Markup Languages, XML and Text Encoding

DiXiT workshop, 24-28 April 2017

Overview

- Background

- Machine Readable Texts

- Text Representation

- Structures and Models

- XML

Background

3

Publishing and XML 1:

Markup as a series of formatting commands

```
.center; .bd
```
Chapter 1
```
.sk;  .in 5
```
This is my paragraph. With a
```
.it  word
```
in italic.
```
.sk;  .in 5
```
…

Chapter 1

    This is my paragraph. With a *word* in italic.

    …

Publishing and XML 2:

Markup as semantic description

```
<head>
Chapter 1
</head>
<p>
This is my paragraph. With an    <emph>word</emph>
in italic.
</p>
<p>...</p>
```

Chapter 1

    This is my paragraph. With a *word* in italic.

OR

This is my paragraph. With a w o r d in italic.

XML (and its less popular parent, SGML)

1986: ISO 8879:1986 Information processing – Text and office systems – Standard Generalized Markup Language (SGML) !!

- Developed primarily for government and corporate materials online and in print
- 1990 WWW based a simple form of SGML.

1998: XML version 1.0 released by the World Wide Web Consortium

- Simplified and more extensible than SGML

The Academy: "Machine-Readable Texts"

Regular list of texts, noting features encoded, and physical format.

Each text used different conventions, and often required specific hardware to process

'Literary Materials in Machine-Readable Form." *Computers and the Humanities* 2, 3 (1968): 133-44.

Casal, Julián del, *Hojas al viento, Nieve, Rimas,* (critical editions with studies of the variants; edited by R. J. Glickman) . Source text identified by author; title; poem, story, or page number; and line number. Titles, subtitles, dedications, chapter headings, and paragraph or stanza numbers are indicated.
Character set: 60; BCD; diacritics: acute accent, umlaut, tilde; punctuation: , . : ; ' " - ¿ ? ¡ ! () []; symbols: *; type differentiation: texts are printed in upper- and lower-case characters. Record size: 132 char., unblocked; density: 556; channels: 7; labeling: non-standard (first record is a header label) . Detailed documentation available. For further information, see under *Dario, Rubén.*
Communicate with Prof. Robert Jay Glickman, Italian and Hispanic Studies, University of Toronto, Toronto 5, Canada.

Caterina da Siena, S., *Libro della divina dottrina*; date, edition, page, line numbers, parts, chapters, paragraphs indicated.
Coding: see under *Alberti, Leon Battista.*
Communicate with Aldo Duro, Accademia della Crusca, Piazza dei Giudici, 1, Firenze, Italy.

Cato the Elder: Fragments: *Orations* (Malcovati, Oratorum Romanorum Fragments) , *Remainder* (Jordan) , *De Agri Cultura* (Mazzarino) . Chapter and fragment numbers enclosed between #'s; chapter titles included as in text.
Punctuation: complete as in text, except for ", % is used; symbols: critical symbols used in text; type differentiation: capitals indicated by $ prefix. Text currently on 8-channel paper tape only (other forms will be made available as soon as practical) . The text of the fragments has been completed; work is progressing on the *De Agri Cultura.*
Communicate with Stephen V. F. Waite, Department of Classics, Dartmouth College, Hanover, N. H. 03755.

7

Computing Humanists and XML

- *Machine readable texts*

- *Electronic text*

- *Computer Corpora*

- *Digital Editions*

- Humanists used descriptive markup as early as 1987

- They contributed significantly to the development of SGML and XML

- Scholarly needs and requirements seemed obscure, but served as models

?Why use XML

- Captures semantic distinctions (not appearance)

- Modeling and computability

- Single input, multiple outputs

- (made for) Electronic publishing

- Interchange and Re-usability

- Sustainability

- Community of peers

- Generalized tools

Structure

Structure is a way of organizing things so that it is possible to:

- Identify them

- Count them

- See what is missing

- Classify them

- Compare them

- Talk about them

Structuring: Turning a Text into Information

Texts contain and display implicit structure(s)

Explicit structure is a way to:

- Identify
- Locate
- Analyze
- Test

Modeling

Adding structure to information is a way of modeling it

Creating a model and then applying it to a document or the features of a document

- allows you to see how it compares with other, similar documents
- Allows you to test your model, and see if it is an accurate abstraction, and therefore useful for further analysis

