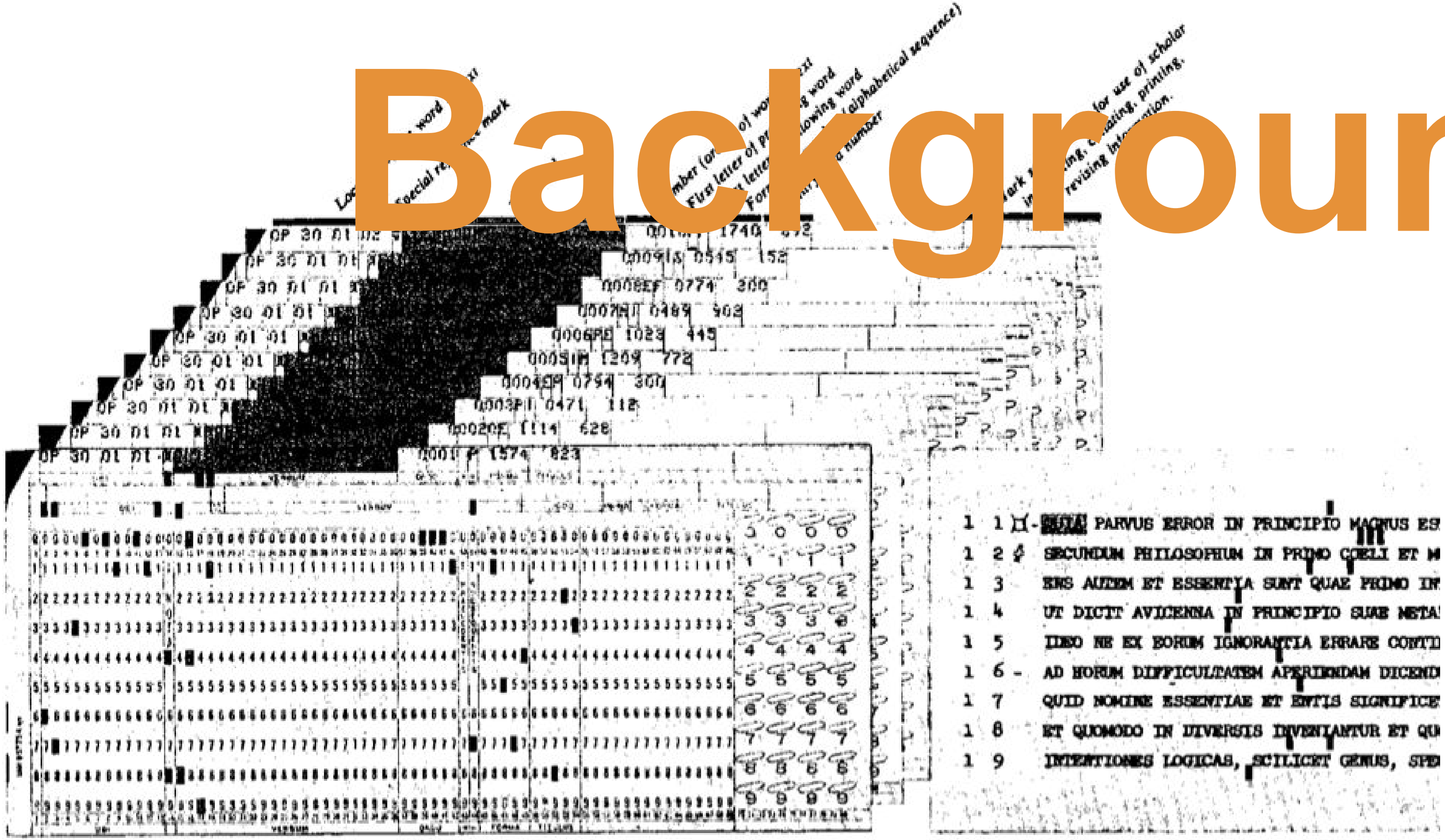
Markup Languages, XML and Text Encoding

DiXiT workshop, 24-28 April 2017

Overview

- Background
- Machine Readable Texts
- Text Representation
- Structures and Models
- XML

Background



Special reference mark
First letter of preceding word
First letter of following word
Form a number
alphabetical sequence)
for use of scholar
revising information.
OP 30 01 01
0001 P 1574 823
0002 E 1114 628
0003 P 0471 112
0004 E 0794 300
0005 P 1209 772
0006 P 1023 445
0007 A 0489 908
0008 E 0774 300
0009 S 0515 152
1740

