Intro to Markup and XML

Elli Mylonas
elli.mylonas@gmail.com

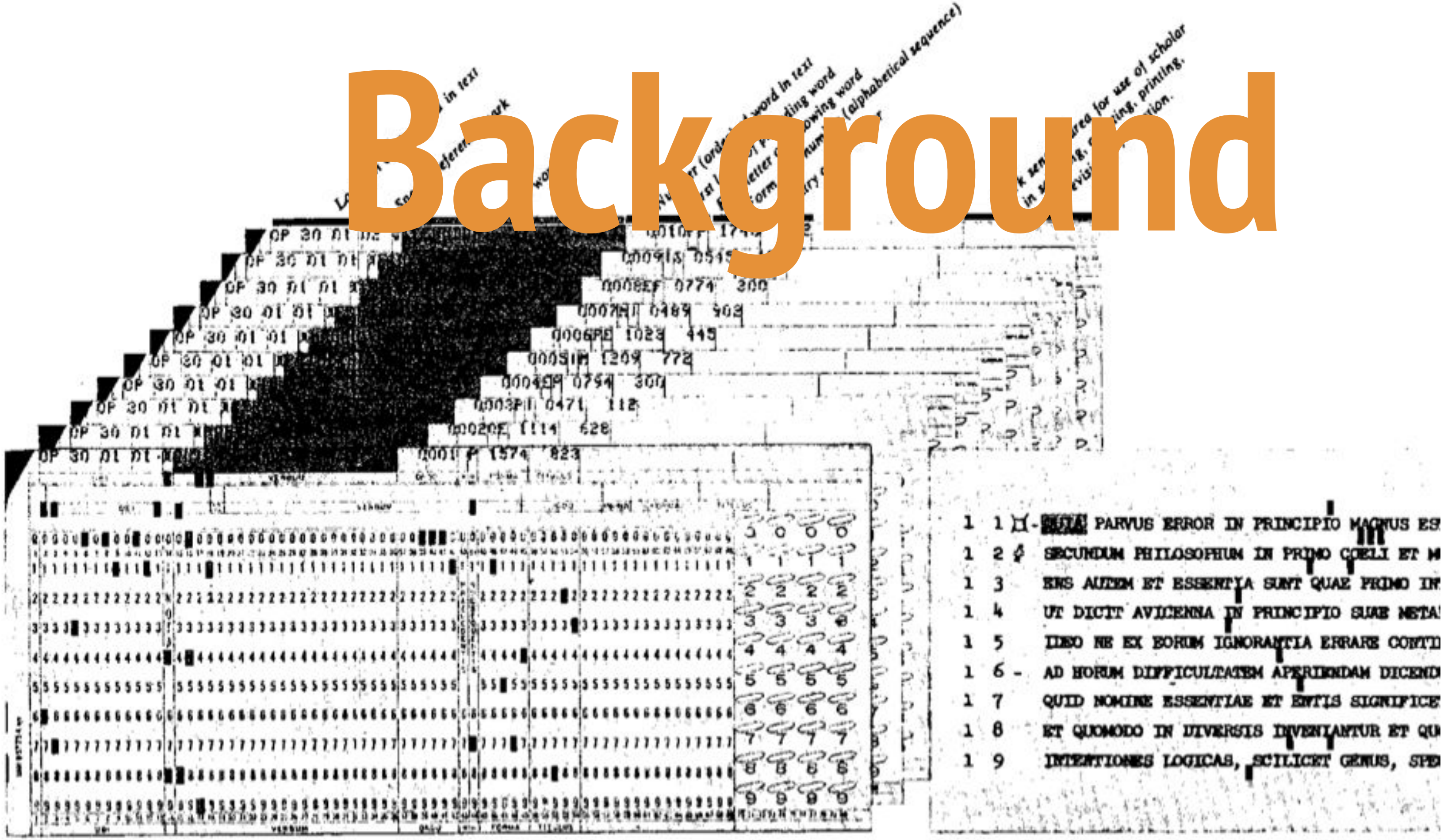
BærUt!
Sustainable Digital Scholarly Editions

This Presentation: http://tiny.cc/BaerUtMarkup

Overview

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

Background



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

Procedural Markup:

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.