Structuring Documents with XML

XML models documents as a tree - a set of elements that can contain other elements.
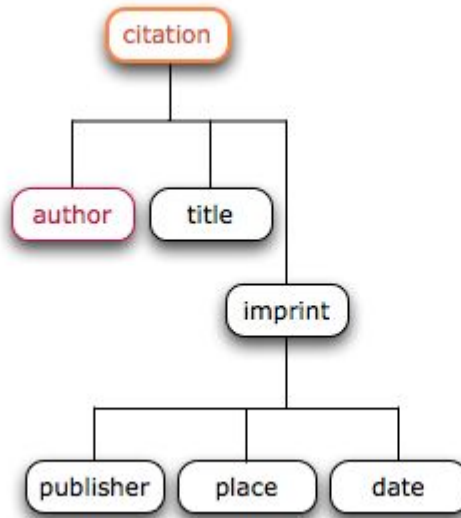
Flanders and Bauman 2016. http://www.wwp.northeastern.edu/outreach/seminars/_current/presentations/xml_intro/xml_newIntro_tutorial_02.xhtml

XML Syntax

XML Notation and Syntax

XML is not in itself an encoding language like TEI and HTML

XML provides the components—notation, grammar and syntax—used to define and describe encoding languages.

XML is a *metalanguage*

Text Sample

Από::Τσίρκας Στρατής, Πρός::Παπαϊωάννου Μ.Μ., 1955-10-19

Αλεξ. 19 Οκτ. 1955

Αγαπητοί μου Τάσο και Μιχάλη,

Γιατί δεν μου γράψατε; Σας στέλνω το 7ο κεφάλαιο και περιμένω τις

παρατηρήσεις σας πριν το στείλω στην

Επιθ. Τέχνης για το τεύχος του Νοέμβρη.

Τι λέτε;

Σήμερα έδοσα ένα αντίγραφο στο Μαλάνο. Ο θεός βοηθός!

Έφτασα στο 10ο κεφάλαιο.

Μένουν άλλα 10 τουλάχιστο.

Τα «συμπεράσματα» θα τα γράψω στο τέλος ή και καθόλου.

Περιμένω ανυπόμονα τη γνώμη σας και … σας φιλώ

Σ.Τσίρκας

Elements—Στοιχεία

An element (στοιχείο) surrounds some text, and consists of a start tag (ετικέτα αρχής) and an end tag (ετικέτα τέλους).

Σήμερα έδοσα ένα αντίγραφο στο

<name>Μαλάνο</name>.

Containment

Elements may nest, but not may not overlap.

<line><sentence>Σας στέλνω το 7ο κεφάλαιο και περιμένω τις παρατηρήσεις σας πριν το στείλω στην**</line>**
<line>Επιθ. Τέχνης για το τεύχος του Νοέμβρη.**</sentence><line>**
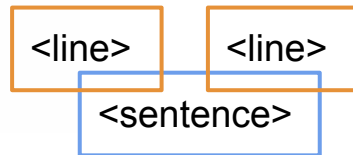
Containment

\<sentence>\<line>Σας στέλνω το 7ο κεφάλαιο και περιμένω τις παρατηρήσεις σας πριν το στείλω στην**\</line>**
\<line>Επιθ. Τέχνης για το τεύχος του Νοέμβρη.**\<line>\</sentence>**

<sentence>

<line>	<line>
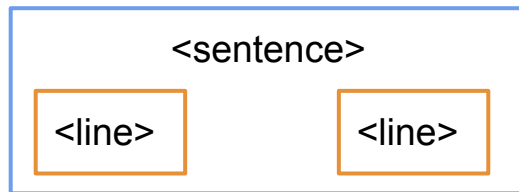
Empty Elements- κενά στοιχεία

If an element (στοιχείο) has no content (?), it may be displayed using the following shorthand:

=

Certain elements never have content as they are used to mark a point in the text, and not a span of text.

Attributes—ιδιότητες

Start tags (ετικέτες αρχής) may have one or more attributes (ιδιότητες) which
provide information about the element or its content.

Σήμερα έδοσα ένα αντίγραφο στο

<name type="person">Μαλάνο</name>.

Attribute values may come from a required list, a recommended list or be
completely open. Some attributes may have more than

Attribute Values— αξίες ιδιοτήτων ??

Attribute values (γρεεκ) may come from

- A closed list of values
- A list of suggested/recommended values
- An open list

Some attributes may have more than one value. Values are separated by a space.

