

XML

What is markup?

Markup is a type of annotation

- Markup has been used for centuries in publishing as a means of providing formatting instructions to printers (human and machine).
- “Instructions for the typesetter that are written on the copy (e.g. underlining words that are to be set in italics).” - Webster’s Online Dictionary

The importance of formatting

- Textual formatting has a lot to do with how we understand written text.
- Texts tend to have a familiar structure.
- We use non-alphanumeric characters to indicate meaning.
- We also use different *styles* to affect meaning.

an example

What is this?

οἱ μὲν ἰππήων στρότον, οἱ δὲ πέσδων,
οἱ δὲ νάων φαῖσ' ἐπὶ γᾶν μέλαιναν
ἔμμεναι κάλλιστον. ἔγω δὲ κῆν' ὅτ-
τω τις ἔραται.

Poetry, obviously

A bit easier to figure out than this:

ΟΙΜΕΝΙΠΠΗΩΝΣΤΡΟΤΟΝΟΙΔΕΠΕΣΔΩΝΟΙΔΕ
ΝΑΩΝΦΑΙΣΕΠΙΓΑΝΜΕΛΑΙΝΑΝΕΜΜΕΝΑΙΚΑΛ
ΛΙΣΤΟΝΕΓΩΔΕΚΗΝΟΤΤΩΤΙΣΕΡΑΤΑΙ

Some thoughts:

- Markup is meta-textual information.
- That is, it provides information on how a text should be interpreted and re-rendered.
- Formatting influences semantics, *i.e.*, it implies meaning.
- So we see (and use) some types of markup all the time...

Examples of markup

- Highlighted sections of text.
- Corrections on a “marked” paper.
- Marginal comments.
- But we might also understand markup to trespass into the text itself.
 1. Punctuation.
 2. Spacing.

Markup Languages

- Markup languages provide standardized ways of annotating and structuring texts and data.
- These languages do things like:
 1. Provide formatting instructions to a rendering engine.
 2. Make semantic distinctions between portions of a text.
 3. Provide datatyping information.
 4. Represent the underlying structure of a document.
 5. Add metadata to a document.
 6. Provide linking mechanisms between documents and within documents.

Let's get a bit less abstract

- HTML stands for HyperText Markup Language.
- So in what senses is this markup?
 1. HTML provides structure
 2. and formatting instructions
 3. and *some* semantics
 4. and allows the addition of metadata
 5. and provides linking mechanisms.
- It does a lot of things very badly though...

Let's get a bit more abstract again...

- HTML is a markup language, in the “proper” sense of the term.
- HTML is an *application* of SGML, Standard Generalized Markup Language.
- SGML, even though it has the words “markup language” in its name is really a syntax for creating markup languages, not a markup language itself.
- XML is a derivative of SGML. In fact, it *is* SGML, with more restrictions than standard SGML.

So XML is:

- A set of rules. If a document conforms to these rules, it is an XML document, or XML *instance*.
- An XML *application* is a grammar that specifies what tags can be used, and where.
- An XML *instance* is a document marked up in XML, **whether it uses a grammar or not.**

Two basic types of XML

- We've been assuming that XML is for texts. Actually it's commonly used for more rigidly structured kinds of data too.
- So you'll hear lots of talk about data-oriented vs. document-oriented XML.
- This is important because the two tend to pull in opposing directions.

Thought experiment

- Let's imagine a database with tables containing data. Maybe a directory with names, addresses, and phone numbers.
- Now, how would we store a book (say *David Copperfield*) in a structure like that?

The text

The screenshot shows a Mozilla Firefox browser window. The title bar reads "Mozilla Firefox". The address bar contains the URL "http://www.gutenberg.net/dirs/etext96/cprfd10.txt". The browser's menu bar includes "Mozilla Firebird Help", "Mozilla Firebird Disc...", and "Plug-in FAQ". The toolbar contains various icons for "Disable", "CSS", "Forms", "Images", "Information", "Miscellaneous", "Outline", "Resize", "Validation", and "View Source". The main content area displays the text of "THE PERSONAL HISTORY AND EXPERIENCE OF DAVID COPPERFIELD THE YOUNGER", specifically "CHAPTER 1 I AM BORN". The text is rendered in a monospaced font. The status bar at the bottom of the window shows "Done".

THE PERSONAL HISTORY AND
EXPERIENCE OF
DAVID COPPERFIELD THE YOUNGER

CHAPTER 1
I AM BORN

Whether I shall turn out to be the hero of my own life, or whether that station will be held by anybody else, these pages must show. To begin my life with the beginning of my life, I record that I was born (as I have been informed and believe) on a Friday, at twelve o'clock at night. It was remarked that the clock began to strike, and I began to cry, simultaneously.

In consideration of the day and hour of my birth, it was declared by the nurse, and by some sage women in the neighbourhood who had taken a lively interest in me several months before there was any possibility of our becoming personally acquainted, first, that I was destined to be unlucky in life; and secondly, that I was privileged to see ghosts and spirits; both these gifts inevitably attaching, as they believed, to all unlucky infants of either gender, born towards the small hours on a Friday night.

I need say nothing here, on the first head, because nothing can show better than my history whether that prediction was verified or falsified by the result. On the second branch of the question, I will only remark, that unless I ran through that part of my inheritance while I was still a baby, I have not come into it yet. But I do not at all complain of having been kept out of this property; and if anybody else should be in the present enjoyment of it, he is heartily welcome to keep it.

I was born with a caul, which was advertised for sale, in the newspapers, at the low price of fifteen guineas. Whether sea-going people were short of money about that time, or were short of faith and preferred cork jackets, I don't know; all I know is, that there was but one solitary bidding, and that was from an attorney connected with the bill-broking business, who offered two pounds in cash, and the balance in sherry, but declined to be guaranteed from drowning on any higher bargain. Consequently the advertisement was withdrawn at a dead loss - for as to sherry, my poor dear mother's own sherry was in the market then - and ten years afterwards, the caul was put up in a raffle down in our part of the country, to fifty members at half-a-crown a head, the winner to spend five shillings. I was present myself, and I remember to have felt quite

Done

Tables

- We could have a “words” table:

id	word
1	Whether
2	I
3	shall
4	turn
5	out
6	to
7	be
8	the
9	hero
10	of
11	my
12	own
13	life
14	,
15	or
16	that

Tables

- And an “order” table:

id	order
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
14	14
15	15
1	16
16	17

- and so on. Unwieldy though, isn't it? What would we do about paragraphs? Chapters? Chapter headings?

The moral of the story:

- Information that is loosely structured is hard to store in highly structured containers, such as relational databases.
- The converse is true also: highly structured information really benefits from the things that highly structured containers have to offer, like data typing, enforcement of relational integrity, constraints, etc.

data vs. document

- So these two kinds of information have fundamental differences...
- And XML handles both...

Poetry again

Some say a host of cavalry, some of infantry,
others of ships, is the most beautiful thing
upon the dark earth. But *I* say it
is whatever one loves.

Let's mark it up

<stanza>

<l>Some say a host of cavalry, some of infantry,</l>

<l>others of ships, is the most beautiful thing</l>

<l>upon the dark earth. But <emphasis>I</emphasis> say it</l>

<l>is whatever one loves.</l>

</stanza>