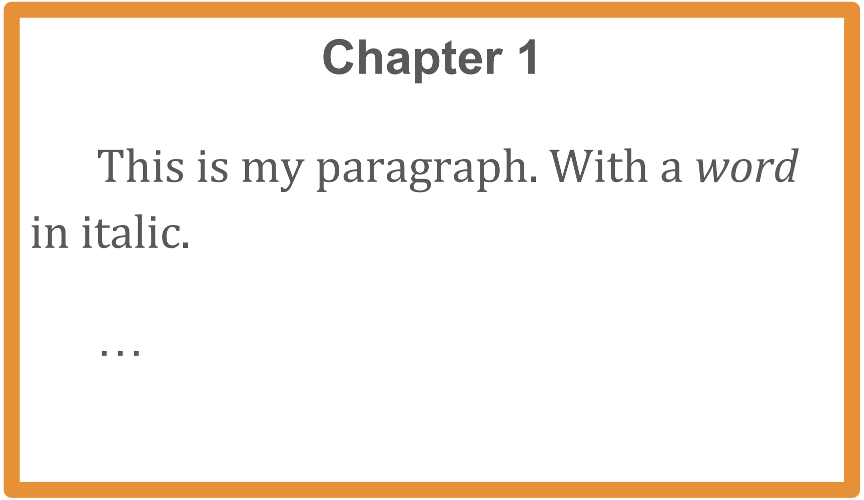
```
1  1 ¤-      PARVUS ERROR IN PRINCIPIO MAGNUS ES
1  2 ♂  SECUNDUM PHILOSOPHUM IN PRIMO COELI ET M
1  3    ENS AUTEM ET ESSENTIA SUNT QUAE PRIMO IN
1  4    UT DICIT AVICENNA IN PRINCIPIO SUAE META
1  5    IDEO NE EX EORUM IGNORANTIA ERRARE CONTI
1  6 -  AD HORUM DIFFICULTATEM APERIENDAM DICEND
1  7    QUID NOMINE ESSENTIAE ET ENTIS SIGNIFICE
1  8    ET QUOMODO IN DIVERSIS INVENIANTUR ET QU
1  9    INTENTIONES LOGICAS, SCILICET GENUS, SPE
```

P. Tasman, "Literary Data Processing," in *IBM Journal of Research and Development*, vol. 1, no. 3, pp. 249-256, July 1957. doi: 10.1147/rd.13.0249
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392686&isnumber=5392679

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

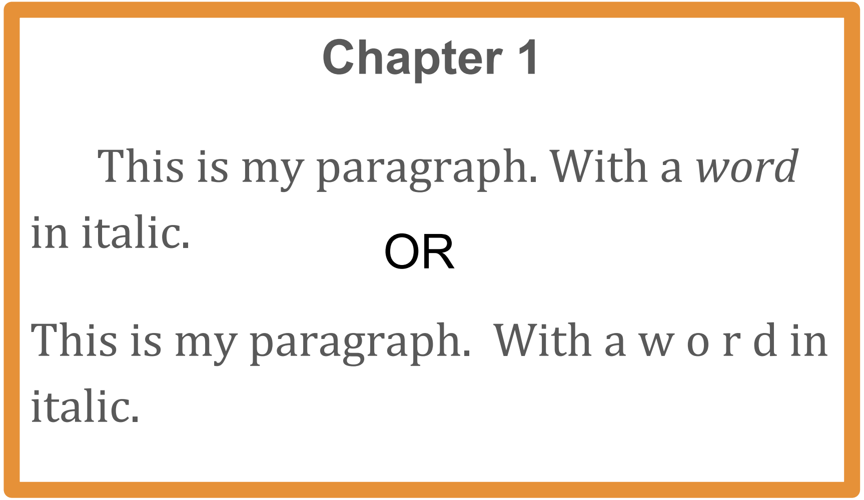
This is my paragraph. With a *word* in italic.

...

