XML

An Introduction

HCL

What is XML ?

- XML stands for EXtensible Markup Language.
- XML is a markup language much like HTML.
- XML was designed to describe data (data is embedded between tags that describe it)
- XML is a cross-platform.

HCL

Example

```
<?xml version="1.0" ?>
<priceList>
    <coffee>
        <name>Mocha Java</name>
        <price>11.95</price>
    </coffee>
    <coffee>
        <name>Espresso</name>
        <price>12.50</price>
    </coffee>
</priceList>
```

HCL

Uses

- XML is used to Exchange Data
- XML can be used to Share Data
- XML can be used to Store Data
- XML can be used to Create new Languages

HCL

XML applications of today

a) WML(Wireless markup language)

b) MathML(Mathematical Markup Language)

c) XHTML

d) XML-RPC

e) EDI (Electronic data interchange)

f) XML document in Web services, deployment descriptors in enterprise application etc.

HCL

Origin

- XML and its related technologies are developed and approved by W3C.
- Released in December 1997.
- SGML (Standard Generalized Markup Language by IBM) was the first language that was used to describe data.
- XML is successor to SGML, simplified and adapted to internet.
- XML as been used to define successor of HTML called XHTML.

HCL

XML-Related Components

- Namespace: used to overcome clashing names of tags
- DTD (document type definitions): gives specifications for the tags in XML
- XML Schema: an alternatives to DTD
- XML parser
- XPath, XLink, XPointer: used for navigating and linking

HCL

API for XML

- A software is written to check if the XML document is well-formed and extract the information between XML tags.
- APIs have been developed in C, C++, java and other languages that help in creating, reading and manipulating XML documents.
- XML tags are not predefined. You must define your own tags for your application.