Attribute values consist of a alphanumeric characters and symbols. **No spaces.**

```
<p xml:lang="grc heb">...
```

`xml:id` Attribute

The `xml:id` attribute is a special attribute used to identify an element.

All elements can have an xml:id attribute.

By definition, an `xml:id` is

- unique within a file
- can have no spaces
- must start with a letter.

`<p xml:id="ch1.s4.p1">quam</p>`

Ετικέτα αρχής — start tag

Ετικέτα τέλους — end tag

`<greeting xml:lang="en">Hello, world!</greeting>`

element name — Όνομα στοιχείου

attribute name — Όνομα ιδιότητας

attribute value — Αξία ιδιότητας

content

element name — Όνομα στοιχείου

25
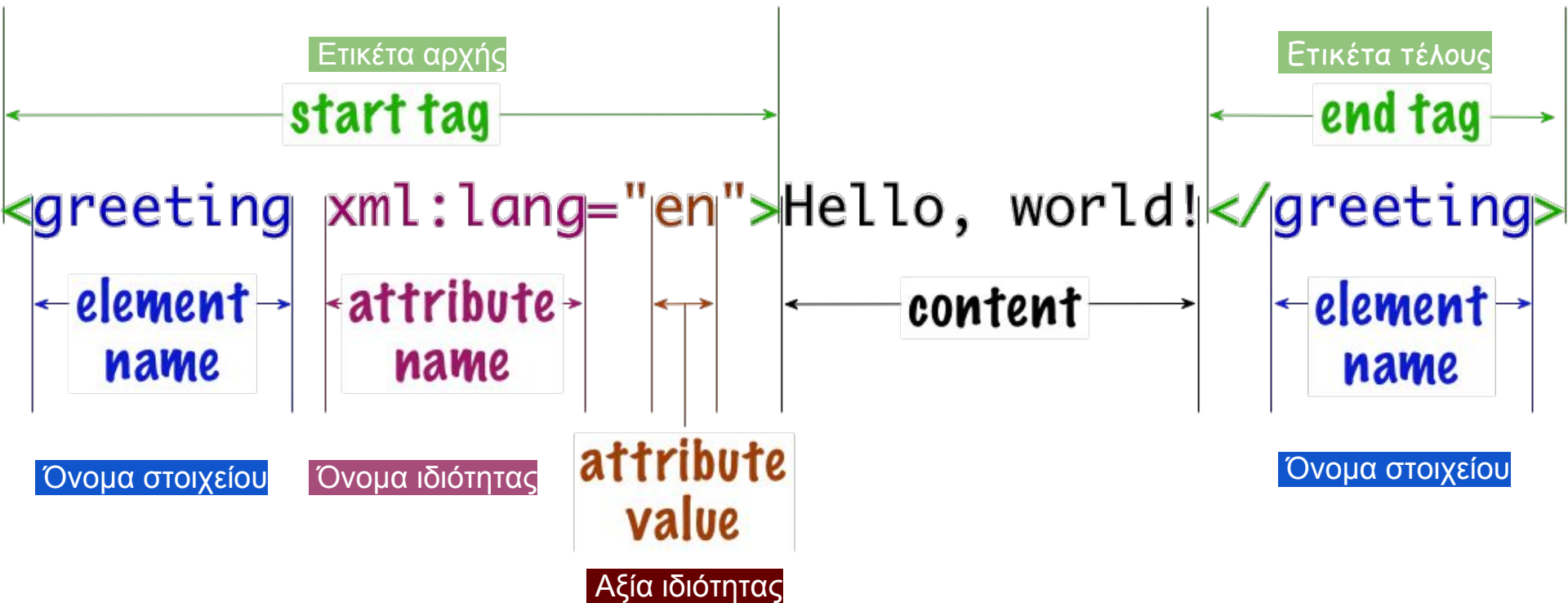
Well Formedness—ορθή μορφοποίηση

When in a document

- There are no missing $<$ $>$ $/$ **"** in tags and around attribute values
- Elements have matching begin and end tags (or are empty)
- All elements nest properly, with no overlap
- There is a single element that contains all other elements (a root element)

then it is considered to be well-formed (ορθά μορφοποιημένο)

XML documents that are not well-formed are incorrect.