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

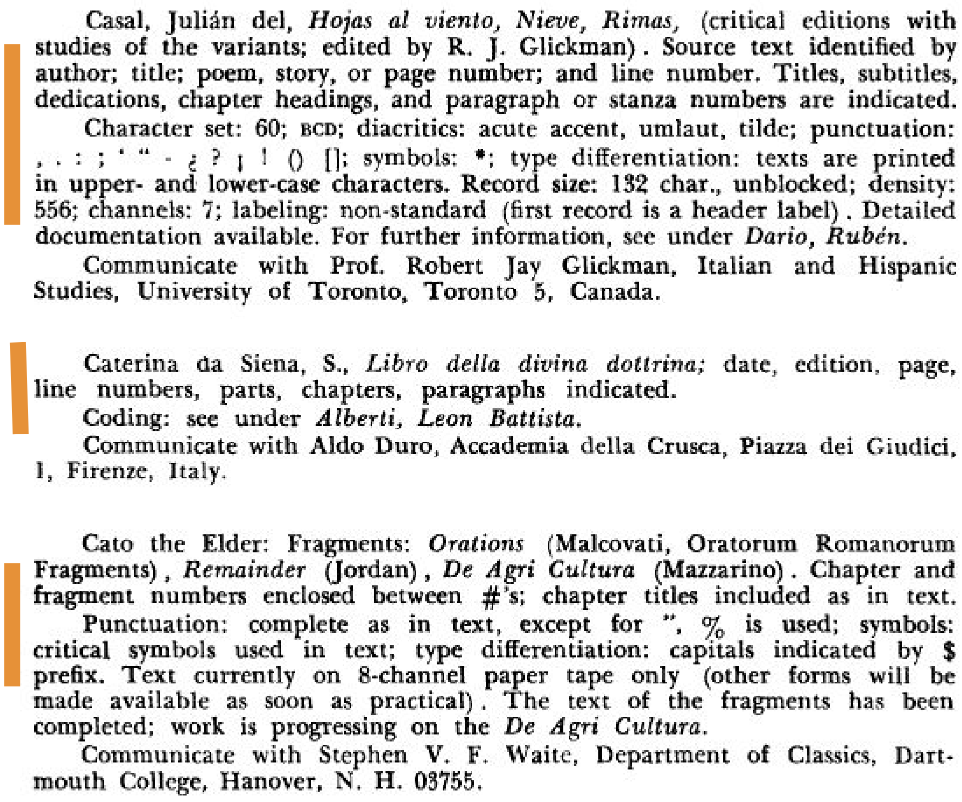
- Simplified and more extensible than SGML

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.

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 to the technical community

?Why use XML

- Captures semantic distinctions (not appearance)
- (designed for) Electronic publishing
- Single input, multiple outputs
- Interchange and Re-usability
- Sustainability
- Modeling and computability
- Community of peers
- Generalized tools

Structure

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

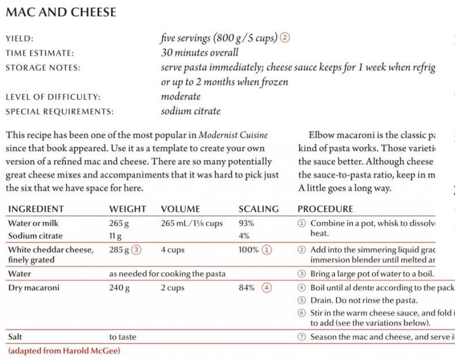
- Identify them
- Count them
- See what is missing
- Classify them
- Compare them
- Talk about them

Structuring: Turning a Text into Information

Texts contain and display implicit structure(s)

Explicit structure is a way to:

- Identify
- Locate
- Analyze
- Test



MAC AND CHEESE

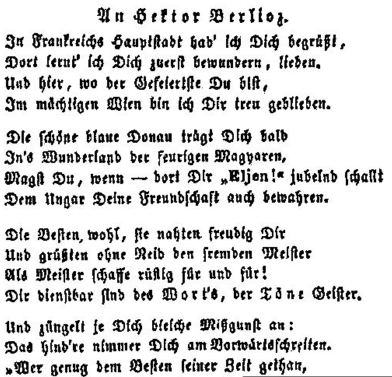
YIELD:	*five servings (800 g / 5 cups)* ②
TIME ESTIMATE:	*30 minutes overall*
STORAGE NOTES:	*serve pasta immediately; cheese sauce keeps for 1 week when refrig or up to 2 months when frozen*
LEVEL OF DIFFICULTY:	*moderate*
SPECIAL REQUIREMENTS:	*sodium citrate*

This recipe has been one of the most popular in *Modernist Cuisine* since that book appeared. Use it as a template to create your own version of a refined mac and cheese. There are so many potentially great cheese mixes and accompaniments that it was hard to pick just the six that we have space for here.