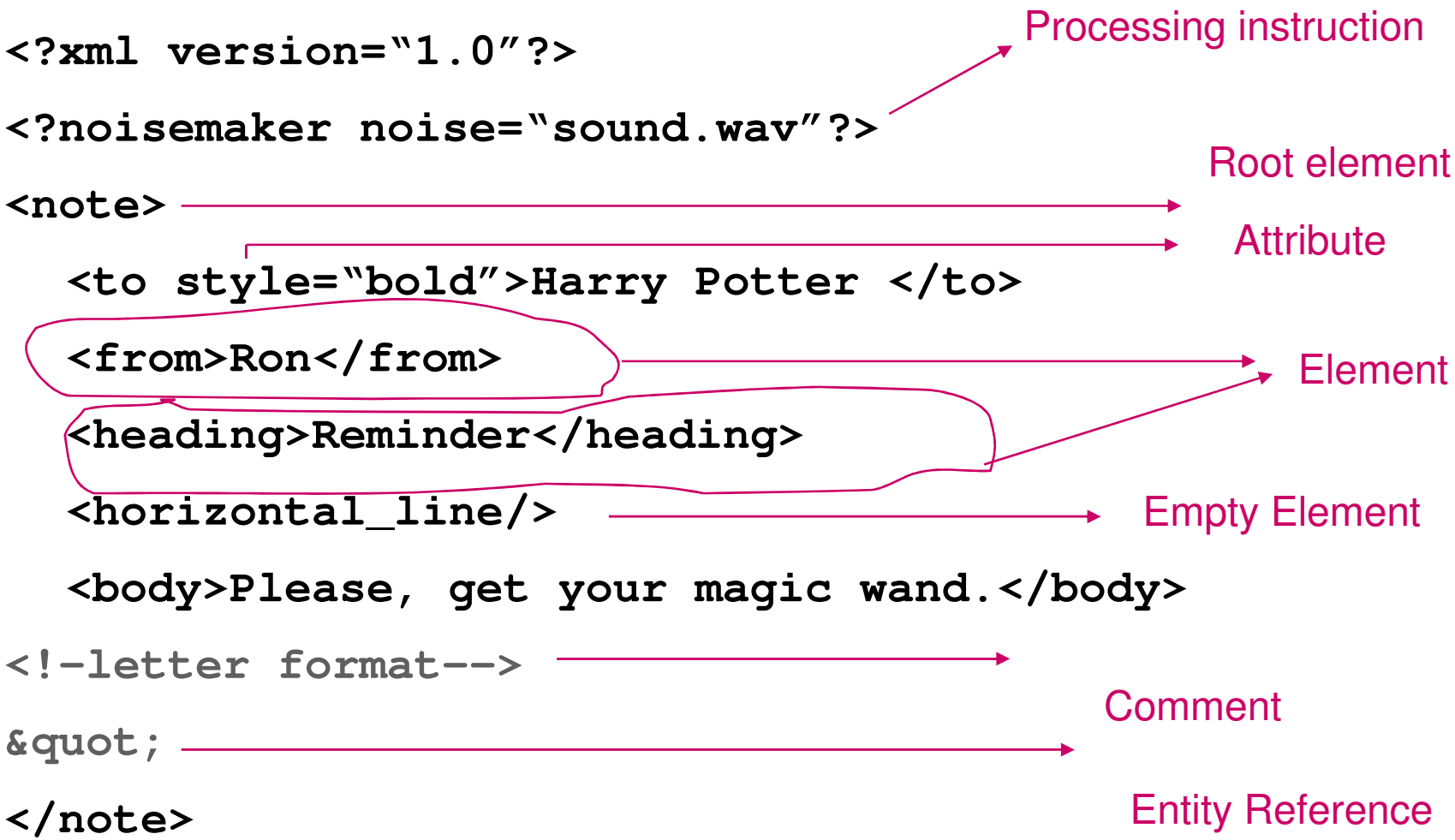
HCL

Tools

We are going to use this

- XPontus :
  - XML Editor that can perform validation(DTD, XML Schema, Relax NG, Batch XML validation), XSL transformations(HTML, XML, PDF, SVG), schema/DTD generation, XML/DTD/HTML/XSL code completion, code formatting.
  - Open source
- XMLSPy
- Oxygen XML
- Exchanger XML Editor

XML document structure



Special markup characters

- <
- >
- &
- ‘
- “

- Use < for <
- Use > for >
- Use & for &
- Use ' for ‘
- Use " for “

Redefining XML

- An XML document is an information unit that can be viewed in two ways:
  - as a linear sequence of characters that contain character data or markup or entity references
  - or as an abstract data structure that is a tree of nodes.

HCL

Well-formed Constraints

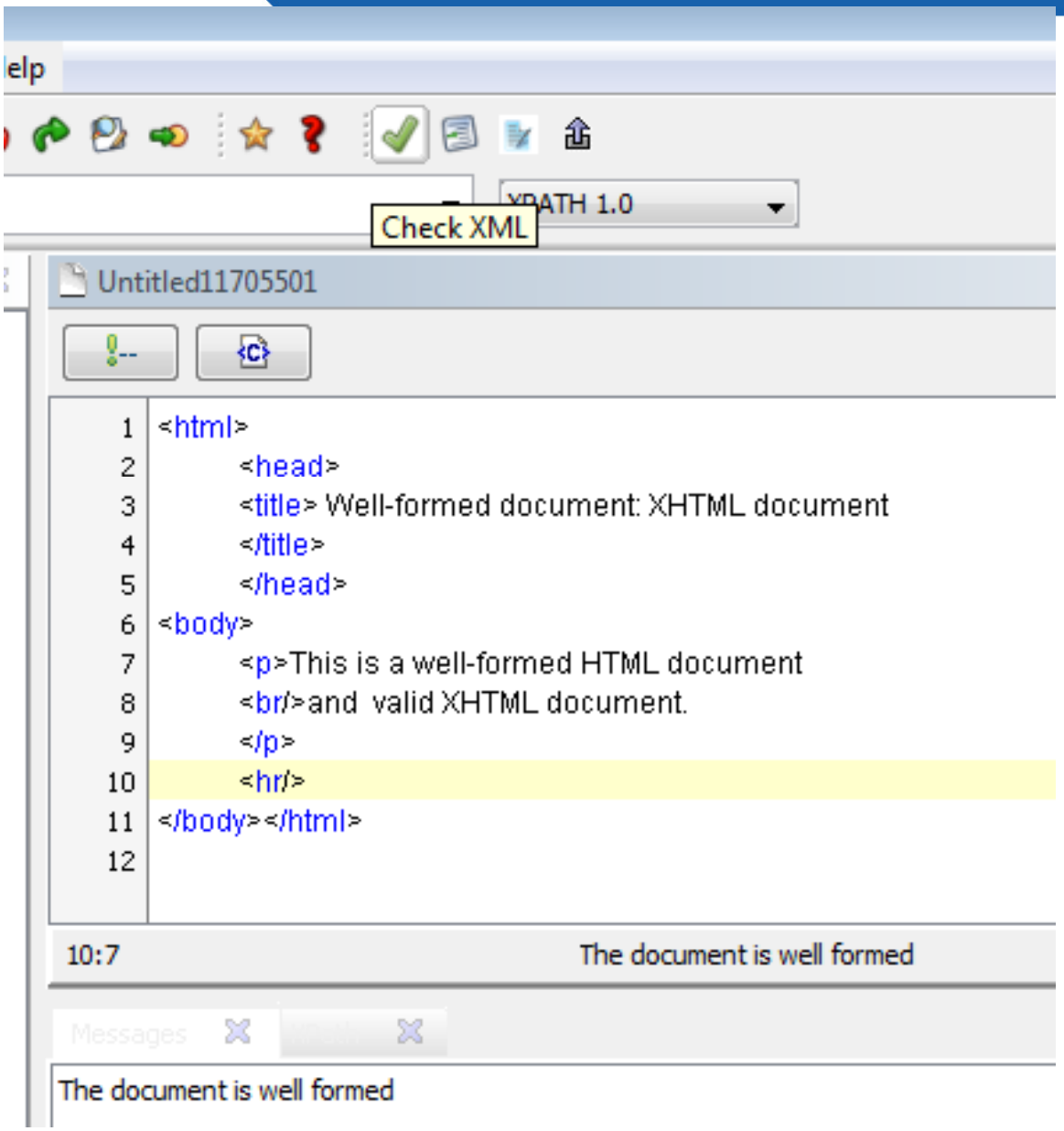
- All XML elements must nest correctly.
- XML tags are case sensitive. The case of the start tag and its corresponding end tag must match.
- All XML elements must be properly nested
- All XML documents must have one and only one Root element
- All the elements (other than the root)must have one and one parent.
- Attribute values must always be quoted
- Empty tags must end with a ‘/’.
- An XML document that confirms to the above rules is called a “Well formed” XML document