Text Sample

```xml
<text>
  <body>
    <div type="letter">
      <opener>
        <dateline>21/6/55</dateline>
        <salute>Αγαπητέ μου <persName>Στρατή</persName></salute>
      </opener>
      <p> Αν έχω τα <bibl type="book"><title>«Γράμματα»</title></bibl> θα ετοιμάσω
μια συλλογή των έργων του <persName><surname>Σκληρού</surname></persName>.
Υπάρχουν κληρονόμοι;</p>
      <p> Έγιναν αυτού συζητήσεις γύρω από το σοσιαλισμό, όπως εδώ στο
«Νουμά»;</p>
      <p>Για την έκδοση εκλογής από το έργο σου: Στη μελέτη μου που θα τη
συνοδεύσει μπορώ να μιλήσω άνετα ή η θέση σου εμποδίζει;</p>
      <p> Ευχαριστώ για τα βιβλία κλπ. Θα σου στείλω κ εγώ βιβλία, χρήσιμα
ίσως.</p>
      <p>Ο κ. <persName><surname>Αλιθέρσης</surname></persName> είχε
γράψει ιστορία της λογοτεχνίας. Υπάρχει αντίτυπο;</p>
      <p>Θα σου ξαναγράψω περισσότερα. Ο <persName><surname>Βουρνάς
</surname></persName>είναι ενθουσιασμένος. </p>
      <closer><signed> Μ.Π.</signed></closer>
    </div>
  </body>
</text>
```

Schema

The XML schema is a set of rules that defines the names of the elements and the relationships in which they can appear.

- A file that has correct XML syntax is well-formed.
- A file that is correct according to a schema is valid.

Note: Element and attribute names and attribute values in a schema do not have real semantics, as far as the XML software is concerned.

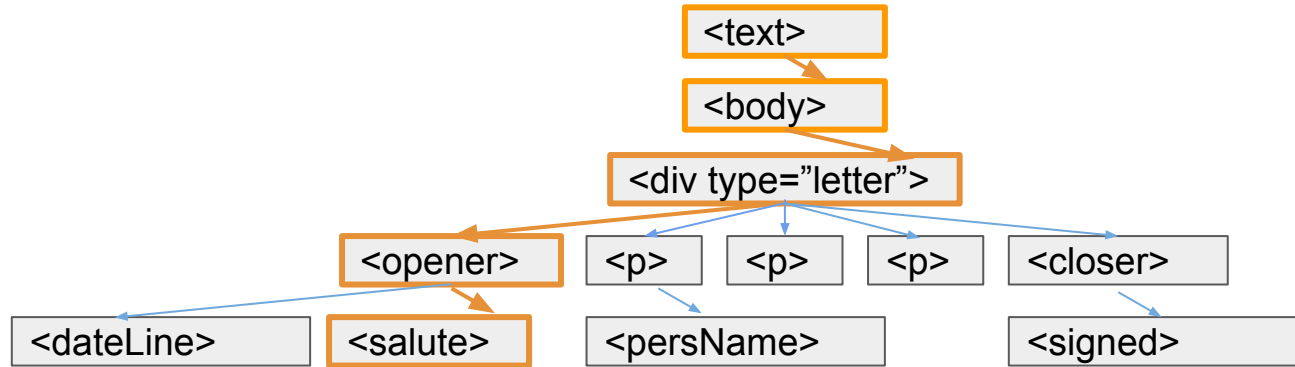   <p> does not mean "paragraph" and <name> does not mean "name" to the software.