Elbow macaroni is the classic pa kind of pasta works. Those varieti the sauce better. Although cheese the sauce-to-pasta ratio, keep in m A little goes a long way.

INGREDIENT	WEIGHT	VOLUME	SCALING	PROCEDURE
Water or milk	265 g	265 mL / 1⅛ cups	93%	① Combine in a pot, whisk to dissolv heat.
Sodium citrate	11 g		4%	
White cheddar cheese, finely grated	285 g ③	4 cups	100% ①	② Add into the simmering liquid grad immersion blender until melted ar
Water	as needed for cooking the pasta			③ Bring a large pot of water to a boil.
Dry macaroni	240 g	2 cups	84% ④	④ Boil until al dente according to the package directions, 5–6 minutes. ⑤ Drain. Do not rinse the pasta. ⑥ Stir in the warm cheese sauce, and fold in any accompaniments you wish to add (see the variations below).
Salt	to taste			⑦ Season the mac and cheese, and serve it immediately.

(adapted from Harold McGee)



An Hektor Berlioz.

In Frankreichs Hauptſtadt hab' ich Dich begrüßt,
Dort lernt' ich Dich zuerſt bewundern, lieben.
Und hier, wo der Gefeiertſte Du biſt,
Im mächtigen Wien bin ich Dir treu geblieben.

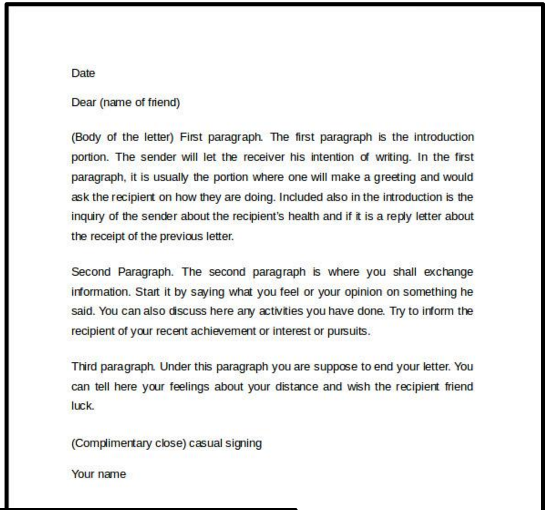
Die ſchöne blaue Donau trägt Dich bald
In's Wunderland der feurigen Magyaren,
Magſt Du, wenn — dort Dir „Eljen!“ jubelnd ſchallt
Dem Ungar Deine Freundſchaft auch bewahren.

Die Beſten, wohl, ſie nahten freudig Dir
Und grüßten ohne Neid den fremden Meiſter
Als Meiſter ſchaffe rüſtig für und für!
Dir dienſtbar ſind des Wort's, der Töne Geiſter.

Und züngelt ſe Dich bleiche Mißgunſt an:
Das hind're nimmer Dich am Vorwärtsſchreiten.
„Wer genug dem Beſten ſeiner Zeit gethan,
Der hat — mein Freund! — gelei

O, folge immer Deinem Génius!
So groß und kühn! — ſo rühmen
Sing' uns von erſter Liebe, erſten
Von wilden Leidenſchaften, ungewi

Sing' uns die Luſt des Carnevals!
Tyrteus gleich laß Freiheitslieder f
Sing' uns von Schlachten, wo in
Für's theure Vaterland die Helden



Date

Dear (name of friend)

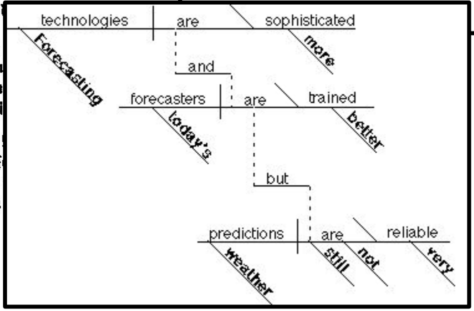
(Body of the letter) First paragraph. The first paragraph is the introduction portion. The sender will let the receiver his intention of writing. In the first paragraph, it is usually the portion where one will make a greeting and would ask the recipient on how they are doing. Included also in the introduction is the inquiry of the sender about the recipient's health and if it is a reply letter about the receipt of the previous letter.

Second Paragraph. The second paragraph is where you shall exchange information. Start it by saying what you feel or your opinion on something he said. You can also discuss here any activities you have done. Try to inform the recipient of your recent achievement or interest or pursuits.

