοιχείο book έχει ένα γνώρισμα με όνομα color.
  - Το γνώρισμα αυτό μπορεί να πάρει μια από τις τιμές red, green, blue.
  - Σε περίπτωση που το γνώρισμα απουσιάζει από ένα στοιχείο book θεωρείται ως εάν να είναι παρών και η τιμή του να είναι η blue (προκαθορισμένη τιμή).

Δηλώσεις Λίστας Γνωρισμάτων (συνέχεια)

- Τιμές της παράμετρου *προκαθορισμός_τιμής:*
  - Μπορεί να πάρει σαν τιμή μια από τις πιθανές τιμές του γνωρίσματος, με τη σημασία που αναφέραμε προηγούμενα.
  - Η τιμή #REQUIRED η οποία επιβάλει την υποχρεωτική εμφάνιση του γνωρίσματος στο αντίστοιχο στοιχείο.
  - Η τιμή #IMPLIED η οποία υποδηλώνει ότι δεν παρέχεται κάποια προκαθορισμένη τιμή (και δεν είναι υποχρεωτική η εμφάνιση του συγκεκριμένου γνωρίσματος)
  - Η τιμή #FIXED ακολουθούμενη από μια συγκεκριμένη τιμή. Στην περίπτωση αυτή όλες οι εμφανίσεις του αντίστοιχου γνωρίσματος στο XML τεκμήριο θα πρέπει να έχουν για τιμή τη συγκεκριμένη τιμή που έχει δηλωθεί μετά από το #FIXED.

Δηλώσεις Λίστας Γνωρισμάτων: Παράδειγμα

- Με τη δήλωση:

  <!ATTLIST form method CDATA #FIXED "POST">

- ορίζεται ότι:
  - το στοιχείο form διαθέτει το γνώρισμα method το οποίο είναι τύπου CDATA και έχει πάντα τη τιμή POST.

Σύνδεση XML με DTD

- Προκειμένου να εξεταστεί η εγκυρότητα ενός XML τεκμηρίου ως προς ένα DTD, θα πρέπει τα δύο αυτά να συσχετιστούν.
- Γενικά έχουμε δύο επιλογές.
  - *Πρώτη επιλογή*: να συμπεριλάβουμε το DTD στο ίδιο αρχείο με αυτό που βρίσκεται το XML τεκμήριο.
  - *Δεύτερη επιλογή*: να τοποθετήσουμε τις δηλώσεις του DTD σε ξεχωριστό αρχείο και στη συνέχεια να συσχετίσουμε κατάλληλα τα δύο αρχεία.

Σύνδεση XML με DTD:Παράδειγμα 1

- DTD ενσωματωμένο στο XML τεκμήριο:

DTD

```
<?xml version="1.0"?>
<!DOCTYPE greeting [
    <!ELEMENT greeting (#PCDATA)>
]>
<greeting>Hello, world!</greeting>
```

Σύνδεση XML με DTD:Παράδειγμα 2

- Σύνδεση με εξωτερικό αρχείο στο οποίο έχει αποθηκευτεί το DTD:

Σύνδεση με DTD

```
<?xml version="1.0"?>
<!DOCTYPE greeting SYSTEM "hello.dtd">
<greeting>Hello, world!</greeting>
```

Η γλώσσα *XML Schema*

- Η γλώσσα *XML Schema* είναι μια γλώσσα XML κατάλληλη για την περιγραφή της δομής XML τεκμηρίων.
- Η XML Schema (όπως και τα DTD) είναι γλώσσα περιγραφής σχήματος.
- Η XML Schema προσφέρει χαρακτηριστικά και δυνατότητες, ισχυρότερα αυτών που παρέχονται από τα DTD.

XML Schema: Παράδειγμα (συνέχεια)

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
    <xs:element name="TAB">
      <xs:complexType>
          <xs:element name="φοιτητής" minOccurs=0
                maxOccurs="unbounded">
            <xs:complexType>
                <xs:sequence>
                    <xs:element name="όνομα" type="xs:string"/>
                    <xs:element name="επώνυμο" type="xs:string"/>
                </xs:sequence>
            </xs:complexType>
          </xs:element>
      </xs:complexType>
    </xs:element>
</xs:schema>
```

<!ELEMENT TAB (φοιτητής*)>

<!ELEMENT φοιτητής (όνομα, επώνυμο)>

<!ELEMENT όνομα (#PCDATA)>

<!ELEMENT επώνυμο (#PCDATA)>

Πρότυπα Μεταδεδομένων

Χρήστος Παπαθεοδώρου (papatheodor@ionio.gr)
Αναπληρωτής Καθηγητής
Ομάδα Βάσεων Δεδομένων και Πληροφοριακών Συστημάτων,
Τμήμα Αρχειονομίας – Βιβλιοθηκονομίας, Ιόνιο Πανεπιστήμιο
και
Μονάδα Ψηφιακής Επιμέλειας,
Ινστιτούτο Πληροφοριακών Συστημάτων και Προσομοίωσης
Ερευνητικό Κέντρο «Αθηνά»

DBIS
database & information systems group
ionian university

Περιεχόμενα

- Εισαγωγή στα μεταδεδομένα
- Dublin Core
- Κωδικοποίηση Κειμένων (Text Encoding Initiative)
- Αρχειακά μεταδεδομένα – Encoding archival description (EAD)

Ψηφιακά τεκμήρια

- Οτιδήποτε υπάρχει σε ψηφιακή μορφή και προσπελαύνεται με τη βοήθεια υπολογιστή
  - κείμενα
  - εικόνες, κινούμενες εικόνες
  - ήχος, βίντεο
  - ιστοσελίδες, προγράμματα

Τύποι δεδομένων Ηλεκτρονικών τεκμηρίων

Κείμενο

.DOC

.TXT

.RTF

.PDF

Εικόνες

.BMP

.GIF

.JPEG

.TIFF

.EPS

Κινούμενες εικόνες

.ANI

.FLI

.FLC

.GIF

Ηχος

.WAV

.MID

.SND

.AUD

Βίντεο

.AVI

.MOV

.MPG

.QT

Ιστοσελίδες

.HTM

.HTML

.XML

Προγράμματα

.COM

.EXE

Πώς εντοπίζουμε και ανακτούμε τα ψηφιακά τεκμήρια;

Μεταδεδομένα

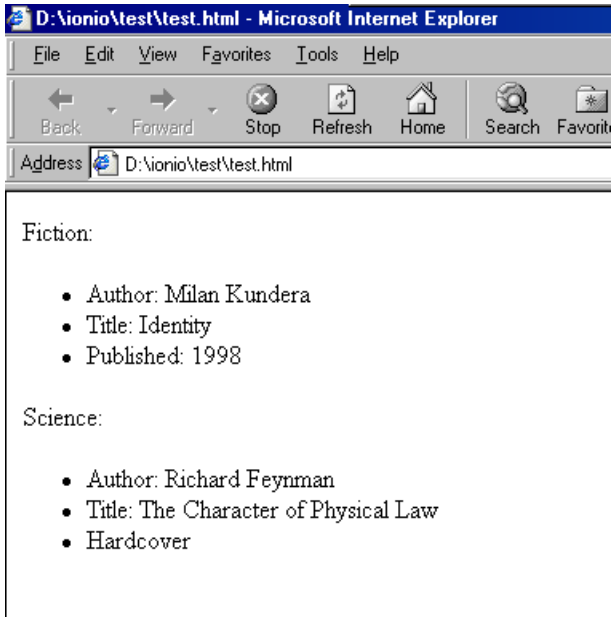
- Τα μεταδεδομένα είναι **δομημένη πληροφορία** που περιγράφει, εξηγεί, εντοπίζει ή διευκολύνει την ανάκτηση, τη χρήση ή την διαχείριση ενός ψηφιακού τεκμηρίου
- Συχνά καλούνται («**δομημένα**») «δεδομένα για άλλα δεδομένα» ή «πληροφορία για άλλη πληροφορία»
- "**Metadata** is machine understandable information about web resources or other things." (Tim Berners-Lee)
- Διάφορα πρότυπα μεταδεδομένων ανάλογα με το είδος του τεκμηρίου και τις πληροφοριακές ανάγκες
  - ανακάλυψη: περιγραφικά, π.χ. τίλος, συγγραφέας
  - διατήρηση: τεχνικά, π.χ. τύπος υλικού, ψηφιακής μορφή

Που Βρίσκονται τα Μεταδεδομένα

- Στην ετικέτα ενός CD (ορατά)
- Στη σελίδα τίλου ενός τεκμηρίου (ορατά)
- Στην κορυφή μιας ιστοσελίδας (ορατά)
- Σε διαφορετική εγγραφή, π.χ. στον κατάλογο μιας βιβλιοθήκης (ορατά)
- Στην ηλεκτρονική μορφή μέσα στον πόρο (ορατά)
- Ενσωματωμένα στο ηλεκτρονικό δημοσίευμα (μέρος της κωδικοποίησης, μη ορατά)

Παράδειγμα: Ιστοσελίδα σε HTML

