E**X**tensible **M**arkup **L**anguage

(XML)

Very brief intro to XML

Why would you encode documents in extensible markup language (XML)?

- XML is a metalanguage (a language about markup languages).
- It is also a formal model that represents an texts in an ordered hierarchy. We think and write with hierarchical structures, and XML is a practical attempt to formalise and represent documents in a machine-readable language.
- Computers can operate quickly and efficiently on trees (ordered hierarchies) much more quickly and efficiently than they can on non-hierarchical text. Large amounts of data can be managed and transformed efficiently if a text is modelled as a tree.

Document Hierarchy

- The hierarchy imposed on documents depends on the state of surviving documents as well as one's research questions. The same document can be encoded in more than one hierarchical (or non-hierarchical, for that matter) structure.

- Presentational versus Descriptive markup: ***Presentational*** describes what text looks like. ***Descriptive*** describes what a textual subcomponent is. The markup used in digital research projects is mostly descriptive.

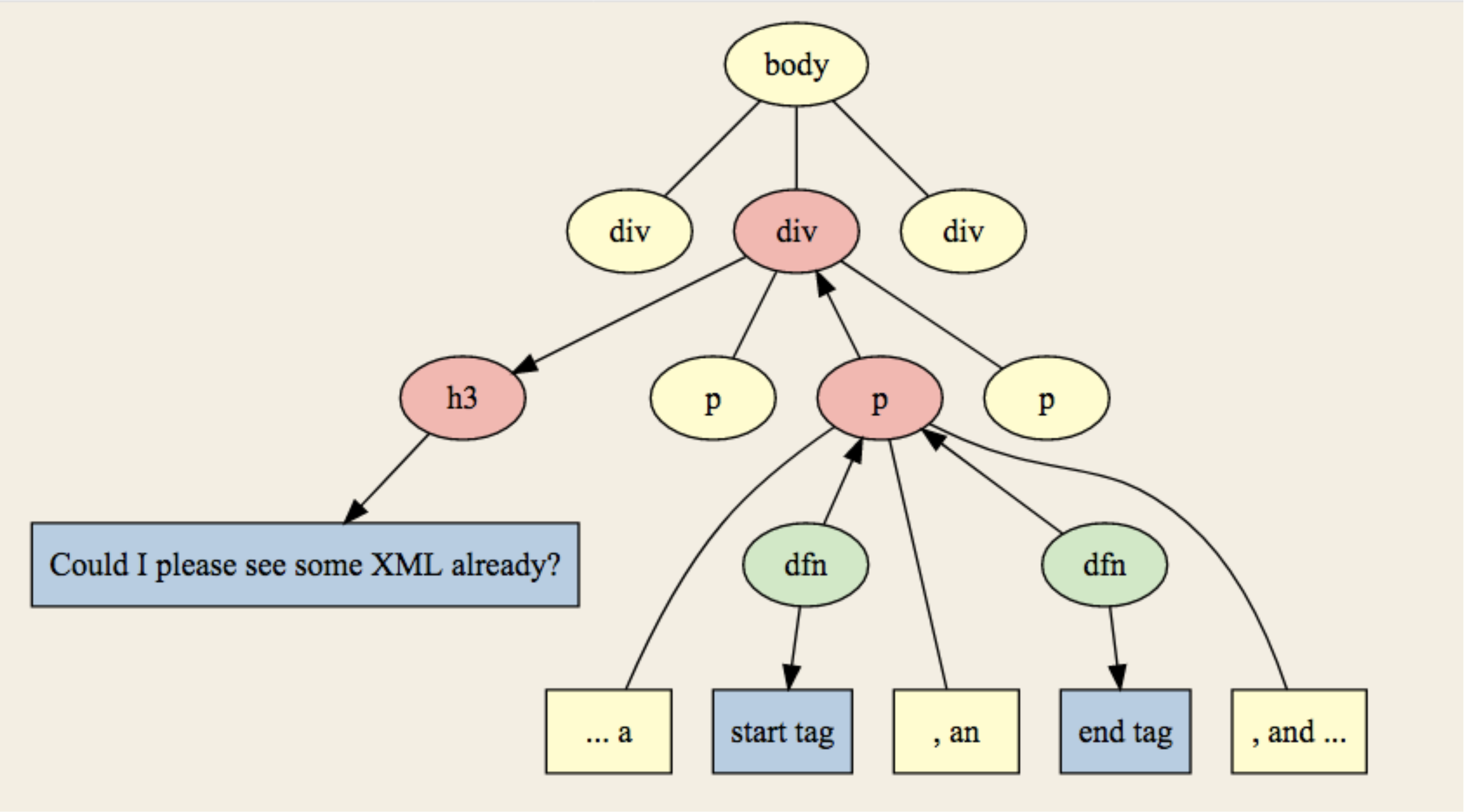


Image courtesy David Birnbaum, <http://dh.obdurodon.org/what-is-xml.xhtml>.

