NTUA

Τεχνολογία Πολυμέσων

5. Διάλεξη 5: XML

XML

- Μεταγλώσσα για την κωδικοποίηση δεδομένων
- Πρόβλημα που επιζητά λύσεις:
  - Κοινή γλώσσα επικοινωνίας των εφαρμογών
- Σημαίνει eXtensible Markup Language
- Σχεδιάστηκε για τη μεταφορά και την αποθήκευση δεδομένων
- Σχεδιάστηκε για να είναι αναγνώσιμη τόσο από ανθρώπους όσο και από μηχανές
- Όσο καλύτερη η κωδικοποίηση τόσο μεγαλύτερες ευκαιρίες επαν-επεξεργασίας και διαχείρισης

Προσπάθειες

- Πρώτη γενναία προσπάθεια: Η SGML- Standard Generalized Markup Language
- Η γλώσσα αυτή είχε τα γενικά χαρακτηριστικά που απαιτούνταν αλλά ήταν **πολύπλοκη** στη χρήση
- Γέννησε δύο άλλες γλώσσες, την HTML και την XML.

Τι είναι η XML;

- Ακόμη, είναι επεκτάσιμη—extensible και παραμετροποίησιμη.
  - Μπορεί δηλαδή ο προγραμματιστής να καθορίσει τα δικά του στοιχεία, να καθορίσει τη δομή του περιεχομένου της κωδικοποίησης.
  - Για παράδειγμα, άλλα στοιχεία θα έχει ένας κώδικας xml που θα περιγράφει τη δομή ενός κειμένου χημείας και άλλα ένας κώδικας που περιγράφει τα χαρακτηριστικά των αποθηκευμένων βιβλίων.
- Markup- ή καλύτερα metamarkup language.
  - Στον ίδιο τον κώδικα περιγράφεται η δομή και το περιεχόμενο των δεδομένων.
  - Περιγραφή των σχέσεων μεταξύ των δεδομένων.

Τι είναι η XML;

- Τι δεν είναι η XML- δεν κάνει τα πάντα.
- Δεν περιέχονται στοιχεία παρουσίασης.
- Δεν είναι μια γλώσσα προγραμματισμού.
- Δεν είναι ούτε πρωτόκολλο μεταφοράς δεδομένων, αλλά ούτε βάση δεδομένων.
- Αλλά, τι προσφέρει;
  - Διαχειρίσιμα δεδομένα—portable data

Διαχείριση δεδομένων…;

- Ένας web browser
- Ένας επεξεργαστής-διορθωτής κειμένου,
- Μία βάση δεδομένων που αποθηκεύει αρχεία δεδομένων (Microsoft SQL Server stores xml data in a new record)
- Ένα σχεδιαστικό πρόγραμμα,
- Ένα πρόγραμμα που «βλέπει» τον κώδικα σαν μια μορφή οικονομικού κειμένου
- Ένα πρόγραμμα διαμοίρασης πληροφοριών που παίρνει τις νέες πληροφορίες
- Ένα πρόγραμμα που έχουμε γράψει μόνοι μας (σε Java, C, Python) και κάνει ότι του πούμε εμείς στα δεδομένα,
- Οτιδήποτε άλλο.

Ελαστικότητα

- Συμπέρασμα: Η XML είναι μια εξαιρετικά ελαστική δομή κωδικοποίησης δεδομένων. "XML is an *extremely* flexible format for data".(Harold and Means, 2002, σελ.8)
- Στην πραγματικότητα, η XML είναι κατάλληλη για την αποθήκευση και ανταλλαγή κάθε είδους δεδομένων που μπορεί να κωδικοποιηθεί σε μορφή κειμένου.
- Είναι ακατάλληλη για την κωδικοποίηση δεδομένων πολυμέσων, όπως οι φωτογραφίες, ο ήχος, το βίντεο και άλλες μορφές πολύπλοκης παράστασης δεδομένων.
- Επιτρέπει την επαναχρησιμοποίηση ή επανα-οργάνωση των δεδομένων.