```
<HTML>
<BODY>
<p> Fiction:</p>
<UL><LI>Author: Milan Kundera</LI>
      <LI>Title: Identity</LI>
      <LI>Published: 1998</LI>
</UL>
<p> Science: </p>
<UL><LI>Author: Richard Feynman</LI>
      <LI>Title: The Character of Physical Law</LI>
      <LI>Hardcover</LI>
</UL>
</BODY>
</HTML>
```



Πώς θα κάνουμε «ορατά» τα μεταδεδομένα από τους υπολογιστές (μηχαναγνώσιμη πληροφορία);

XML: κωδικοποίηση μεταδεδομένων

- Η HTML ελέγχει τον τρόπο εμφάνισης των ηλεκτρονικών τεκμηρίων, δεν ασχολείται με τη φύση τους, την περιγραφή τους, τη σημασία τους
- e**X**tensible **M**arkup **L**anguage (XML) αποτελεί μια εξαιρετικά απλή διάλεκτο της γλώσσας Standard Generalized Markup Language (SGML), η οποία αναπτύχθηκε με στόχο να διευκολύνει το χειρισμό, επεξεργασία, διακίνηση και αποθήκευση τεκμηρίων στον Παγκόσμιο Ιστό (web)

Αναγκαιότητα Προτύπων Μεταδεδομένων

- Χρειάζονται για να έχουμε **κοινή αντίληψη** των δεδομένων που περιγράφονται με αυτά
- Μας προσφέρουν μεγαλύτερη **δομή**
  - Μας περιορίζουν στην **ευελιξία**
  - Επεξεργάζονται ευκολότερα **μηχανικά**
- Είναι αναγκαιότερα σε **ψηφιακά** αντικείμενα
  - Αφού στα συμβατικά έχουμε αισθητήρια αντίληψη για αναγνώριση, θέση, θέματος, …

Πόσα Πρότυπα Μεταδεδομένων

- Δεν υπάρχει **ένα** μοναδικό διεθνές πρότυπο για μεταδεδομένα, γιατί:
  - Χρειαζόμαστε διαφορετικά **επίπεδα πολυπλοκότητας**, από πλούσιες μέχρι απλές περιγραφές
  - **Υπάρχουν** κάμποσα σχήματα μεταδεδομένων, για διαφορετικά επίπεδα και απαιτήσεις
  - **Επεκτείνουμε** τα υπάρχοντα πρότυπα

Πρότυπα Μεταδεδομένων

- AACR2
- MARC
- Text Encoding Initiative - TEI Header (1990)
- Dublin Core - DC (1995)
- Encoded Archival Description – EAD (1996)
- Open Archives Initiative - OAI
- Visual Resources Association Core– VRA (1997)
- Federal Geographic Data Committee for Digital Geospatial Metadata – FGDC
- Data Documentation Initiative – DDI (1997)
- Gateway to Educational Materials – GEM (1999)
- Government (Global) Information Locator Service - GILS
- Metadata Encoding and Transmission Standard – METS
- Metadata Object Description Schema – MODS
- Computer Interchange of Museum Information – CIMI
- Interoperability of Data in E-Commerce Systems – INDECS
- Online Information Exchange – ONIX (2000)
- Extended Markup Language – XML, MARCXML
- Australian Recordkeeping Metadata Schema (RKMS)

Κατηγορίες Μεταδεδομένων

- **Διαχειριστικά (administrative)**

Με πληροφορίες για τη διαχείριση και διατήρηση της εγγραφής, όπως δημιουργία, μετατροπή και σχέσεις με άλλες εγγραφές, π.χ. κάτοχος, ημερομηνίες δημιουργίας ή/και μεταβολής, γλώσσα εγγραφής, διαχείρισης δικαιωμάτων κλπ.

Δομικά (structural)

Με πληροφορίες για την αποθήκευση και παρουσίαση

Περιγραφικά (descriptive)

Με πληροφορίες που περιγράφουν ιδιότητες και περιεχόμενο του αντικειμένου στο οποίο αναφέρεται η εγγραφή και συμβάλλουν στην ανάκτηση (ανακάλυψη), π.χ. τίτλος, συγγραφέας, θεματικοί όροι, περίληψη

Επιλογή μεταδεδομένων

Είναι σημαντικό να καθοριστούν

- το **πεδίο εφαρμογής**
- τα **κριτήρια**
- οι **μορφές** των ψηφιακών αντικειμένων και συλλογών
- ο **βαθμός λεπτομέρειας** στην περιγραφή και παρουσίαση

προκειμένου να αξιολογηθούν τα πρότυπα μεταδεδομένων και επιλεγεί το κατάλληλο.

Ερώτηση: Ποια μπορεί να είναι τα βασικά στοιχεία που θα περιγράφουν μια δεδομένη ψηφιακή συλλογή;

Απάντηση (παράδειγμα)

Υποχρεωτικά

- Τίτλος
- Συγγραφέας
- Θέμα
- Ηλεκτρονική διεύθυνση

Επιθυμητά

- Εκδότης
- Ημερομηνία
- Κατηγορία
- Σχόλια

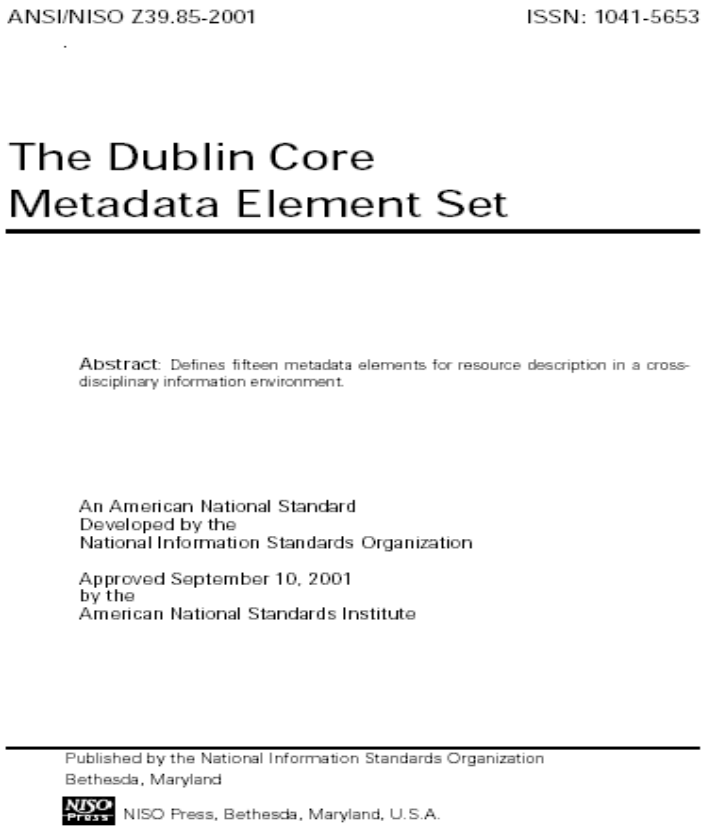
Αλλα σχετικά με συλλογή

- Αναγνωριστικά
- Γλώσσα
- Έκδοση
- Πνευματικά δικαιώματα
- ……

Dublin Core

DBIS
database & information systems group
ionian university

Το Πρότυπο «Dublin Core»



ANSI/NISO Z39.85-2001 ISSN: 1041-5653

The Dublin Core
Metadata Element Set

Abstract: Defines fifteen metadata elements for resource description in a cross-disciplinary information environment.

An American National Standard
Developed by the
National Information Standards Organization

Approved September 10, 2001
by the
American National Standards Institute

Published by the National Information Standards Organization
Bethesda, Maryland

NISO Press, Bethesda, Maryland, U.S.A.

- Προτυποποίηση
  - ISO 15836-2003
  - US: NISO Z39.85-2001
  - Ευρώπη: αναγνώριση από το CEN/ISSS Workshop Agreement 13874-2000
- Συστήνεται ανταλλαγή με RDF/XML

Χαρακτηριστικά του «Dublin Core»

- «Dublin Core» σημαίνει «Dublin, Ohio» / OCLC
- Πρωτοβουλία για να βελτιώσει την ανακάλυψη πόρων στο Διαδίκτυο
- Κοινός παρονομαστής για επικοινωνία / διαλειτουργικότητα
- Ένα υσμα για σύγκλιση των προτύπων
- Επεκτάσιμο, για να καλύψει τις επιπρόσθετες ανάγκες ανακάλυψης πόρων των διαφορετικών εφαρμογών / περιοχών

Dublin Core – Simple

- Το Dublin Core έχει 15 στοιχεία. Κάθε ένα από αυτά είναι προαιρετικό και επαναλαμβανόμενο
- Τα 15 στοιχεία χωρίζονται σε 3 κατηγορίες:
  - Περιεχόμενο:
    - Περιγράφουν το αντικείμενο
  - Πνευματική Ιδιοκτησία:
    - Περιγράφουν το copyright και τη δημιουργία
  - Στιγμιότυπο:
    - Περιγράφουν την εισαγωγή και διαχείριση

DC – Περιεχόμενο

- Τίτλος / **Title** – (ονομασία πηγής)
- Θέμα / **Subject**, π.χ. λέξεις-κλειδιά, ταξινομικοί κωδικοί
- Περιγραφή / **Description**
  - Π.χ. περίληψη, περιεχόμενα, περιγραφή εικόνας
- Πηγή (ή «Προέλευση) / **Source** – (παραγωγής)
- Γλώσσα / **Language** – (του περιεχομένου)
- Σχέση / **Relation** – (αναφορά σε σχετική πηγή)
  - Π.χ. έκδοση του …
- Κάλυψη / **Coverage** – (γεωγραφική ή χρονική)

DC – Πνευματική Ιδιοκτησία

- Δημιουργός / **Creator** – (πρόσωπο, οργανισμός, υπηρεσία)
- Εκδότης / **Publisher** – (πρόσωπο, οργανισμός, υπηρεσία)
- Συντελεστής (ή «Συνεργάτης» ή «Υπεύθυνος συμβολής») / **Contributor** – (πρόσωπο, οργανισμός, υπηρεσία που συμβάλλει στο περιεχόμενο)
  - Π.χ. μεταφραστής, εικονογράφος, κριτής
- Δικαιώματα / **Rights** – (κείμενο σχετικά με την πνευματική ιδιοκτησία)

DC – Στιγμιότυπο

- Ημερομηνία / **Date**
  - Π.χ. δημιουργίας, έκδοσης, μετάφρασης, πρόσκτησης, καταλογογράφησης, …
- Τύπος / **Type** – (κατηγορία, σχετικά με το περιεχόμενο)
  - Π.χ. ποίημα, λεξικό, software, home-page
- Μορφότυπο / **Format** – (φυσική ή ψηφιακή μορφή)
  - Π.χ. Macintosh-software, pdf, html, διαστάσεις, διάρκεια
- Αναγνωριστικό (ή Προσδιοριστής ή Κωδικός Ταύτισης) / **Identifier**
  - Μοναδικό προσδιοριστικό, π.χ. URL, ISBN, …

Στόχοι του Dublin Core

- Απλότητα δημιουργίας και διατήρησης
  - Μη ειδικοί να δημιουργούν περιγραφικές εγγραφές για αποτελεσματική ανάκτηση σε δικτυωμένο περιβάλλον
- Κοινά κατανοητή (διαθεματική) σημασιολογία
  - Σύγκλιση κοινών, περισσότερο γενικών στοιχείων
  - Αυξημένη ορατότητα και προσβασιμότητα
  - Κατάλληλο και για τον μη ειδικό της αναζητητής
    - Τον «ψηφιακό τουρίστα»

Χρήση του Dublin Core

- Είναι βασικός **πυρήνας** στοιχείων
- ***Δεν*** είναι υποκατάστατο σε πλουσιότερα περιγραφικά πρότυπα
- Παρέχει 15 «παράθυρα» ή ευρύχωρα «καλάθια» από πλουσιότερη περιγραφή πόρων
  - Φανερώνει πλούσιες περιγραφές σε απλή μορφή
  - Σημασιολογικά σταυροδρόμια, αντιστοιχίσεις σε υπάρχοντα δεδομένα

Τα Στοιχεία Μεταδεδομένων του Dublin Core

- Διεπιστημονική ομοφωνία σε απλά σύνολα στοιχείων για ανακάλυψη πόρων
  - 15 στοιχεία (πεδία μεταδεδομένων)
  - όλα προαιρετικά
  - όλα επαναλαμβανόμενα
- Δεν προορίζεται για περιγραφή περίπλοκων πόρων
  - Η αρχική ιδέα των «απλών αντικειμένων – σαν τεκμήρια»
  - Απλότητα στη σημασιολογία, ευκολία χρήσης
- Παρέχει βασική «σημασιολογική διαλειτουργικότητα»
  - Μεταξύ επιστημονικών περιοχών, μεταξύ γλωσσών
  - Δεν παρέχει λεπτομερείς κανόνες καταλογογράφησης
- Επιτρέπει επεκτασιμότητα – σε άλλες κατηγορίες πόρων

Dublin Core Qualified

DBIS
database & information systems group
ionian university

Dublin Core – Qualified

Ως επέκταση του απλού (simple ή unqualified) Dublin Core έχουμε το εξειδικευτικό (qualified) Dublin Core, που προσφέρει:

Βελτίωση της σημασιολογικής ακρίβειας του Dublin Core ορίζοντας τους **εξειδικευτές** (qualifiers)

- Εξειδίκευση του στοιχείου (element refinement)
  - Για περισσότερη λεπτομέρεια στην περιγραφή
- Σύστημα κωδικοποίησης (encoding scheme)
  - Σαν χρήση απλών κανόνων καταλογογράφησης

Ποικιλίες Εξειδικευτών: Εξειδίκευση Στοιχείων

- Προσφέρει μεγαλύτερη **λεπτομέρεια** σε αυτόν που τη χρειάζεται
- Κάνει την σημασία ενός στοιχείου στενότερη ή πιο ειδική
  - «Date *Created*» και «Date *Modified*»
  - «*IsReplacedBy* Relation» και «*Replaces* Relation»
  - Δεν την τροποποιεί ή επεκτείνει, σε καμία περίπτωση
- Αν το λογισμικό δεν καταλαβαίνει κάποιο εξειδικευτή, μπορεί με ασφάλεια να τον αγνοήσει!

Dublin Core: μη Εξειδίκευση Στοιχείων

- Creator
- Subject
- Publisher
- Contributor
- Type
- Identifier
- Source
- Language
- Rights

Dublin Core: Εξειδίκευση Στοιχείων (1/2)

Στοιχείο	Εξειδίκευση
Title	Alternative
Description	Table Of Contents Abstract
Date	Created Valid Available Issued Modified
Format	Extent Medium
Coverage	Spatial Temporal

Dublin Core: Εξειδίκευση Στοιχείων (2/2)

Στοιχείο	Εξειδίκευση
Relation	Is Version Of Has Version Is Replaced By Replaces Is Required By Requires Is Part Of Has Part Is Referenced By References Is Format Of Has Format

Κωδικοποίηση Τιμών (encoding schemes)

- Δηλώνει ότι μια τιμή είναι
  - Ένας όρος από ελεγχόμενο λεξικό (π.χ., *Library of Congress Subject Headings*)
  - Χαρακτήρες μορφοποιημένοι με συγκεκριμένο τρόπο (π.χ., «2003-05-02» σημαίνει «2 Μαΐου», όχι «5 Φεβρουαρίου»)
- Ακόμα και αν ένα σχήμα δεν είναι γνωστό από τον ερμηνευτή (λογισμικό), η τιμή πρέπει να είναι «κατάλληλη» και χρησιμοποιήσιμη για ανακάλυψη πόρων

Dublin Core: μη Κωδικοποίηση Τιμών

- Title
- Creator
- Description
- Publisher
- Contributor
- Rights

Κωδικοποίηση Τιμών (1/2)

Στοιχείο	Σχήμα κωδικοποίησης
Date	W3C-DTF DCMI
Type	DCMI Type Vocabulary
Format	IMT
Identifier	URI
Source	URI
Language	ISO 639-2 RFC 1766
Relation	URI

Κωδικοποίηση Τιμών (2/2)

Στοιχείο	Σχήμα κωδικοποίησης
Subject	LCSH MeSH DDC UDC LCC
Coverage	DCMI Point ISO 3166 DCMI Box DCMI Period W3C-DTF

Κωδικοποίησης Τιμών του «Type»

- Text (βιβλία, άρθρα, email, fax, …)
- Image (εικόνες, κινούμενες εικόνες, διαγράμματα, …)
- Sound (μουσικό CD, αρχείο μουσικής, ομιλία, …)
- Interactive Resource (υπηρεσίες συζήτησης, …)
- Service (τραπεζικές υπηρεσίες, υπηρεσία Z3950, …)
- Software (λογισμικό υπολογιστή)
- Dataset (πίνακες, βάσεις δεδομένων, …)
- Event (συνάντηση, έκθεση, συνέδριο, δίκη, γιορτή, …)
- Collection (συλλογή τεκμηρίων)

Παραδείγματα Dublin Core

DBIS
database & information systems group
ionian university

Κωδικοποίηση Dublin Core σε HTML

- Οι κανόνες σύνταξης των μεταδεδομένων υπαγορεύονται από τις γλώσσες κωδικοποίησης: HTML, XML, …
- Το πρόθεμα DC πριν από ένα στοιχείο δεδομένων προσδιορίζει ότι αυτό ορίζεται με το πρότυπο Dublin Core
- Γενικά, η σύνταξη είναι ως εξής:

<meta name = “PREFIX.ELEMENT_NAME” lang=“LANG”
scheme=“SCHEME” content = “ELEMENT_VALUE”>

- Για παράδειγμα:

<meta name = “DC.Title” content = “Το λίγο του κόσμου”>

<meta name = “DC.Creator” content = “Δημουλά, Κική”>

Παράδειγμα, HTML με DC