…

Descriptive Markup:

Markup as semantic description

Structured text

```
<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 word in italic.

Descriptive Markup as a Standard

Technical documentation, government and legal documents needed to be:

- Machine readable
- Easy to maintain
- Reusable
- Sustainable over long periods of time

Resulting in a markup language that was

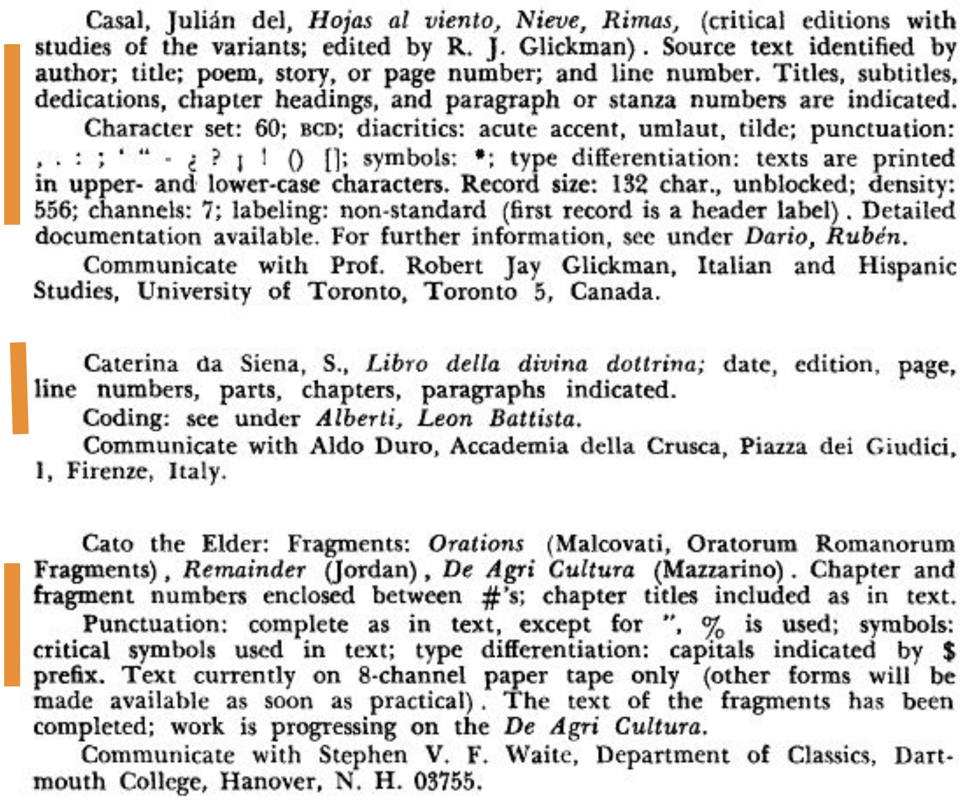
- Descriptive
- Standardized
- Formally defined and verifiable
- Independent of any particular output device
- Modifiable

Sharing Machine-Readable Texts—A long time ago

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

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

‘Literary Materials in Machine-Readable Form.” *Computers and the Humanities* 2, no. 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.

XML (and its parent, SGML)

- 1986: ISO 8879:1986 Information processing – Text and office systems – Standard Generalized Markup Language (SGML) published as a standard
  - Developed primarily for government and corporate materials online and in print
- 1990 WWW proposed by Tim Berners-Lee, using markup based on a simple form of SGML.
- 1998: XML version 1.0 released by the World Wide Web Consortium
  - Simplified and more extensible than SGML
  - Used for data interchange and for text formatting
- XML becomes the language of the Web

Early SGML Adopters in the Humanities

Oxford Text Archive

Women Writers Project

Valley of the Shadow

Whitman Archive

Perseus Project

Canterbury Tales Project

Linguistics

Dictionaries (OED)

Summary: 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:

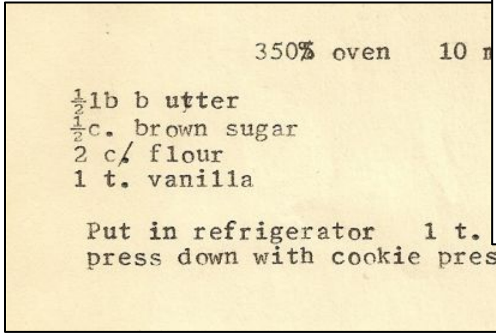
- Identify them
- Count them
- See what is missing
- Classify them
- Compare them
- Talk about them

Models: Turning a Text into Information