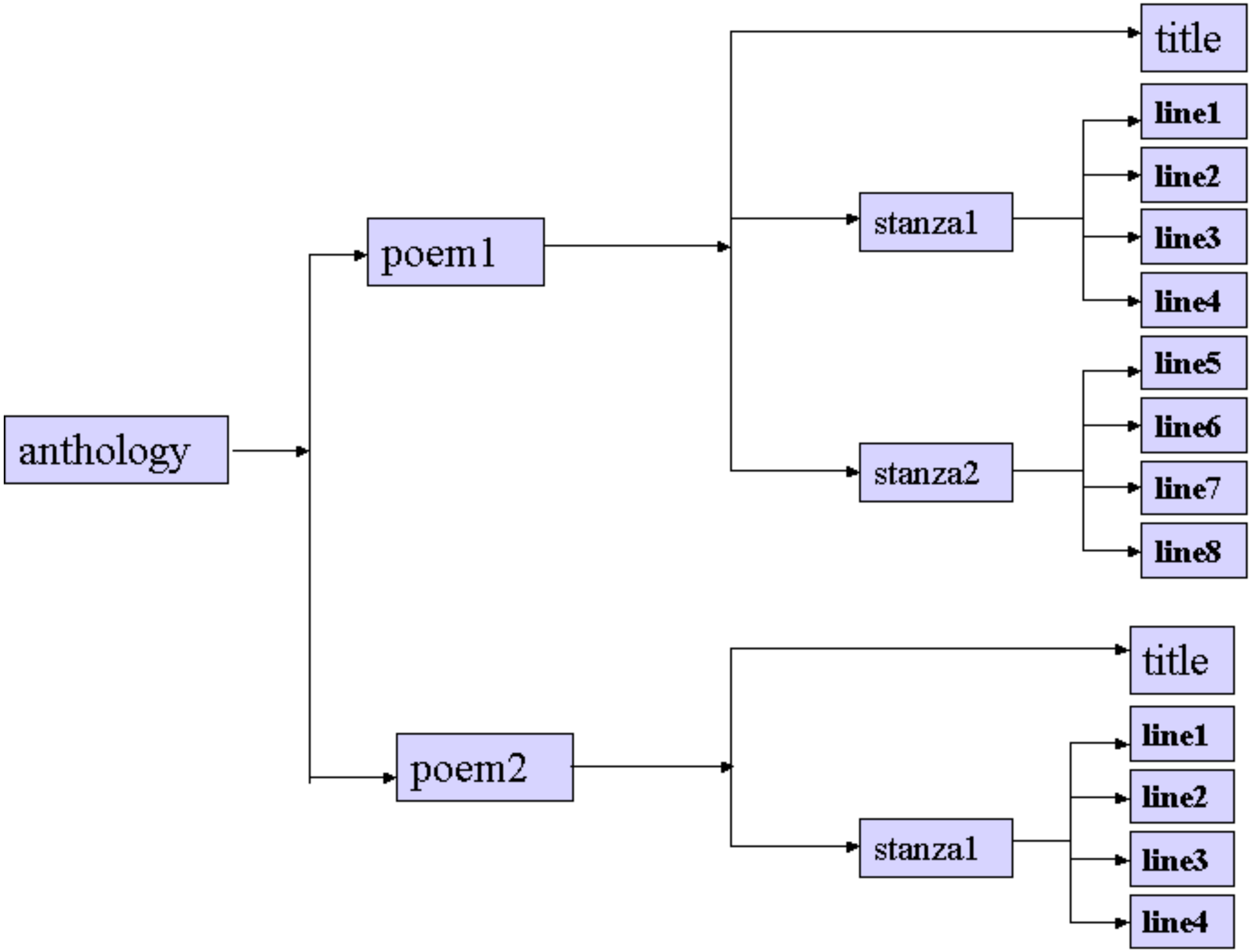


Image courtesy "A Gentle Introduction to XML," *TEI Guidelines* <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html>.

HTML

- Predefined tags e.g., <p> <h1> <table>
- Designed to display data in web browser
- Presentation language

XML

- You define your own tags (that’s the extensible part)
- Designed to describe a text document
- Data description language

Bleak House

Chapter 1

In Chancery

LONDON. Michaelmas Term lately over, and the Lord Chancellor sitting in Lincoln’s Inn Hall. Implacable November weather. As much mud in the streets as if the waters had but newly retired from the face of the earth, and it would not be wonderful to meet a Megalosaurus, forty feet long or so, waddling like an elephantine lizard up Holborn Hill. Smoke lowering down from chimney-pots, making a soft black drizzle, with flakes of soot in it as big as full-grown snow-flakes — gone into mourning, one might imagine, for the death of the sun. Dogs, undistinguishable in mire. Horses, scarcely better; splashed to their very blinkers. Foot passengers, jostling one another’s umbrellas in a general infection of ill-temper, and losing their foot-hold at street-corners, where tens of thousands of other foot passengers have been slipping and sliding since the day broke (if the day ever broke), adding new deposits to the crust upon crust of mud, sticking at those points tenaciously to the pavement, and accumulating at compound interest.