```
<html>
<head>
<title>Το λίγο του Κόσμου </title>
<meta name = “DC.Title”  content = “Το λίγο του κόσμου”>
<meta name = “DC.Creator” content = “Δημουλά, Κική”>
<meta name = “DC.Publisher” content = “Στιγμή”>
<meta name = “DC.Date.Issued” content = “1990”>
<meta name = “DC.Language” scheme = “ISO639-2” content = “gre”>
</head>
<body><pre>
    …Υπήρξα περίεργη και μελετηρή. …
</pre></body>
</html>
```

Dublin Core σε XML

S. Abiteboul, P. Buneman, D. Suciu "*Data on the Web: From Relations to Semistructured Data and XML*" Morgan Kaufmann Publishers, 2000.

```
<?xml version="1.0" encoding="UTF-8"?>
<book
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:isbn ="http://www.isbn.org/def">
        <dc:title> Data on the Web: From Relations to Semistructured Data and XML</dc:title>
        <dc:creator> S. Abiteboul</dc:creator>
        <dc:creator> P. Buneman</dc:creator>
        <dc:creator> D. Suciu</dc:creator>
        <dc:publisher>Morgan Kaufmann Publishers <dc:publisher>
        <dc:date>2000</dc:date>
        <dc:type>Text </dc:type>
        <isbn:number>0-201-06673-4</isbn:number>
</book>
```

Dublin Core – Επεκτασιμότητα

- Η αρχιτεκτονική του Dublin Core υποστηρίζει πιο εξελιγμένες λύσεις μεταδεδομένων
  - Αλλά προσοχή: προτιμάμε επέκταση του DC ή επιλογή άλλων πλουσιοτέρων σχημάτων;
- Έχουμε επεκτάσεις του Dublin Core σε εξειδικευμένους τομείς (domain-specific)
- Ομάδες εργασίας (του DCMI) αναπτύσσουν *Προφίλ Εφαρμογών* (*application profiles*) για ειδικούς τομείς

Προσαρμογή Μεταδεδομένων

- Τα μεταδεδομένα **προσαρμόζονται** σε πολλαπλές παραλλαγές / ποικιλίες
  - Αποτυπώνοντας την **ανομοιομορφία** των δημιουργών και διατηρητών μεταδεδομένων
  - Προσφέροντας για κάθε **κοινότητα** (συγκεκριμένο χώρο) εξειδικευμένη λειτουργικότητα, δημιουργία, διαχείριση, πρόσβαση, …
- Τηρώντας τα αντίστοιχα **πρότυπα** …
- Μεγαλύτερη εξειδίκευση = μεγαλύτερη λειτουργικότητα = λιγότερη διαλειτουργικότητα!

Τι είναι το Προφίλ Εφαρμογής

- Ένα σχήμα μεταδεδομένων που ενσωματώνει ένα **σύνολο στοιχείων** από ένα ή περισσότερα **σύνολα στοιχείων μεταδεδομένων** (ή λεξιλογίων ή χώρους ονομάτων – namespaces)
- Ένα σύνολο από **πολιτικές** που ορίζουν πώς τα στοιχεία πρέπει να εφαρμόζονται στο πεδίο της εφαρμογής
- Ένα σύνολο **οδηγιών** που ξεκαθαρίζουν τις πολιτικές που αφορούν τα στοιχεία

Κατάλογος υπερσυνδέσμων

- http://www.getty.edu/research/institute/standards/intrometadata/2_articles/index.html
- Metadata Demystified: A guide for publishers. NISO Press, July 2003. http://www.niso.org/standards/resources/Metadata_Demystified.pdf
- Arms, Caroline R. “Some Observations on Metadata and Digital Libraries”. In: *Conference on Bibliographic Control for the New Millennium,* September 2000. http://lcweb.loc.gov/catdir/bibcontrol/arms_paper.html
- 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice-Metadata Research and Application, 28 Sept. – 2 Oct. 2003, Washington USA. http://www.ischool.washington.edu/dc2003/index.html
- Guidance on the Structure, Content, and Application of Metadata Records for Digital Resources and Collections. Report of the IFLA Cataloguing Section Working Group on the Use of Metadata Schemas. Draft – for Worldwide Review 27 October, 2003. http://www.ifla.org/VII/s13/guide/metaguide03.pdf
- Dublin Core Metadata Initiative (DCMI). http://dublincore.org/
- The Dublin Core Metadata Element Set. ANSI/NISO Z39.85-2001. http://www.niso.org/standards/resources/Z39-85.pdf
- Metadata Encoding and Transmission Standard http://www.loc.gov/standards/mets/
- Metadata Object Description Schema" (MODS) http://www.loc.gov/standards/mods/
- MARC Standards Website. http://lcweb.loc.gov/marc/

Κατάλογος υπερσυνδέσμων

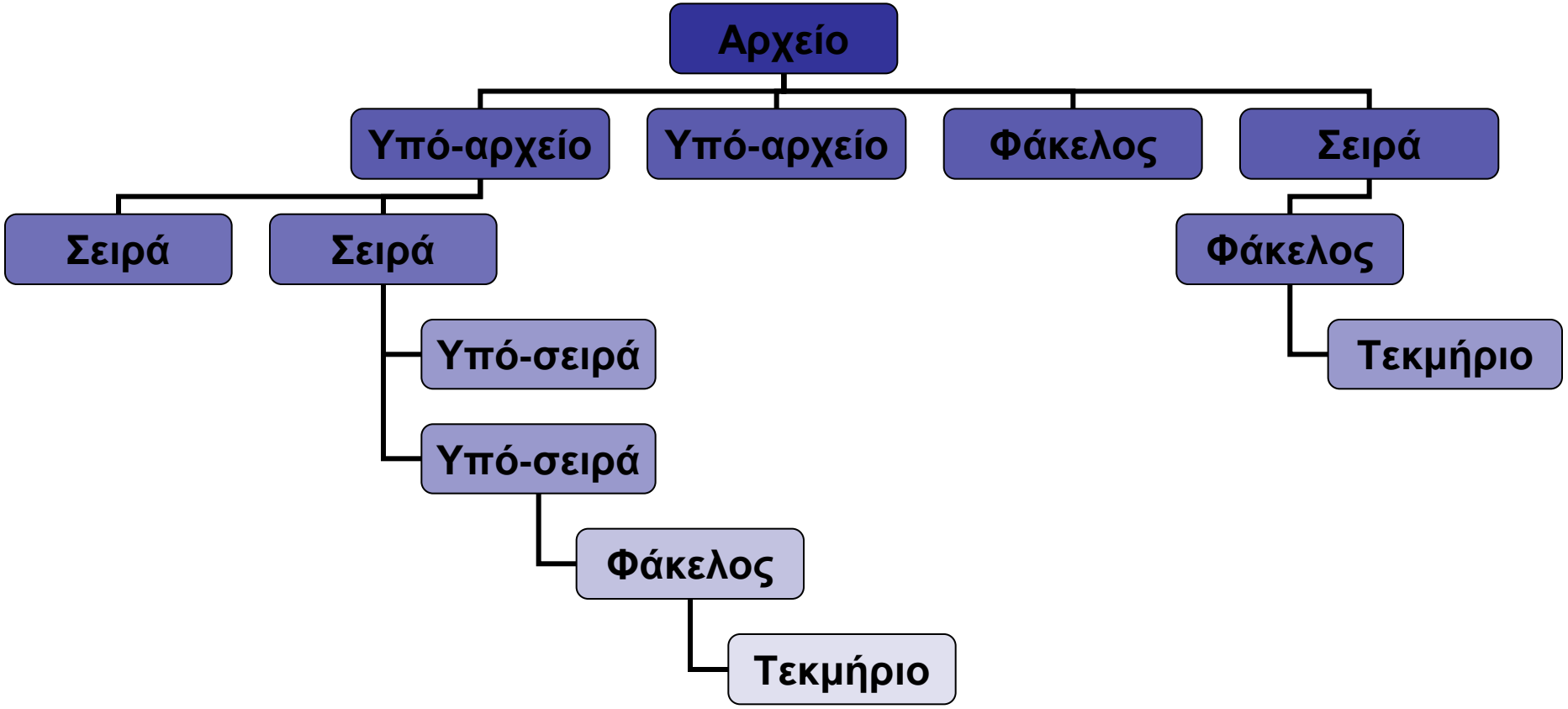
- OCLC, Bibliographic formats and standards. http://www.oclc.org/bibformats/
- UKOLN's metadata site http://www.ukoln.ac.uk/metadata/
- IFLA http://ifla.inist.fr/II/metadata.htm
- IFLA Functional Requirements for Bibliographic Records. www.ifla.org/VII/s13/frbr/frbr.pdf
- "The Value of Metadata". http://www.fgdc.gov/publications/documents/metadata/metabroc.html

Encoded Archival Description (EAD)

DBIS
database & information systems group
ionian university

Αρχειακά μεταδεδομένα

Αρχείο είναι το σύνολο των τεκμηρίων ανεξαρτήτως χρονολογίας, σχήματος και ύλης που έχει δεχθεί ή παραγάγει οποιοδήποτε φυσικό ή νομικό πρόσωπο, οποιοσδήποτε οργανισμός, δημόσιος ή ιδιωτικός, στα πλαίσια των δραστηριοτήτων του.