Navigating an XML File Using xpath

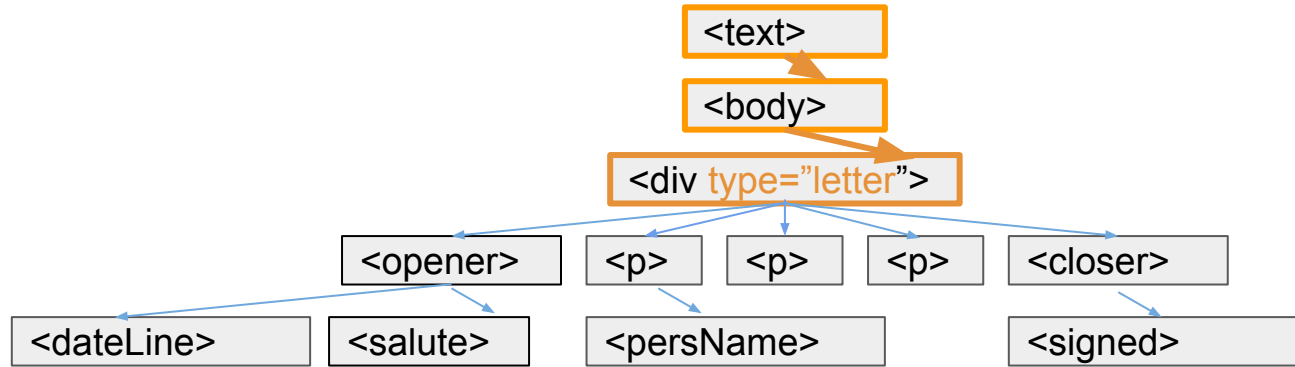A document that is encoded using XML can be visualized in the form of an upside down tree.
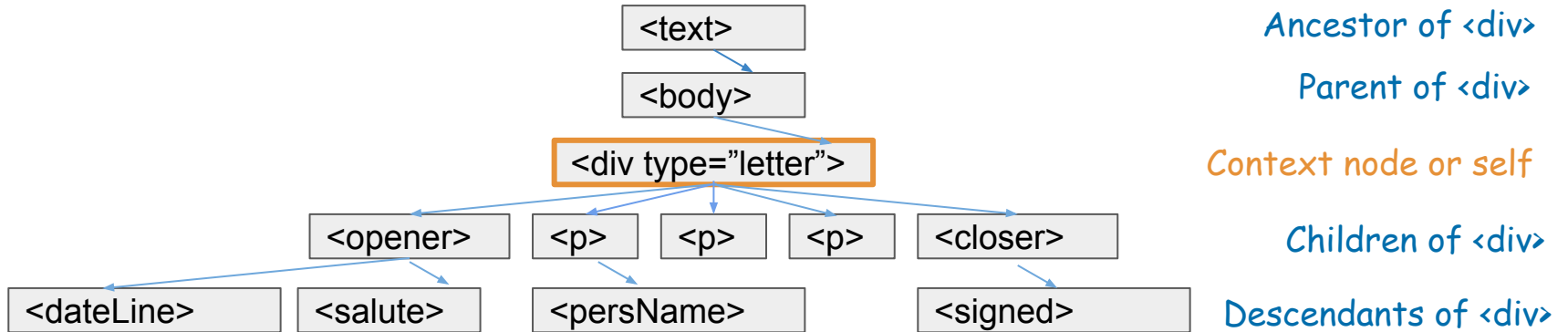
It is possible to identify elements by their location in the tree, specifically, how to navigate to that location.
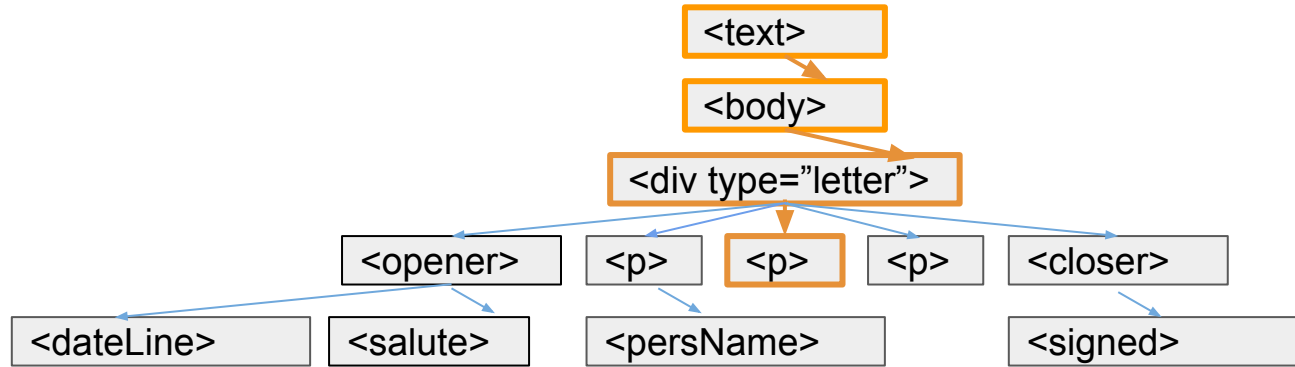
Navigating an XML File Using xpath (1)

/text/body/div/opener/salute

Navigating an XML File Using xpath (2)

`/text/body/div/@type`

Navigating an XML File Using xpath (3)

`/text/body/div/p[2]`= second paragraph

`/text/body/div/*[2]`= second child of <div>

Exercises