Markup Basics
An introduction to XML markup

Alejandro Bia

Miguel Hernández University
http://www.umh.es/
Alicante - Spain

Markup (1)

Markup (or encoding) is any means of making explicit an interpretation of a text.

Scriptio continua

Scriptio continua: ("Continuous script") is a classical style of writing without spaces between words or sentences, with all the text in upper case, and with no punctuation:

ISELLAKEYBOARDWITHTHESPACEBARBROKENVERYCHEAP

This style was often found in Greek manuscripts. It may have been used in order to conserve expensive manuscript material like vellum. It is used currently in Thai and other Southeast Asian abugidas, though with sentence breaks.

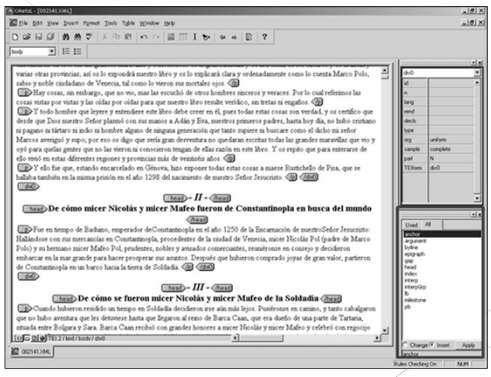
Markup (2)

Markup: Encoding a text for computer processing is like transcribing a manuscript from scriptio continua:

I SELL A KEYBOARD WITH THE SPACE BAR BROKEN, VERY CHEAP.

- a process of **making explicit what is conjectural or implicit.**
- a process of guiding the user as to how the content of the text should be interpreted.
- It is also making it easier to process by automatic means.

How a marked-up text looks like (using a specialized text editor)



This is XML markup

```
<div>
  <author type="first author">Ann (Ward) Radcliffe</author>
  <dates>1764-1823</dates>

  <quotation>
    <line>Fate sits on these dark battlements and frowns,</line>
    <line>And as the portal opens to receive me,</line>
    <line>A voice in hollow murmurs through the courts</line>
    <line>Tells of a nameless deed.</line>
    <source>The Mysteries of Udolpho.</source>
  </quotation>

</div>
```

LaTeX: Another way of of encoding text

```
\title{This is a PHD thesis in LATEX format}
\author{Whoever}

\maketitle

\begin{abstract}
The abstract document appears before any front matter.  As can be seen, it
is doubled-spaced in the final document.  Theses \emph{shouldn't} have to be
doubled-spaced, should they?  As you can see, it makes them awful!
\end{abstract}

\tableofcontents
\listoffigures
\listoftables

\include{Acknowledgements}
\mainmatter

% ---------------------------------------------------------------------------
\part{Summary}
% ---------------------------------------------------------------------------
\part{Acknowledgements}
% ---------------------------------------------------------------------------
\part{Prologue}
\chapter{Motivations}
When these ideas first came to light \footnote{The "dynamic and interactive environment"
      described by the authors of this abstract will be invaluable to librarians.}...

\chapter{Objectives and approach}
```


Markup (4)

- **Markup language:** a set of markup conventions used altogether for encoding texts.
- A markup language must specify:
  - what markup allows,
  - what markup requires,
  - how markup is to be distinguished from text,
  - what the markup means.

Markup (5)

- Coombs, Renear and DeRose argue that:
  "available word processors distract authors from their tasks of research and composition, toward concern with typographic and other tasks"
  (Coombs, Renear, DeRose:1987).
- They say that excessive concern on WYSIWYG technology, made us ignore the most important issue of representing document structure.

A classification of markup

Three types of markup:

- ***Punctuational***
- ***Presentational or procedural***
- ***Descriptive or structural ******

Markup Approaches

With a broader view,
we can distinguish 6 types of markup:

- Punctuational
- Presentational (WYSIWYG)
- Procedural (Troff, TeX, LaTeX)
- Descriptive (GML, SGML)
- Referential (embed, include; SGML)
- Metamarkup (SGML)

Data and Metadata Structuring

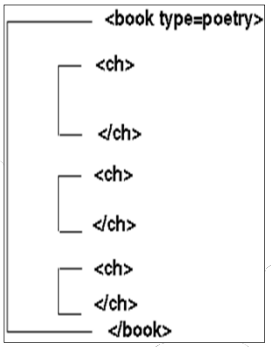
We use structured markup mainly for :

- Data: main body of literary works
- Metadata: bibliographical and processing information related to the literary work

Using structured markup will allow sophisticated searches in both data and metadata.

The OHCO Model

- **OHCO:** Ordered Hierarchy of Content Objects
- A **book** is a good example of what can be represented this way.
- Each Content Object is marked at its beginning and end with a tag.
- Ordering is usually needed, but unordered sets are OK sometimes.



Elements of markup (1)

In XML and SGML, tags are used to mark up the structure of the documents. In these tags we can easily distinguish three parts:

- **Elements** (name of the tag)
- **Attributes** (the elements' features)
- **Attribute Values**

- **Also internal Comments**

Elements of markup (2)

For example:

element --> "div"

attribute --> "type"

value --> "chapter"

Hence, the tag would look like:

```
<div type="chapter"> ... </div>
```


Elements of markup (3)

Empty elements:

<pb/>

This is equivalent to: <pb></pb>

(pb: means *page break***)**

Elements of markup (4)

Comments:

```
<!-- Anything you write here will not be
   seen in the final rendering of the
            document -->
```


Elements vs. Attributes (1)

- **They are both information containers**
  (except for empty elements, which are point markers)
- **But "attributes" are what the word means: attributes/features/characteristics/properties of the element they belong to**
- **Can be considered as metadata to the element, as they provide additional information about the element**

- **If an attribute does not provide information about the element, maybe it should not be an attribute, but a nested element.**

Elements vs. Attributes (2)

- **Example: How many date phrases do you see?**

The day of the battle was the 20th of March of 1854, which was also the general's birthday.