Αρχείο
Υπό-αρχείο
Υπό-αρχείο
Φάκελος
Σειρά
Σειρά
Σειρά
Φάκελος
Υπό-σειρά
Τεκμήριο
Υπό-σειρά
Φάκελος
Τεκμήριο

Εργαλεία έρευνας (Finding aids)

- Τελικό προϊόν της αρχειακής περιγραφής
- Πληροφοριακό εργαλείο το οποίο περιλαμβάνει μεταδεδομένα που εξυπηρετούν την αναζήτηση, διαχείριση και ερμηνεία ενός αρχείου και εξηγούν το περιβάλλον δημιουργίας του

Ξεκινώντας...

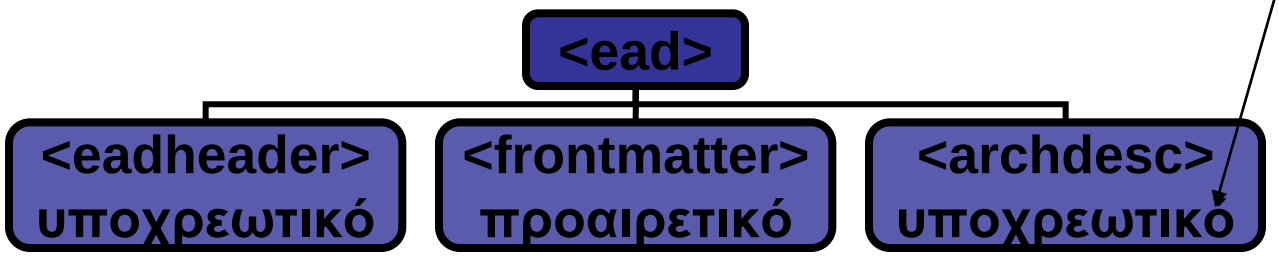
- Απαραίτητες πληροφορίες
  - Τίτλος
  - Όνομα δημιουργού και σκοπός δημιουργίας
  - Ημερομηνίες
  - Μοναδικός αριθμός ταυτοποίησης
  - Διαδικασία πρόσκτησης
  - Περιορισμοί πρόσβασης και / ή αναπαραγωγής
  - Υλικό που έχει αποκοπεί από το κύριο μέρος

EAD

- Περιγραφικά μεταδεδομένα
- Έμπνευση: Text Encoding Initiative (TEI, 1987)
- Πρώτη προσπάθεια: Πανεπιστήμιο Μπέρκλεϋ της Καλιφόρνια (1993)
- SGML/XML DTD
- EAD version 1.0: 1998
- EAD version 2002: η νέα έκδοση
- Η ανάπτυξη του EAD ξεκίνησε από τις ΗΠΑ
- Γρήγορα υπήρξε ενδιαφέρον από τη παγκόσμια αρχειακή κοινότητα
- Society of American Archivists (http://www.archivists.org/) και Library of Congress (http://www.loc.gov/ead/)

<ead>

Περιλαμβάνει επισκόπηση του αρχείου (<did>), διαχειριστικές και συμπληρωματικές πληροφορίες (π.χ. <bioghist>, <accessrestrict> κτλ), περιγραφή συστατικών μερών (<dsc>)



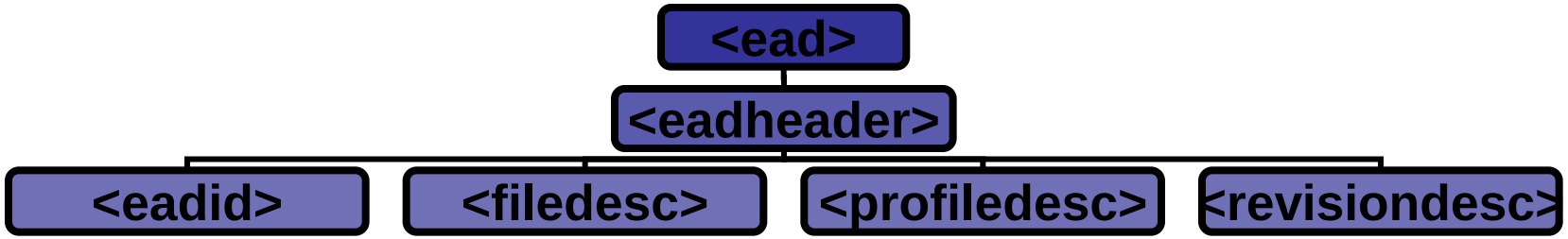
Πληροφορίες για το εργαλείο έρευνας **και όχι για το αρχείο.** Μεταδεδομένα ηλεκτρονικού εργαλείου έρευνας

Χρησιμοποιείται για την παροχή πληροφοριών έκδοσης, όπως της σελίδας τίτλου του έντυπου εργαλείου έρευνας ή άλλων προκαταρτικών πληροφοριών

<eadheader>

- Υποχρεωτικό στοιχείο
- Βασισμένο στο TEI header
- Πληροφορίες για το εργαλείο έρευνας **και όχι για το αρχείο**
- Μεταδεδομένα ηλεκτρονικού εργαλείου έρευνας
- «Ηλεκτρονική σελίδα τίτλου»

<eadheader>



<eadheader>

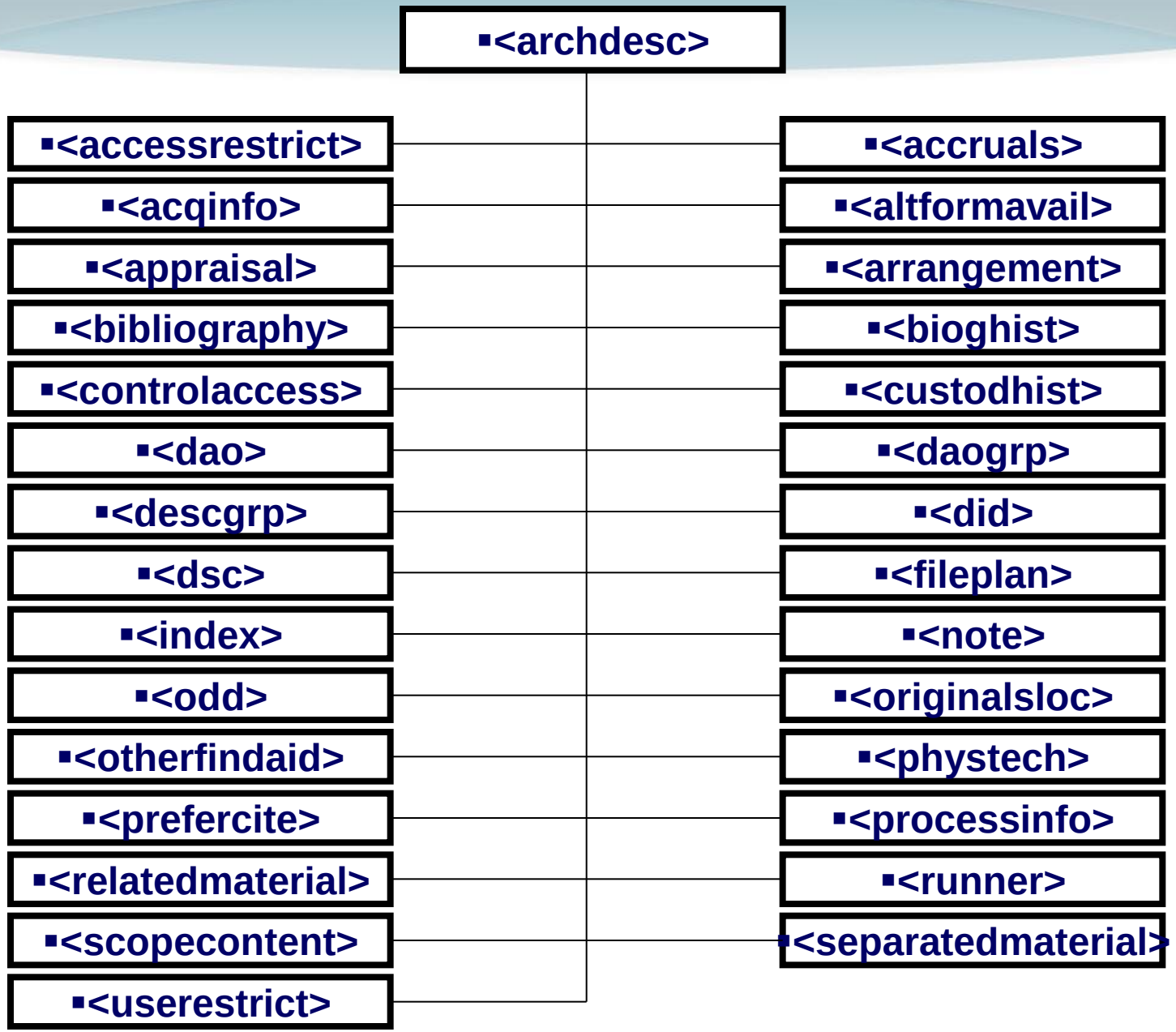
- <eadid>
  - Υποχρεωτικό στοιχείο
  - Κωδικός ταύτισης του εργαλείου έρευνας
- <filedesc>
  - Υποχρεωτικό στοιχείο
  - Βασικές βιβλιογραφικές πληροφορίες για το εργαλείο έρευνας
  - <titlestmt>, <editionstmt>, <publicationstmt>, <seriesstmt>, <notestmt>

<eadheader>

- <profiledesc>
  - Πληροφορίες για την κωδικοποίηση του εργαλείου έρευνας
  - <creation>, <langusage>, <descrules>
- <revisiondesc>
  - Πληροφορίες για αλλαγές στο εργαλείο έρευνας
  - <list>, <change>

<archdesc>

- Υποχρεωτικό στοιχείο
- Εργαλείο έρευνας
- Περιλαμβάνει
  - Επισκόπηση του αρχείου (<did>)
  - Διαχειριστικές και συμπληρωματικές πληροφορίες (π.χ. <bioghist>, <accessrestrict> κτλ)
  - Περιγραφή συστατικών μερών (<dsc>)


▪<archdesc>
▪<accessrestrict>
▪<acqinfo>
▪<appraisal>
▪<bibliography>
▪<controlaccess>
▪<dao>
▪<descgrp>
▪<dsc>
▪<index>
▪<odd>
▪<otherfindaid>
▪<prefercite>
▪<relatedmaterial>
▪<scopecontent>
▪<userestrict>
▪<accruals>
▪<altformavail>
▪<arrangement>
▪<bioghist>
▪<custodhist>
▪<daogrp>
▪<did>
▪<fileplan>
▪<note>
▪<originalsloc>
▪<phystech>
▪<processinfo>
▪<runner>
▪<separatedmaterial>

<archdesc>

- <did>
  - Υποχρεωτικό στοιχείο
  - Εμφανίζεται και μέσα στην περιγραφή των συστατικών μερών (<c> και <c01> - <c12>)
  - Βασικές πληροφορίες για το σύνολο του αρχείου (υψηλότερο επίπεδο)
  - <unittitle>, <origination>, <langmaterial>, <physdesc>, <unitdate>, <unitid>, <abstract>, <physloc>, <repository> ...

<archdesc>

- <accessrestrict>
  - Διαθεσιμότητα του αρχείου
- <acqinfo>
  - Πηγή πρόσκτησης και συνθήκες πρόσκτησης
- <bioghist>
  - Βιογραφία ή διοικητική ιστορία δημιουργού
- <controlaccess>
  - Σημεία πρόσβασης
  - <corpname>, <famname>, <geogname>, <persname>, <subject>, <title>, <genreform> ...

<archdesc>

- <scopecontent>
  - Επισκόπηση περιεχομένου του αρχείου (σε μορφή κειμένου)
- <separatedmaterial>
  - Υλικό από την ίδια πηγή που έχει αποκοπεί φυσικά ή εκκαθαριστεί από το αρχείο
- <userestrict>
  - Πληροφορίες για τις συνθήκες που επηρεάζουν τη χρήση του αρχείου εφόσον έχει επιτραπεί η πρόσβαση

<archdesc>

- <dsc>
  - Προαιρετικό στοιχείο
  - Αναλυτική περιγραφή του αρχείου
  - Περιγραφή από το γενικό στο ειδικό
  - Συστατικά μέρη: <c> ή <c01> - <c12>

<archdesc>

- Για τη σύνδεση με ένα ή περισσότερα ψηφιακά αντικείμενα
  - <daogrp> Digital Archival Object Group
    - <daodesc> Digital Archival Object Description
    - <daoloc> Digital Archival Object Location