Texts contain and display implicit structure(s)

Explicit structure is a way to:

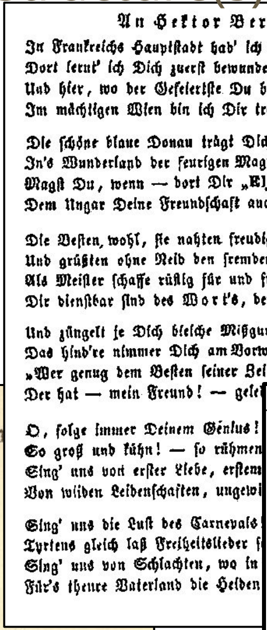
- Identify
- Locate
- Interpret
- Analyze
- Test



350% oven 10 m

½lb b utter
½c. brown sugar
2 c/ flour
1 t. vanilla

Put in refrigerator 1 t.
press down with cookie press



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.

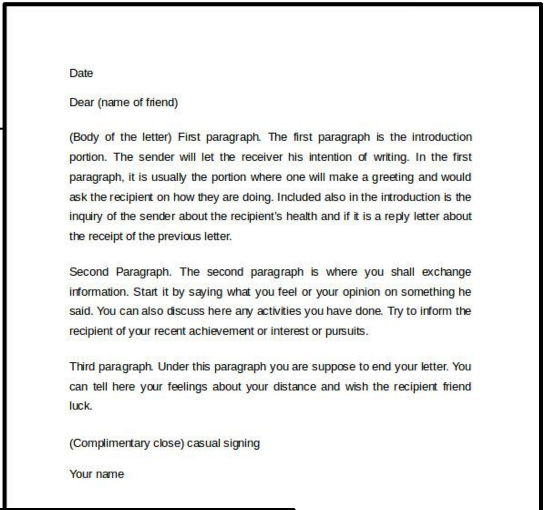
Die ſchöne blaue Donau trägt Dich bald
In's Wunderland der feurigen Magyaren,
Magſt Du, wenn — dort Dir „Eljen!“ jubelnd ſchallt
Dem Ungar Deine Freundſchaft auch bewahren.

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 je 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, erſtem
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)

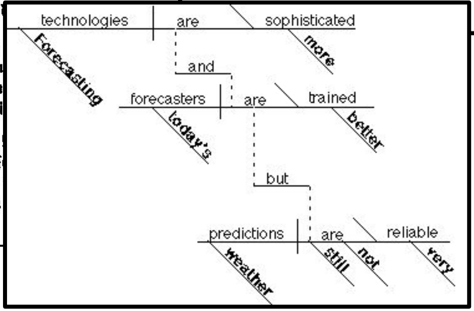
(Body of the letter) First paragraph. The first paragraph is the introduction portion. The sender will let the receiver his intention of writing. In the first paragraph, it is usually the portion where one will make a greeting and would ask the recipient on how they are doing. Included also in the introduction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



technologies
are
sophisticated
more
Forecasting
and
forecasters
are
trained
today's
better
but
predictions
are
reliable
weather
still
not
very

Modeling Information

Adding structure to a text reflects a theory about or an interpretation of the information it expresses.

Adding structure to text implements a specific model and point of view

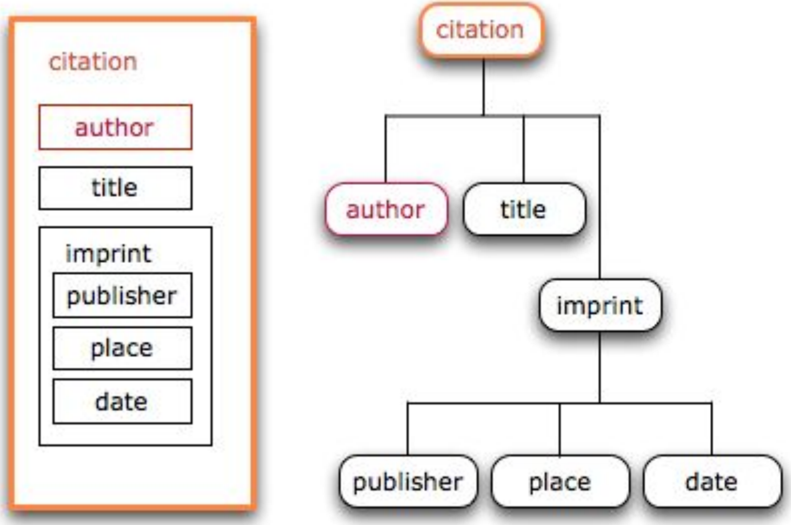
When you apply a model that identifies the significant features of a document, it:

- 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

XML Syntax

Structuring Documents with XML

XML models documents as a tree - a set of elements that can contain other elements.