Third paragraph. Under this paragraph you are suppose to end your letter. You can tell here your feelings about your distance and wish the recipient friend luck.

(Complimentary close) casual signing

Your name



Modeling

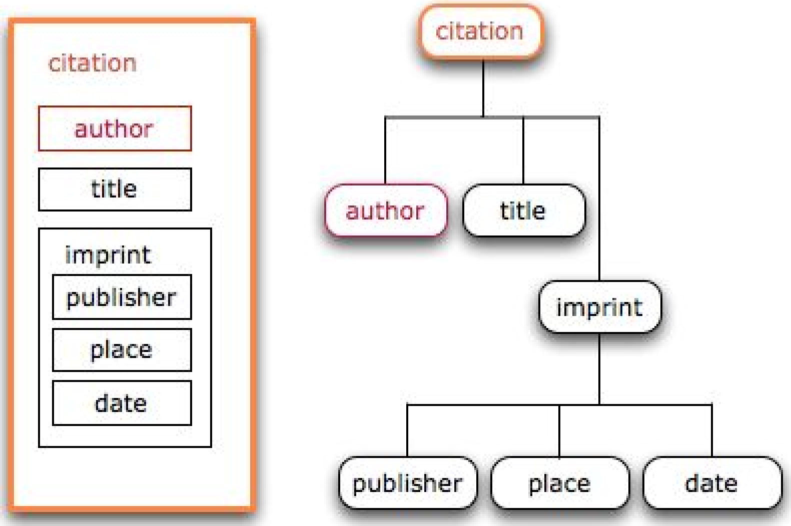
Adding structure to information is a way of modeling it

When you create a model and then apply it to a document or the features of a document it:

- Allows you to see how your document 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.



```
<?xml version="1.0"
encoding="UTF-8"?>
<citation>
  <author>Katherine Hayles</author>
  <title>Writing Machines</title>
  <imprint>
    <publisher>MIT Press</publisher>
    <place>Cambridge, MA</place>
    <date>2002</date>
  </imprint>
</citation>
```

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).

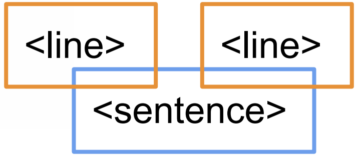
Σήμερα έδοσα ένα αντίγραφο στο

<name>Μαλάνο</name>.

Containment

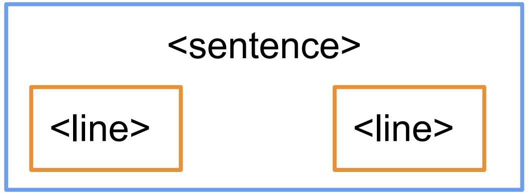
Elements may nest, but not may not overlap.

```
<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>
```



Empty Elements- κενά στοιχεία

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

```
<pagebreak></pagebreak>

= <pagebreak />
```

Certain elements, such as a page break marker, never have content as they are used to mark a point in the text, and not a span of text. These are referred to as *milestone* elements.

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>
```

An element may have more than one attribute.