```
<daogrp>
        <daodesc>
          <head>Image Sampler</head>
          <p>Explanatory paragraph</p>
        </daodesc>
        <daoloc href="//images/Baker109.jpeg">
          <daodesc>
            <p>Ella Baker, head-and-shoulders portrait</p>
          </daodesc>
        </daoloc>
</daogrp>
```

EAD – Γνωρίσματα

- Πρόσθετες πληροφορίες για τα στοιχεία
- Μεταδεδομένα για τα στοιχεία
- Επεξεργασία από λογισμικό
- «Κωδικοποιημένα δεδομένα»
- Τιμές
  - Σταθερές
  - Προτεινόμενες
  - Ανεξάρτητες

Γνωρίσματα

- level
  - Ιεραρχικό επίπεδο του τεκμηρίου που περιγράφεται
  - <archdesc> (Υποχρεωτικό), <c> και <c01> - <c12>
  - Τιμές: collection, fonds, class, recordgrp, series, subfonds, subgrp, subseries, file, item, otherlevel
  - π.χ. <archdesc level="fonds">

    <c level="item">

    <c02 level="subseries">

Γνωρίσματα

- type
  - Διαθέσιμο για αρκετά στοιχεία
    - <unitdate> (bulk, inclusive, other)
    - <archdesc> (inventory, register, other)
    - <dsc> (analyticover, combined, in-depth, othertype)

Γνωρίσματα

- audience (external, internal)
- countrycode (π.χ. GR)
- datechar (π.χ. creation, accumulation, ή modification)
- findaidstatus (π.χ. “edited-full-draft”)
- relatedencoding (π.χ. MARC 21)
- encodinganalog (π.χ. 500)
- langcode (π.χ. gre)
- mainagencycode (π.χ. ELIA)
- normal (π.χ. 1756/1868)

Παράδειγμα