Fog everywhere. Fog up the river, where it flows among green aits and meadows; fog down the river, where it rolls defiled among the tiers of shipping and the waterside pollutions of a great (and dirty) city. Fog on the Essex marshes, fog on the Kentish heights. Fog creeping into the cabooses of collier-brigs; fog lying out on the yards, and hovering in the rigging of great ships; fog drooping on the gunwales of barges and small boats. Fog in the eyes and throats of ancient Greenwich pensioners, wheezing by the firesides of their wards; fog in the stem and bowl of the afternoon pipe of the wrathful skipper, down in his close cabin; fog cruelly pinching the toes and fingers of his shivering little ’prentice boy on deck. Chance people on the bridges peeping over the parapets into a nether sky of fog, with fog all round them, as if they were up in a balloon, and hanging in the misty clouds.

Bleak House

Chapter 1

In Chancery

LONDON. Michaelmas Term lately over, and the Lord Chancellor sitting in Lincoln’s Inn Hall. Implacable November weather. As much mud in the streets as if the waters had but newly retired from the face of the earth, and it would not be wonderful to meet a Megalosaurus, forty feet long or so, waddling like an elephantine lizard up Holborn Hill. Smoke lowering down from chimney-pots, making a soft black drizzle, with flakes of soot in it as big as full-grown snow-flakes — gone into mourning, one might imagine, for the death of the sun. Dogs, undistinguishable in mire. Horses, scarcely better; splashed to their very blinkers. Foot passengers, jostling one another’s umbrellas in a general infection of ill-temper, and losing their foot-hold at street-corners, where tens of thousands of other foot passengers have been slipping and sliding since the day broke (if the day ever broke), adding new deposits to the crust upon crust of mud, sticking at those points tenaciously to the pavement, and accumulating at compound interest.