HTML

- Χρησιμοποιεί **tags** ανάμεσα στο κείμενο...
  - ...για να περιγράψει το **layout** της σελίδας

```
<p> <b>Alan</b>, 42 years, <i>agb@abc.com</i>
```

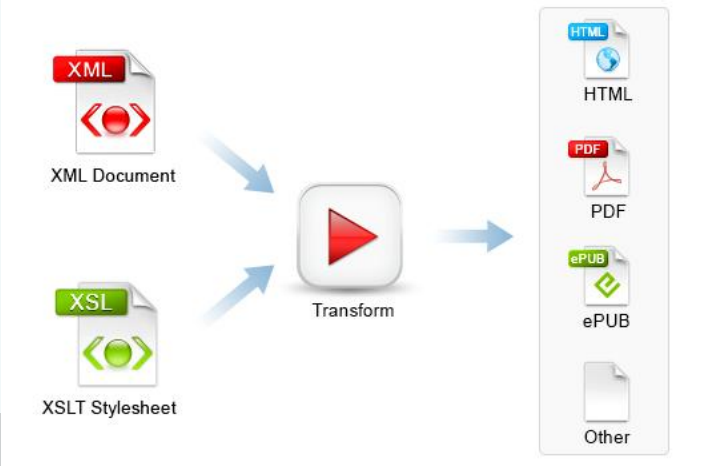
- **Δεν** διευκολύνει άλλα προγράμματα να κατανοήσουν την δομή και το περιεχόμενο μιας σελίδας
  - Ο wrapper «σπάει» αν το italic <i> αλλάξει σε teletype <t>
- Το πρόβλημα είναι οτι σχεδιάστηκε ειδικά για να περιγράψει την **παρουσίαση** και όχι το περιεχόμενο
  - Καθένας θα ήθελε να υπάρχει «ένα ακόμη tag» στην HTML προκειμένου να βοηθηθεί η δική του εφαρμογή

XML

- Σχεδιάστηκε ειδικά για να περιγράψει το **περιεχόμενο** (content) και όχι την παρουσίαση μιας σελίδας
- Βασικές διαφορές από την HTML
  - Μπορεί κανείς να ορίσει **νέα tags** κατά βούληση
  - Τα tags μπορούν να εφωλιασθούν δομικά σε οποιοδήποτε βάθος
  - Ενα έγγραφο XML μπορεί προαιρετικά να περιέχει μια περιγραφή της **γραμματικής** του
- Τα tags δομούν το περιεχόμενο