```
<name type="person" role="writer">Μαλάνο</name>
```

Attribute Values—αξίες ιδιοτήτων

Attribute values (αξίες ιδιοτήτων) 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.

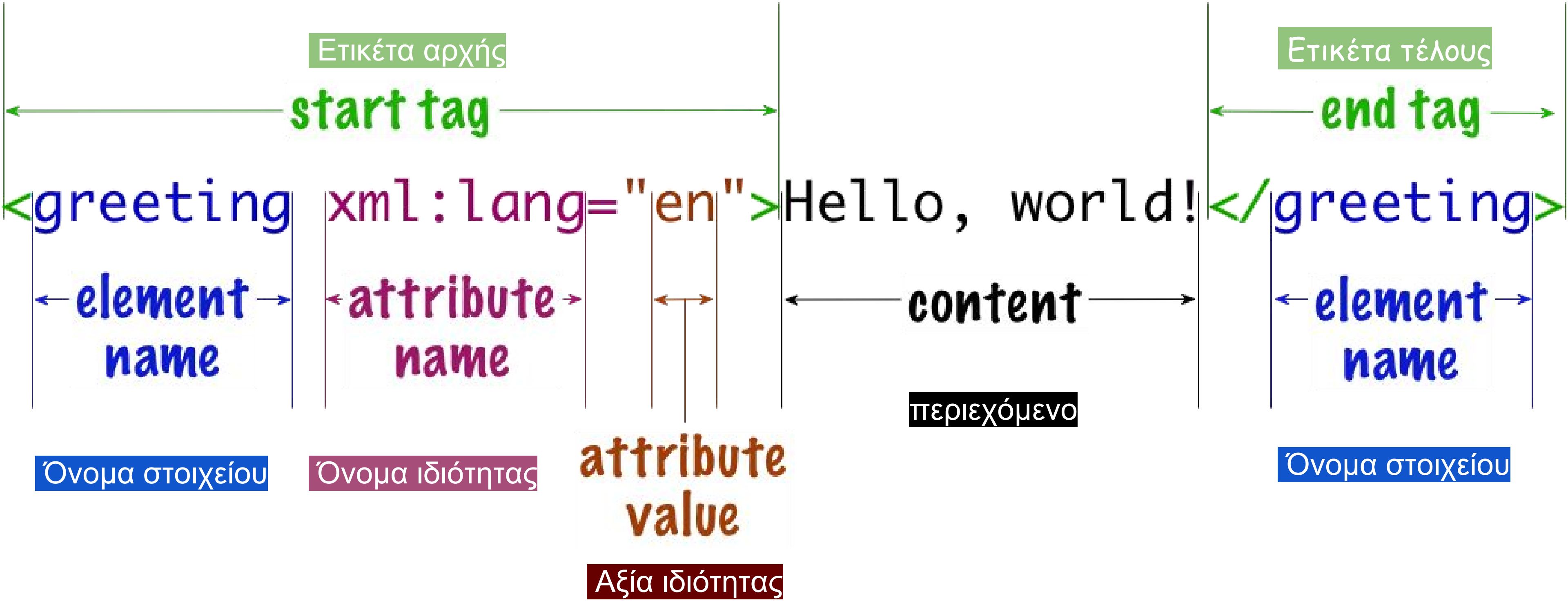
All elements may have an xml:id attribute.

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>
```

Anatomy of an Element



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

Ετικέτα αρχής

Ετικέτα τέλους

Όνομα στοιχείου

Όνομα ιδιότητας

Αξία ιδιότητας

περιεχόμενο

Όνομα στοιχείου

Flanders and Bauman 2016. http://www.wwp.northeastern.edu/outreach/seminars/_current/presentations/xml_intro/xml_newIntro_tutorial_02.xhtml

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.

Schema

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

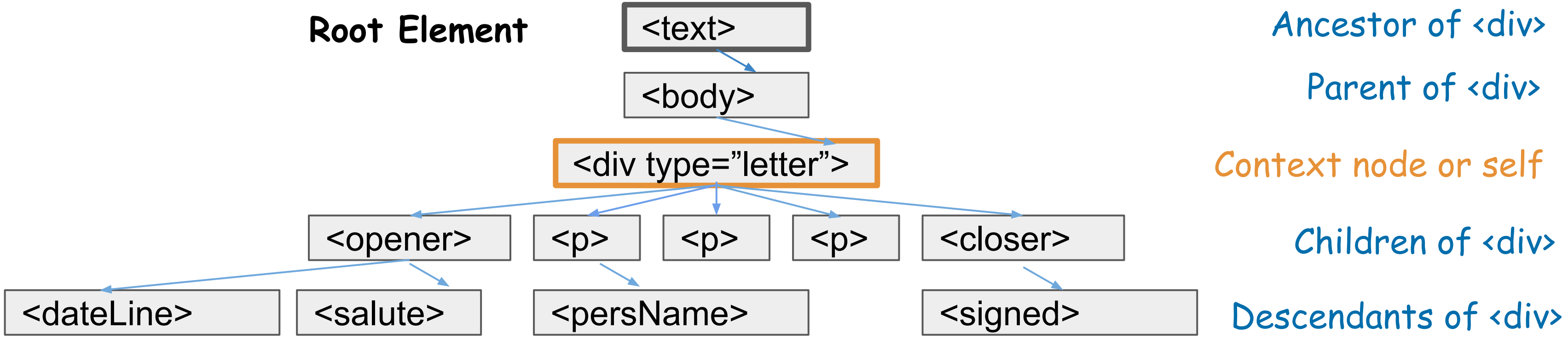
- A file that has correct XML syntax is well-formed.
- A file that is correct according to a schema is valid.

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.