Fog everywhere. Fog up the river, where it flows among green aits and meadows; fog down the river, where it rolls defiled among the tiers of shipping and the waterside pollutions of a great (and dirty) city. Fog on the Essex marshes, fog on the Kentish heights. Fog creeping into the cabooses of collier-brigs; fog lying out on the yards, and hovering in the rigging of great ships; fog drooping on the gunwales of barges and small boats. Fog in the eyes and throats of ancient Greenwich pensioners, wheezing by the firesides of their wards; fog in the stem and bowl of the afternoon pipe of the wrathful skipper, down in his close cabin; fog cruelly pinching the toes and fingers of his shivering little ’prentice boy on deck. Chance people on the bridges peeping over the parapets into a nether sky of fog, with fog all round them, as if they were up in a balloon, and hanging in the misty clouds.

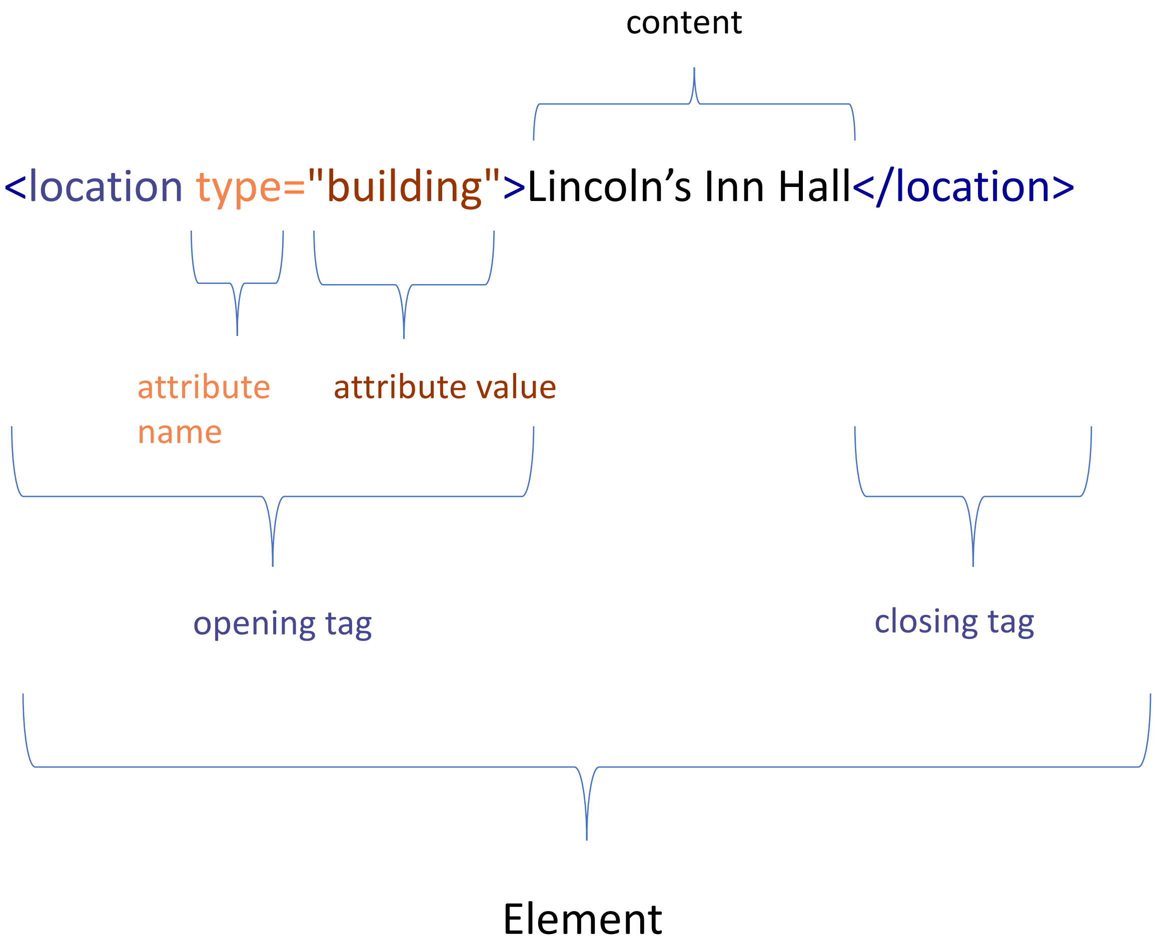
Michaelmas Term lately over, and the Lord Chancellor sitting in <location type="building">Lincoln’s Inn Hall</location>.