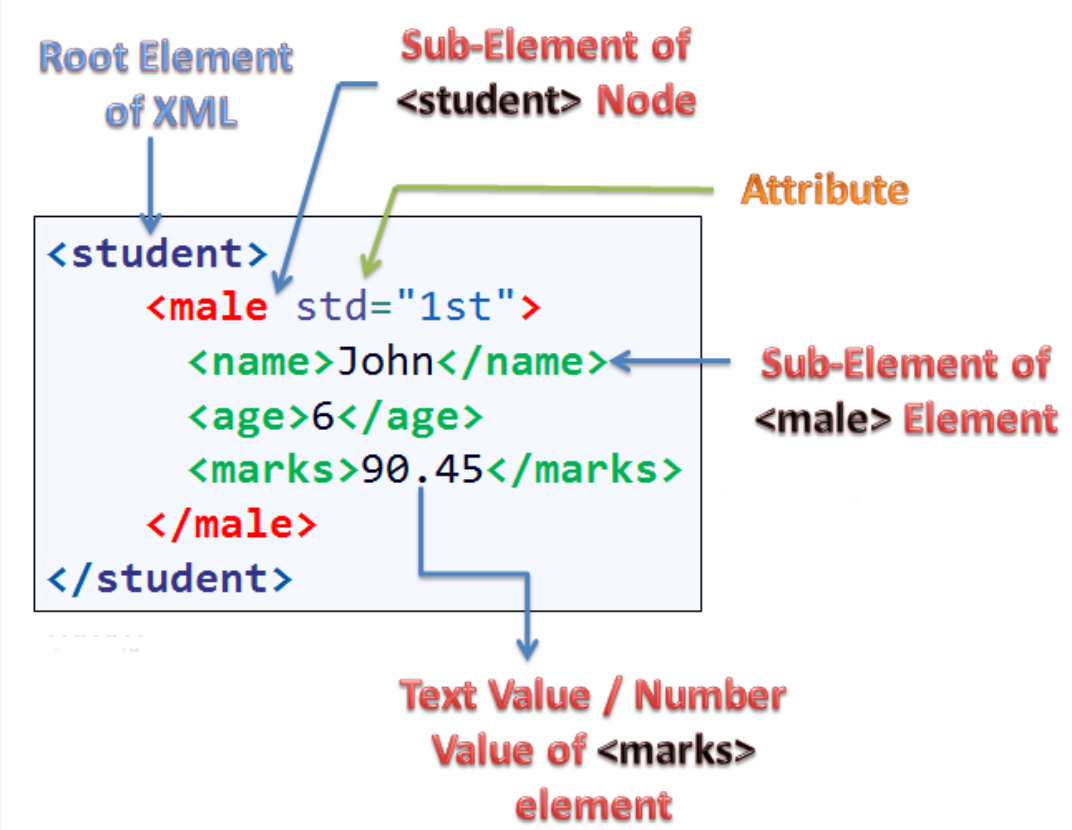
```
<person><name>Kostas</name>…</person>
```

  - Το πως θα εμφανιστούν ορίζεται ξεχωριστά από κάποιο **stylesheet** (XSL)



XML (2)

- Βασικό συστατικό της XML είναι το **element**
  - Κείμενο που περικλείεται από ένα ζεύγος tags
  - Start-tag και end-tag (**markups**)



Elements και tags

- Τι μπορεί να υπάρχει ανάμεσα στα start-tag και end-tag;
  - Απλό κείμενο
  - Αλλα elements
  - Οποιοδήποτε μίγμα των δύο παραπάνω !
- Τα tags στην XML:
  - Ορίζονται από τους χρήστες, **δεν** υπάρχουν προκαθορισμένα tags όπως στην HTML
  - Ανοίγουν και κλείνουν πάντα με την «σωστή» σειρά (σαν παρενθέσεις)
  - Εξαίρεση: **empty tag**, πχ. `<married/>` (*το / στο τέλος*)

```
<married/> = <married></married>
```

- Element, element **content**, και **subelement**

XML attributes

- Ενα element μπορεί να περιέχει μηδέν ή περισσότερα **attributes**
  - Δεν σχετίζονται με τα attributes στις σχεσιακές βάσεις
  - Περιγράφουν ιδιότητες (**properties**) του element

```
<product>
   <name language="French">trompette no 6</name>
   <price currency="Euro"> 420.12 </price>
   <address format="XL1245" language="French">
      <street>31 rue Croix-Bosset</street>
      <zip>92874</zip>
   </address>
</product>
```