Elements vs. Attributes (2)

- **Example: How many date phrases do you see?**

The day of the battle was the 20th of March of 1854, which was also the general's birthday.

Elements vs. Attributes (2)

- **Example: How many date phrases do you see?**

** The <date>day of the battle</date> was <date>the 20th of March of 1854</date>, which was also <date>the general's birthday</date>. **

Elements vs. Attributes (2)

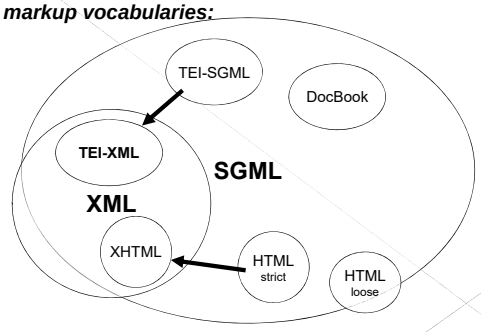
- **Example: How many date phrases do you see?**

** The <date date="20/03/1854">day of the battle</date> was <date date="20/03/1854">the 20th of March of 1854</date>, which was also <date date="20/03/1801">the general's birthday</date>. **

- **NOTE: The general was born in 1801,so by the time of the battle he was 53 years old.**

Markup languages and vocabularies

*The world of **markup languages** and **markup vocabularies:***



What is HTML?

- **HTML** *(HyperText Markup Language)* is the *lingua franca* for publishing hypertext on the WWW.
- **HTML** uses tags such as ** ** and ** ** to structure text into paragraphs, lists, etc.
- **Is a markup vocabulary for the Web**
- **HTML → presentational** markup (layout & appearance)
- **It's not meant for descriptive** markup (data structure)

What is SGML?

- **SGML** *(Standard Generalized Markup Language)* is an international standard for the definition of device-independent, system-independent methods of representing texts in electronic form.
- **SGML** is used for the description of marked-up electronic texts.
- **SGML** is a markup language, that is, a means of formally describing how to create markup vocabularies and how to use them.

SGML features

- There are three characteristics which separate **SGML** from other markup languages:
  - Its emphasis on descriptive rather than procedural markup
  - Its document-type concept
  - Its independence of any system for representing the script in which a text is written

What is XML?

- **XML** *(eXtensible Markup Language)* is another markup language like SGML, …BUT SIMPLER TO USE.
- It was thought to create documents for the Web.
- XML **IS NOT** a tagset, it is a markup language: it allow us to create a tagset and defines how tags must be used.

What is TEI?

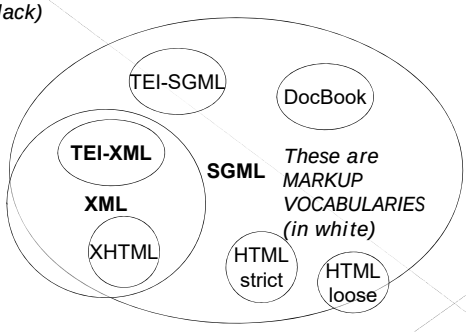
- TEI is a markup vocabulary: a set of tags with a specific name and meaning thought to mark a text.
- Difference between TEI and HTML:
  - HTML marks the format of the document (i.e. how it should look on a browser);
  - TEI marks the logical structure of the document (divisions, paragraphs, titles, etc.).

What is XHTML?

- XHTML is HTML 4.0 but following the rules of XML
- Among them:
  - Tags must be always in lowercase
  - All opened tags must be closed
  - There are empty elements:
    -
 instead of
  - Only HTML 4.0 elements are allowed:
    - instead of <b>
    - <emph> instead of <i>

Markup languages and vocabularies

These are MARKUP LANGUAGES (in black)

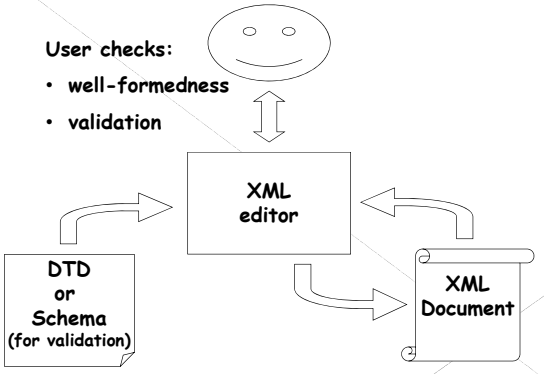


Well-formedness and Validation

An XML document is:

- **Well-formed:**
  - **When it complies to the rules of XML**
  - **If it is not well-formed, it is not XML**
- **Valid:**
  - **When it complies to the rules of a DTD or Schema**
  - **We can then say thet it belongs to a "document type"**
  - **To be valid it must first be well-formed**
  - **The use of a DTD/Schema is not mandatory but usually necessary**

Well-formedness and Validation



XML Document Structure

- The XML declaration
- The DOCTYPE declaration
- Association to STYLESHEETS
- The document BODY

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE TEI SYSTEM "cervantes2013.dtd">
<?xml-stylesheet href="teixlite2013.css" type="text/css"?>
<TEI>
   …
   Here goes the metadata and the body of the text...
   …
</TEI>
```