```
<archdesc level="fonds">
    <did>
        <unittitle>Αρχείο Γραμματείας / Υπουργείου επί των Οικονομικών </unittitle>
        <unitid countrycode="GR" repositorycode="GAK">256.12</unitid>
        <unitdate normal="1833/1862">1833-1862</unitdate>
        <langmaterial>
            <language langcode="fre">Κυρίως γαλλικά</language>
            <language langcode="gre">και δευτερευόντως
         Ελληνικά </language>
        </langmaterial>
        <origination>
            <corpname>Ανάκτορα, Γραμματεία / Υπουργείο επί των
        Οικονομικών</corpname>
        </origination>
    </did>
```

<accessrestrict> Δεν υπάρχουν περιορισμοί πρόσβασης </accessrestrict>

<acqinfo> Η πρόσκτηση από τα Γ.Α.Κ. πολλών αρχείων του δημόσιου τομέα του 19ου αιώνα ήταν ιδιαίτερα δυσχερής... </acqinfo>

<bioghist> Στις 25 Ιανουαρίου / 1 Φεβρουαρίου 1833 καθιερώθηκε, με διάταγμα της Αντιβασιλείας, οι διευθύνοντες στα Υπουργεία να φέρουν τον τίτλο του «Γραμματέως της Επικρατείας»... </bioghist>

<controlaccess>

<controlaccess>

<subject>Ελληνική Επανάσταση</subject>

<subject>Οικονομικά</subject>

</controlaccess>

</controlaccess>

<scopecontent> Το αρχείο της Γραμματείας διαρθρώνεται ως εξής:...

</scopecontent>