- Οπως και τα elements, τα attributes ορίζονται από τον **χρήστη**

Attributes και elements

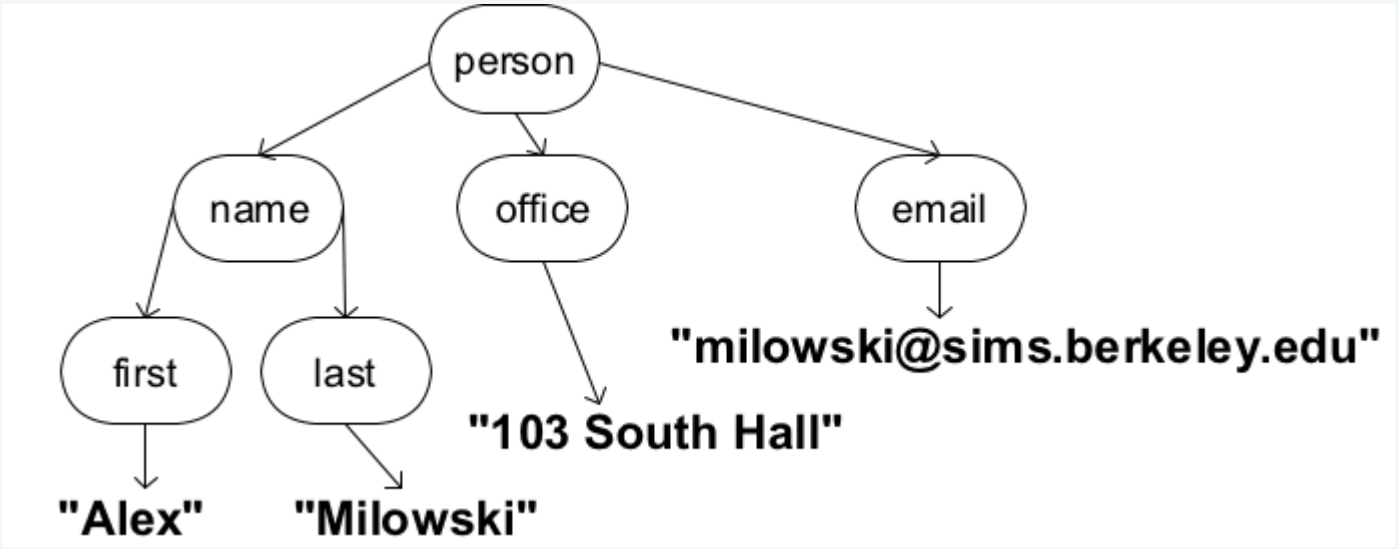
- **Διαφορές** με elements:
  - Η τιμή του attribute είναι πάντα ένα string σε εισαγωγικά, ενώ το element μπορεί να περιέχει άλλα elements
  - Ενα element μπορεί να έχει το πολύ ένα attribute με ένα όνομα, ενώ μπορεί να έχει πολλά subelements με το ίδιο όνομα
- Τα attributes:
  - Φανερώνουν την καταγωγή της XML σαν document markup γλώσσας
  - Εισάγουν κάποια **δυσκολία** στην ανταλλαγή πληροφορίας: **αναπαράσταση σαν attribute ή σαν subelement;**

```
<person name="Alan" age="42" email="agb@abc.com"/>

<person age="42">
   <name>Alan</name>
   <email>agb@abc.com</email>
</person>
```

XML Graph

- Ενα μοντέλο για XML data: XML Graph
  - Οι κόμβοι αντιστοιχούν στα elements



- Εύκολη η μετατροπή, ιδίως αν έχουμε δένδρο
  - Μπορούμε να έχουμε γράφο; - **ΝΑΙ**

XML References

- Η XML έχει έναν μηχανισμό ώστε ένα element να μπορεί να «δείχνεται» από περισσότερους από έναν «πατέρες»
  - Ορίζοντας έτσι **γράφο** αντί για δένδρο
- Ο μηχανισμός είναι attributes τύπου **ID**, **IDREF**, και **IDREFS**
  - Τα ID αναθέτουν ταυτότητες στα elements
  - Τα IDREF «δείχνουν» στα ID από άλλα elements, **οπουδήποτε**

```
<state id="s2" idrefs="c2 c7 c12 c32">Nevada</state>
...
<city id="c7">
   <cname>Carson City</cname>
   <state-of idref="s2"/>
</city>
```

- **Ποιά** attributes είναι τύπου ID / IDREF;
  - Πάντως όχι κατ' ανάγκη αυτά που λέγονται id / idref
  - Αλλά αυτά που ορίζονται από το **DTD** (στην συνέχεια)

Διάταξη

- Τα δύο XML **δεν** είναι ισοδύναμα:

```
<person><firstname>John</firstname>
        <lastname>Smith</lastname></person>
<person><lastname>Smith</lastname>
        <firstname>John</firstname></person>
```