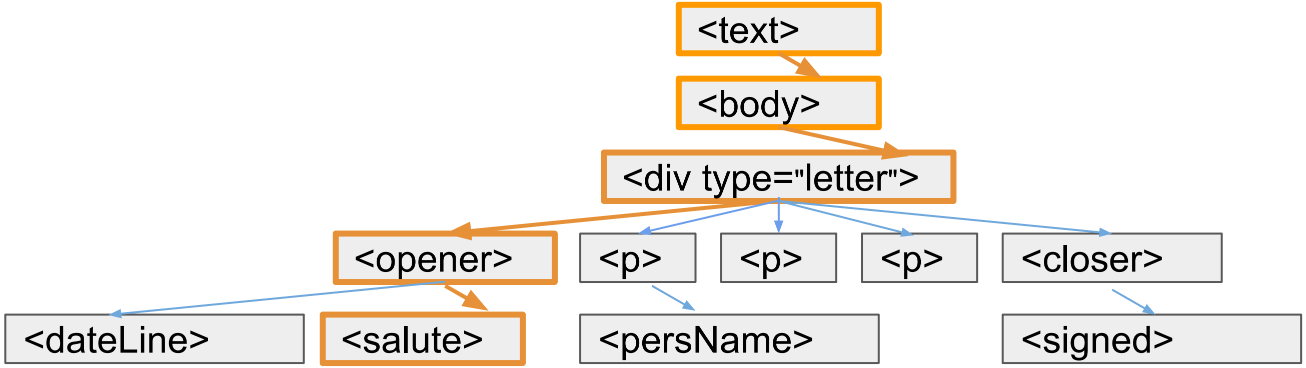
Navigating an XML File Using xpath

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 indicating how to navigate to them across the tree.

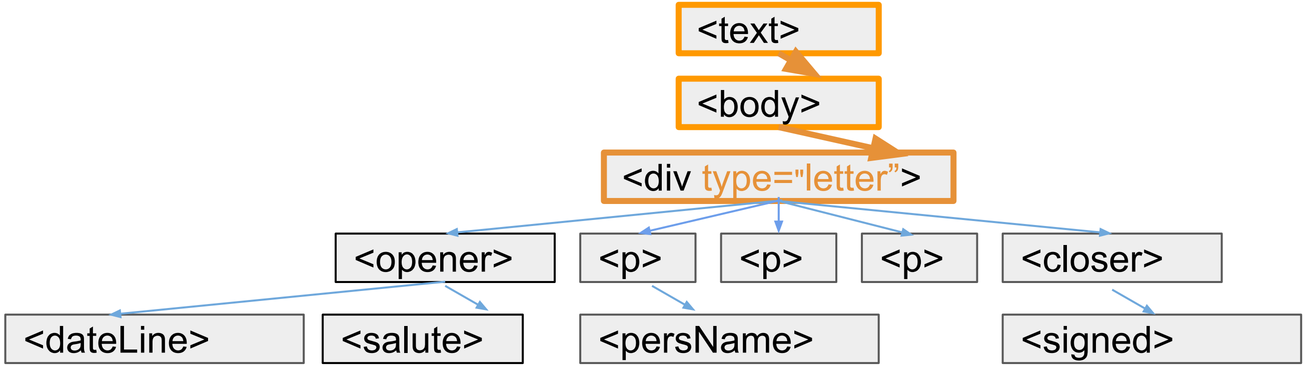
Navigating an XML File Using xpath (1)



`/text/body/div/opener/salute`

(points to <salute> and its contents)

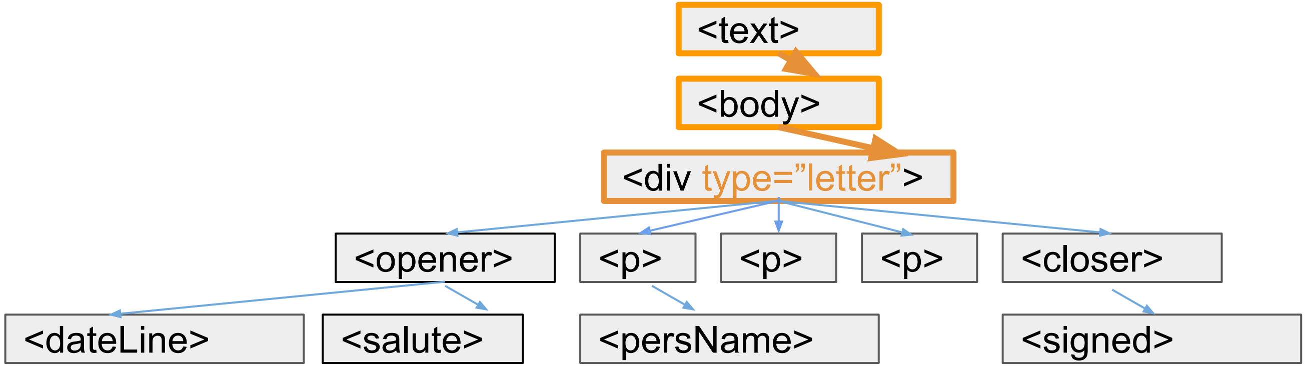
Navigating an XML File Using xpath (2)