```
<dsc>
    <c01 level="subfonds">
        <did>
            <unittitle>Υπο-αρχείο Ανακτόρων</unittitle>
            <unitdate>1833-1844</unitdate>
            <origination>
                <corpname>Ανάκτορα</corpname>
            </origination>
        </did>
        <c02 level="series">
            <did>
                <unittitle>Δάση</unittitle>
                <unitdate>1833-1844</unitdate>
            </did>
            <c03 level="file">…</c03>
        </c02>
    </c01>
</dsc>
```

Χρήσιμες ηλεκτρονικές πηγές

- Encoded Archival Description (EAD) - Official EAD Version 2002 Web Site. (http://www.loc.gov/ead/)
- EAD Help Pages (http://jefferson.village.virginia.edu/ead/)
- Encoded Archival Description - An Introduction and Overview (http://www.dlib.org/dlib/november99/11pitti.html)
- THE EAD COOKBOOK 2002 (http://www.iath.virginia.edu/ead/ead2002cookbookhelp.html)
- Recommended Best Practices for Encoded Archival Description Finding Aids at the Library of Congress (http://www.loc.gov/ead/practices/lcp2002.html)

Χρήσιμες ηλεκτρονικές πηγές

- RLG Best Practice Guidelines for Encoded Archival Description (http://www.rlg.org/en/pdfs/bpg.pdf)
- RLG EAD Report Card (http://www.rlg.org/ead-report-card/)
- Encoded Archival Context (http://www.iath.virginia.edu/eac/)
- ISAD (G) (http://www.ica.org/biblio/cds/isad_g_2e.pdf)
- ΔΙΠΑΠ (Γ) (http://www.eae.org.gr/dipap.pdf)

Text Encoding Initiative (TEI)

DBIS
database & information systems group
ionian university

Ορισμοί

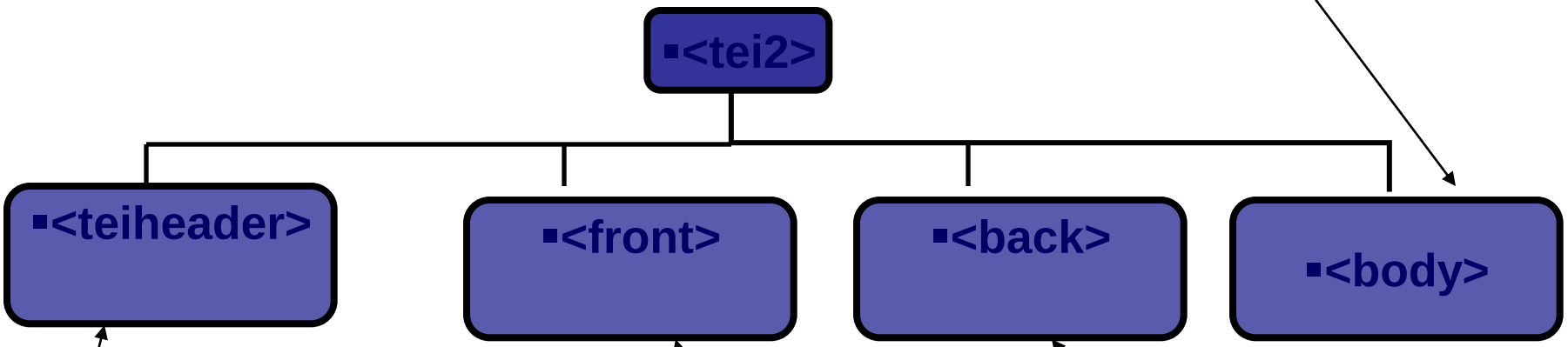
- Κωδικοποίηση κειμένου (text markup, encoding) = Διαδικασία διάκρισης δομικών ή σημασιολογικών (semantic) χαρακτηριστικών κειμένου με βάση κάποιους κανόνες.
- Text encoding initiative: SGML-DTD
- Στόχος του είναι να δημιουργήσει ένα περιβάλλον για την κωδικοποίηση κειμένων ακαδημαϊκού ενδιαφέροντος, έτσι ώστε να μπορούν να μεταγράφονται και να διατηρούνται ανεξάρτητα από την εκάστοτε τεχνολογία.

Ανασκόπηση

- 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.)

Η δομή του TEI

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



- Δεν αποτελεί μέρος του υπό κωδικοποίηση κειμένου, αλλά παρέχει πληροφορίες (μεταδεδομένα) για αυτό

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

- Περιέχει παραρτήματα κ.λπ. που ακολουθούν το κυρίως κείμενο

Βασική δομή