XML Anatomy

I’ve tagged the text *Lincoln’s Inn Hall*. The opening tag and the closing tag mark the limits. I’ve added more information with an attribute name and attribute value (which go inside the opening tag). Whenever you have an attribute name, you must have an attribute value, set off from the name with an = symbol and double quotation marks:

Michaelmas Term lately over, and the Lord Chancellor sitting in <location type="building">Lincoln’s Inn Hall</location>.

XML Anatomy



content

<location type="building">Lincoln’s Inn Hall</location>

attribute name

attribute value

opening tag

closing tag

Element

N.B. Empty elements are those without content. Instead of writing an opening and closing tag with no content in between them, we write a single tag with the forward slash to the right of the element name (e.g. <lb></lb> can be represented as <lb/> to indicate a line break.)

Attributes vs Elements

```
<location type="building" borough="camden" material="brick" constructed="1489">Lincoln’s Inn Hall</location>
```

```
<location><borough place="Camden"><consDate year="1489"><redbrick>Lincoln’s Inn Hall</redbrick></consDate></borough></location>
```

XML Rules

- XML documents must be *well formed*
- XML documents must be *valid*
- XML elements must *nest*, and *never overlap*
- XML documents must have a single *root element* and can be expressed as a hierarchy, or tree.

Well formed XML

- In XML, all text must be delimited (nothing hangs around outside the root element)
- For an XML document to be well formed, there can be no overlap.

These examples are not well formed. Why?

Example 1:

```
<lg type="stanza">
   <l><s>A bird came down the walk</l>
   <l>He did not know I saw</s></l>
   <l><s>He bit an angleworm in half</l>
   <l>And ate the fellow raw </s></l>
</lg>
```

These examples are not well formed. Why?

Example 2:

It <verb infinitive="to be">was<verb> a <adjective>dark</adjective> and stormy</adjective> night.

Valid XML

- A valid XML document uses correct vocabulary—only includes elements and attributes specified in a schema
- A valid XML document uses correct grammar—the elements are in the right place, in the right order

Validity:

Rules for letters

- A letter must begin with a date, followed by a salutation, at least one paragraph and a signature
- A date may contain transcribed text
- A paragraph may contain names and transcribed text
- A signature may contain names and transcribed text
- A salutation may contain names and transcribed text

```
<letter id="l1">
<date>2012-02-12</date>
<salutation>Dear Harry,</salutation>
<paragraph>Symmetrical dates are so
elegant!</paragraph>
<signature>Yours, Larry</signature>
</letter>
```

Validity:

Rules for letters

- A letter must begin with a date, followed by a salutation, at least one paragraph and a signature
- A date may contain transcribed text
- A paragraph may contain names and transcribed text
- A signature may contain names and transcribed text
- A salutation may contain names and transcribed text

```
<letter id="l2">
<salutation>Dear
        <name>Larry</name>,
</salutation>
  <paragraph>Dates just reveal
your enslavement to the space-time
continuum.</paragraph>
<signature>Yours,
        <name>Harry</name>
</signature>
</letter>
```

Validity:

Rules for letters

- A letter must begin with a date, followed by a salutation, at least one paragraph and a signature
- A date may contain transcribed text
- A paragraph may contain names and transcribed text
- A signature may contain names and transcribed text
- A salutation may contain names and transcribed text

```
<letter id="l3">
<salutation>Dear
<name>Harry</name>,
</salutation>
<paragraph>My triskaidekaphobia
is acting up.</paragraph>
<signature>Yours, Larry</signature>
<date>2012-02-13</date>
</letter>
```

Validity:

Rules for letters

- A letter must begin with a date, followed by a salutation, at least one paragraph and a signature
- A date may contain transcribed text
- A paragraph may contain names and transcribed text
- A signature may contain names and transcribed text
- A salutation may contain names and transcribed text

```
<letter id="l4">
<date>2012-02-14</date>
<salutation>Dear Larry,</salutation>
<paragraph>Happy
<name>Valentine</name>'s
Day!</paragraph>
<paragraph>I was just kidding about
the space-time
continuum.</paragraph>
<paragraph>You know
who...</paragraph>
</letter>
```

Validity:

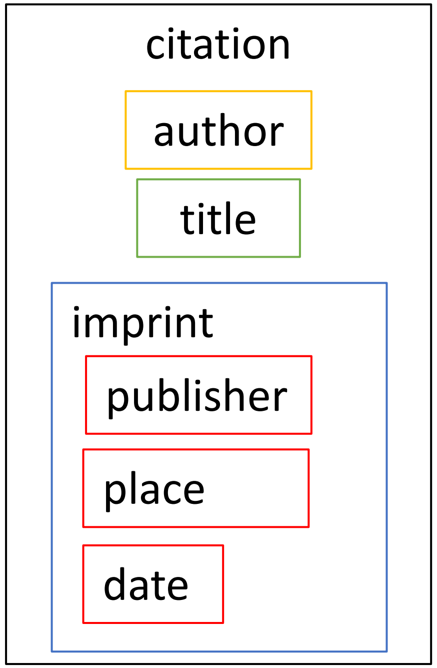
Rules for letters

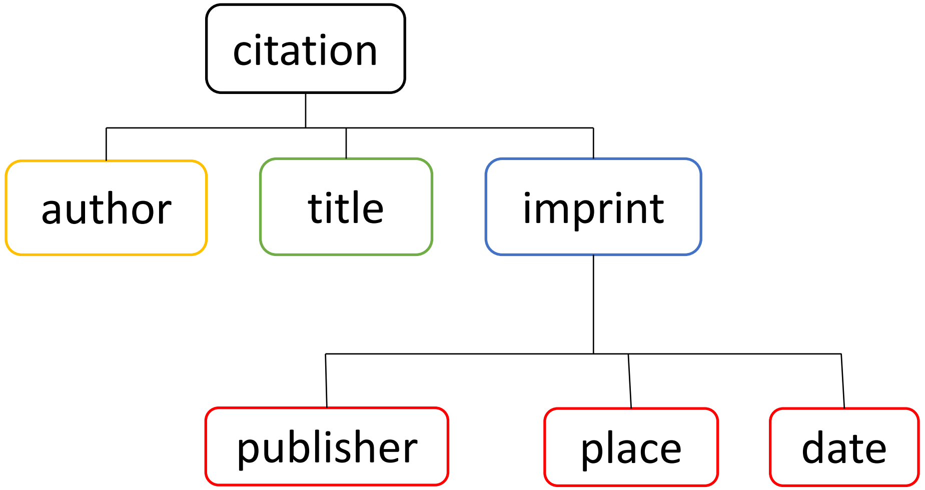
- A letter must begin with a date, followed by a salutation, at least one paragraph and a signature
- A date may contain transcribed text
- A paragraph may contain names and transcribed text
- A signature may contain names and transcribed text
- A salutation may contain names and transcribed text

```
<letter "l5">
<paragraph>Ack! I forgot! Is <date>
14 Feb.</date> always Valentine's
Day? I should put it on my
calendar</paragraph>
</letter>
```

Elements, nesting, one root

- XML documents have elements, attributes and values.
- All XML-based languages’ elements nest, and never overlap
- XML documents have a single root element and can be expressed as a hierarchy, or a tree.





```
<?xml version="1.0" encoding="UTF-8"?>
<citation>
   <author>Charles Dickens</author>
   <title>Bleak House</title>
   <imprint>
      <publisher>Wordsworth
Classics</publisher>
      <place>Ware</place>
      <date>1993</date>
   </imprint>
</citation>
```