- Τα XML attributes **δεν** είναι διατεταγμένα
  - Τα παρακάτω είναι **ισοδύναμα**:

```
<person firstname="John" lastname="Smith"/>
<person lastname="Smith" firstname="John"/>
```

- Η απαίτηση για διάταξη **δυσκολεύει** την αποδοτική διαχείριση των δεδομένων XML
  - Συχνά η διάταξη αγνοείται σε εφαρμογές ανταλλαγής πληροφορίας

Μίξη elements και κειμένου

- Το παρακάτω επιτρέπεται στην XML
  - Λέμε ότι το person έχει **mixed content**

```
<person>
   This is my best friend
   <name>Alan</name>
   <age>42</age>
   I am not too sure about the following email
   <email>agb@abc.com</email>
</person>
```

- Δείχνει την εγγραφο-κεντρική καταγωγή της XML
  - Κάπως «αφύσικο» από την πλευρά των βάσεων δεδομένων
  - Προτιμούμε το **element content**

Επιπλέον στοιχεία της XML

- Εγγραφο-κεντρικά στοιχεία που **δεν** χρειάζονται στην ανταλλαγή δεδομένων
- Σχόλια

```
<!– this is comment -->
```

- Processing Instructions

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

- Αρχική γραμμή

```
<?xml version="1.0"?>
```

- Προαιρετικό Document Type Definition (DTD)
  - Ορίζει την **γραμματική** του κειμένου

```
<?xml version="1.0"?>
<!DOCTYPE name [markupdeclarations]>
<name>…</name>
```

← root element

DTD

- **Document Type Definition** (**DTD**): αναπόσπαστο μέρος της XML
  - Προτάθηκε σαν μια γραμματική για τα XML έγγραφα
  - Σε κάποιο βαθμό μπορεί να ιδωθεί **σαν σχήμα** για δεδομένα μορφοποιημένα σε XML
- Ενα DTD που περιγράφει δυαδικά δένδρα:

```
<!ELEMENT node (leaf | (node, node))>
<!ELEMENT leaf (#PCDATA)>
```

- Τα DTDs μοιάζουν να ορίζουν **τύπους** δεδομένων

```
<!DOCTYPE db [
  <!ELEMENT db (person*)>
  <!ELEMENT person (name,age,email)>
  <!ELEMENT name (#PCDATA)>
...]>
```

- PCDATA - Parsed Character Data
- CDATA - (Unparsed) Character Data

Δήλωση attributes στο DTD

- Σύνδεση των attributes με τα elements στα οποία εμφανίζονται

```
<product>
  <name language="French"
        department="Music">trompette</name>
  <price currency="Euro"> 420.12  </price>
</product>

<!ATTLIST name language   CDATA #REQUIRED
               department CDATA #IMPLIED>
<!ATTLIST price currency  CDATA #IMPLIED>
```

- Ορισμός των attributes
  - #REQUIRED υποχρεωτικό, #IMPLIED προαιρετικό
  - Τύπος CDATA = string
  - Τύποι ID, IDREF, IDREFS

Well-formed και valid

- **Well-formed** XML έγγραφα
  - Τα tags πρέπει να είναι σωστά εμφωλιασμένα
  - Τα attributes ενός element πρέπει να είναι μοναδικά
- **Valid** XML έγγραφο
  - Είναι well-formed
  - Εχει κάποιο DTD
  - Συμμορφώνεται με αυτό το DTD
- **Περιορισμοί** του DTD σαν σχήμα για δεδομένα XML
  - Δεν υπάρχουν ατομικοί τύποι (πχ. integer)
  - Δεν υπάρχουν περιορισμοί διαστήματος (πχ. 0-140)
  - Ο τύπος ενος element είναι global (πχ. ίδιο name και σε person και σε course;)
  - Δεν προσδιορίζει τον τυπο των IDREFs