```xml
<?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 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*

Elements

An element surrounds some text, and consists of a start tag and an end tag. These tags surround the element’s content .

```
Isn’t this a beautiful day,
<name>Charlie Brown</name>.
```

[Charles Schulz. “Peanuts” 10/7/1967]

PEANUTS
ISN'T THIS A BEAUTIFUL DAY, CHARLIE BROWN?
10-7

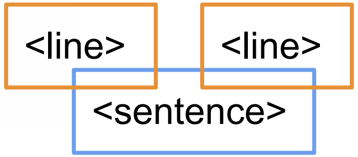
Containment

Elements may nest, but not may not overlap.

```
<line>I signed it <q>From</line>
<line>your sweet Babbooette.</q></line>
```

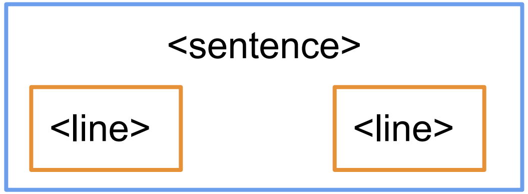
DID YOU GET MY VALENTINE? I SIGNED IT "FROM YOUR SWEET BABBOOETTE"

[Charles Schulz. "Peanuts" 2/14/1991]



Containment

```
<line>I signed it <q>From</q></line>
<line><q>your sweet
Babbooette.</q></line>
```



<sentence>

<line>

<line>

Empty Elements

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

```
<lb></lb>
 <lb />
```

} are equivalent

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.

```
<lb n="1"/><name key="Melior">Mel
<lb n="2" break="no"/>ior</name>
```

CA.Malibu.JPGM.L.84.AN.1.62, USEpigraphy

Attributes

PEANUTS
MISS OTHMAR WANTS YOU TO POUND ERASERS AGAIN, CHARLIE BROWN...
2-15

Start tags may have one or more attributes which provide information about the element or its content.

```
Miss Othmar wants you to pound erasers again,
<name type="person">Charlie Brown</name>...

<date when="1969-02-19">2-19</date>
```

An element may have more than one attribute.

```
<name type="person" role="teacher">Miss
Othmar</name>
```

Attribute Values

Attribute values may come from

- A closed list of values
- A list of suggested/recommended values
- Unrestricted

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

Attribute values consist of a alphanumeric characters and symbols.

```
<material type="#stone #metal">
```