`/text/body/div/@type`

(points to value of @type = “letter”)

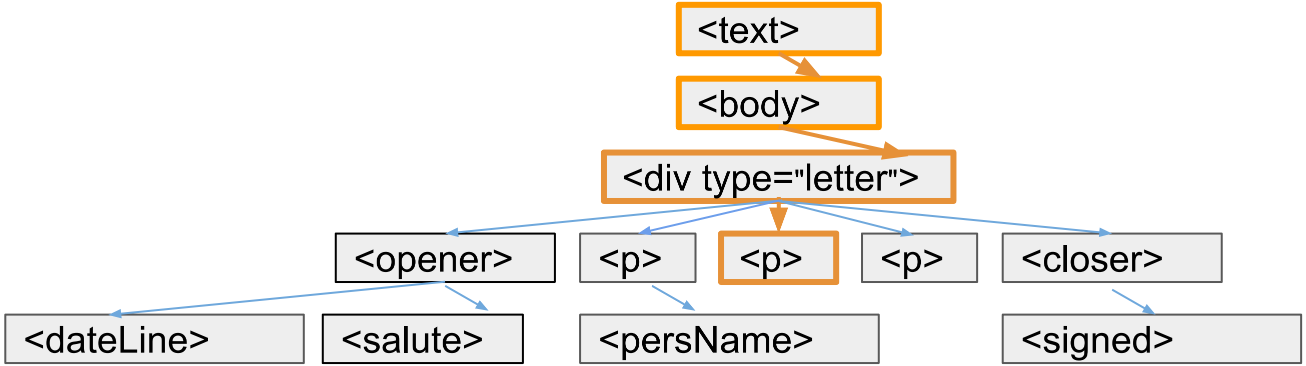
Navigating an XML File Using xpath (3)



`/text/body/div[@type]`

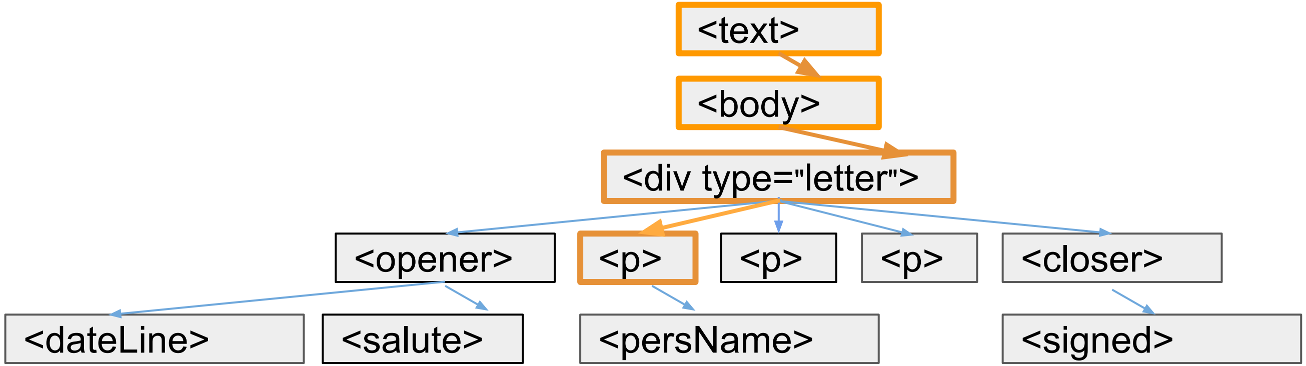
(points to <div> if it has a @type attribute)

Navigating an XML File Using xpath (4)



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

Navigating an XML File Using xpath (5)



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

Exercises