```
<?xml version="1.0"?>
<!DOCTYPE TEI.2 SYSTEM "http://faculty-web.at.northwestern.edu/
    english/mmueller/TeiXBaby/TeiXBaby.dtd">
<!ELEMENT TEI.2 (teiHeader, text)>
<!ELEMENT text (front?, body, back?)>
<TEI.2>
    <teiHeader> [ TEI Header information ] </teiHeader>
        <text>
            <front> [ front matter ... ] </front>
            <body> [ body of text ... ] </body>
            <back> [ back matter ... ] </back>
        </text>
</TEI.2>
```

Βασικά στοιχεία

- <teiHeader>

  Δεν αποτελεί μέρος του υπο κωδικοποίηση κειμένου, αλλά παρέχει πληροφορίες (μεταδεδομένα) για αυτό.
  - Στοιχεία του teiHeader element: **fileDesc, profileDesc, revisionDesc**, **langUsage language**
  - Στοιχεία του fileDesc: **titleStmt, publicationStmt, sourceDesc**
- <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

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

1. Το body ενός κειμένου χωρίζεται από elements
2. Τα elements χωρίζονται σε <p> (παράγραφος), <q> (εδάφιο με εισαγωγικά), <l> (γραμμή), <lg> (ομάδα γραμμών), <sp> (λόγος) και <speaker> (ομιλητής)
3. Τα <p> και παρόμοια στοιχεία περιλαμβάνουν κείμενο (#PCDATA), το οποίο κωδικοποιείται από στοιχεία φράσεων

Στοιχεία δομής

- div: υποδιαίρεση μέχρι 7 επίπεδα, εφαρμόζεται και στα front, back. Γνωρίσματα:
  - type: ‘Book’, ‘Chapter’, ‘Part’, κ.λπ.
  - id: μοναδικός κωδικός υποδιαίρεσης
  - n: όνομα ή αριθμός υποδιαίρεσης
- head: ο τίτλος της υποδιαίρεσης, <!ELEMENT head #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>
```

Γνωρίσματα

- Στο TEI-DTD υπάρχουν τα ακόλουθα γνωρίσματα που εφαρμόζονται γενικά στα στοιχεία:

<!ATTLIST element
id ID #IMPLIED
n CDATA #IMPLIED
lang IDREF #IMPLIED (γλώσσα)
rend CDATA #IMPLIED (τυπογραφική αναπαράσταση π.χ. <q lang="FR" rend="italics" >

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 (1/2)

- <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 (2/2)

- <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 1980</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.