`xml:id` Attribute

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

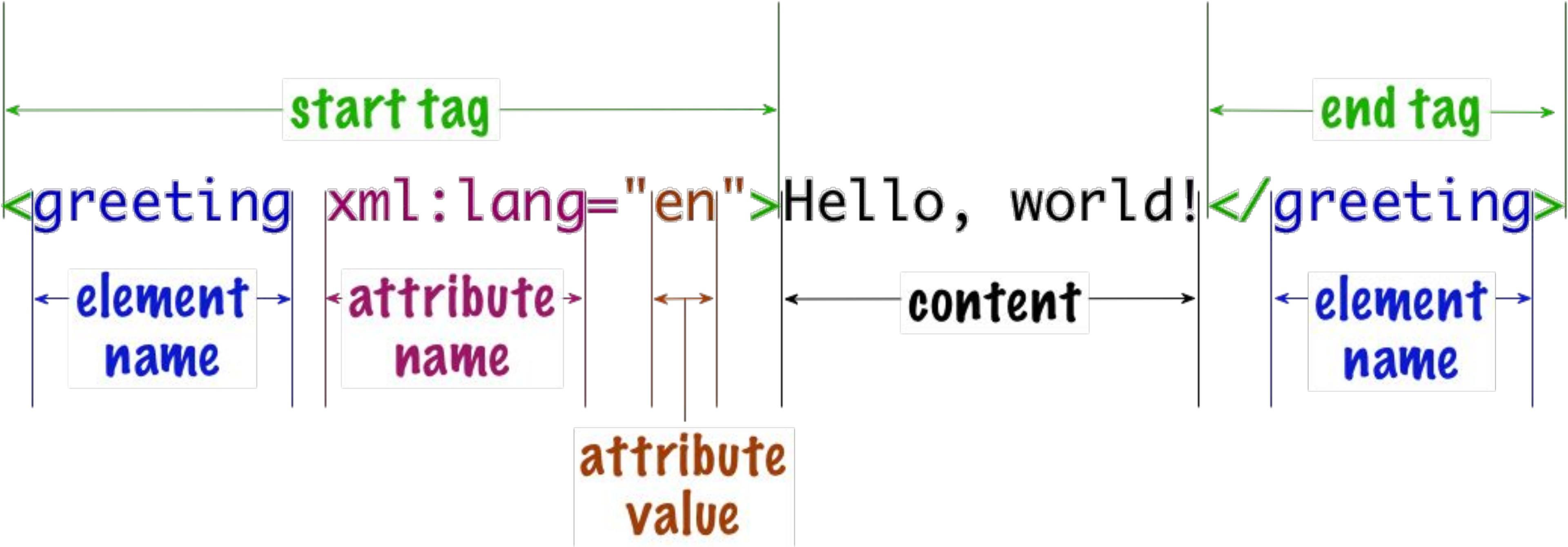
All elements may have an `xml:id` attribute.

By definition, an `xml:id` is

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

```
<persName xml:id="charlie_brown">...
<div xml:id="ch001">...
```

Anatomy of an Element



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

Comments

Comments may appear anywhere in an XML file. They are generally ignored by the XML parser.

Comments are very useful during the encoding process, but should not be relied on to convey information in the long term. That is better conveyed using markup to indicate editorial decisions.

```
<!-- EM 2025-09-19 This is a comment -->
```

Note some best practices evident in the sample comment above.

Well Formedness

When in a document

- There are no missing `< > / "` in tags and around attribute values
- Elements have matching start 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**

XML documents that are not well-formed are incorrect.

Schemas

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.

Some Schemas

- Menota (Medieval Nordic materials; TEI customization)
- MEI (Music Encoding Initiative; TEI customization)
- DocBook (for publishing)
- KnitML
- Epidoc (for digital epigraphy)
- Comic Book Markup Language (CBML)
- HTML (although not formally defined using a schema language)

XML files are text files

XML is written in Unicode

It is possible to open and edit an XML file with any text editor

XML aware editors like Oxygen and VSC are helpful as they

- *can check well formedness*
- *perform validation*
- *assist in editing with file name completion*

Top of an XML File

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model
href="http://epidoc.stoa.org/schema/latest/tei-epidoc.rng"
schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
```

l.1 The **XML declaration** identifies the file as being in XML

l.2 Points to the schema file for the document.

l.3 The root element of the file always has the **namespace(s)** of all the schemas used in the document in xmlns attribute. (more on that if necessary)

A Tiny Well Formed XML file

```xml
<?xml version="1.0" encoding="UTF-8"?><letter
type="friendly">
    <salute>Dear Parents</salute>
    <p1>I am having a good time a camp</p1>
    <p2>Please send cookies</p2>
    <closer>hugs and kisses
        <signed>Scooter</signed>
    </closer>
</letter>
```

A Tiny Valid XML File

```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="letter.rng"?>
<letter type="friendly">
    <salute>Dear Parents</salute>
    <p n="01">I am having a good time a camp</p>
    <p n="02">Please send cookies</p>
    <closer>hugs and kisses
        <signed>Scooter</signed>
    </closer>
</letter>
```

Exercise

For anyone who wants practice with XML and Oxygen, we can perform one or two simple encoding exercises this afternoon.