HCL

Well-formed XML defined

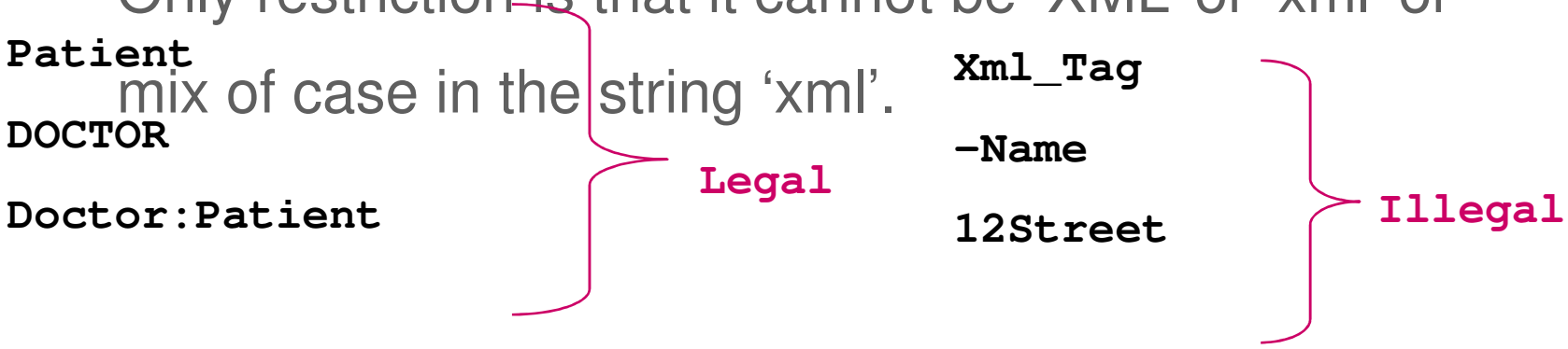
Well-formed XML data conforms to the XML syntax specification, and includes no references to external resources (unless a DTD is provided). It is comprised of elements that form a hierarchical tree, with a single root node (the document element).

Example: XHTML



More about the XML names

- XML names are names given for elements and attributes
- All XML names must begin with a letter or '_' or :.
- Letter could be any alphabets in English or any language supported by UNICODE.
- Only restriction is that it cannot be 'XML' or 'xml' or mix of case in the string 'xml'.



Element

- Basic building block of the XML document
- May have

  - Character data
  - Attributes
  - Character references
  - Entity references
  - Comments
  - PIs
  - CDATA section

- The root element is also called the Document Element.

Attributes

- Attributes are used to attach the information about the element.
- Attribute is a name-value pair
- Attribute values can be any text, entity reference or character reference.
- Attribute values cannot contain special characters.
- Only one instance of attribute name is allowed.

HCL

Character references

- Characters that cannot be typed into a document straight away but must be displayed, can be represented as character references.
- Example: copy right symbol: ©, ®
- **<special> © Worldcom Pvt (India) Ltd</special>**
- Used for representing a single character.
- It is comprised of a decimal or hexa-decimal number between **`&#'** and **`;'**

HCL

Entity references

- 5 built in entity references **`<`** **`>`** `etc.`
- Apart from these 5 entities, number of other entity references are also defined like

  `&copy;` **` `** etc.

HCL

CDATA section

- Character data that you don’t want to be parsed can be kept in CDATA section.

```
<code>
<![CDATA[
      if(a>b && a<10) doThis();
]]>
</code>
```

Comment and PI

- Comment can be given between `<!--` and `-->`
- Example: `<!-- this is a comment -->`
- Processing instruction is used to pass some hints/files to the application along with the xml document.
- PI is given between two '?'
- Example:

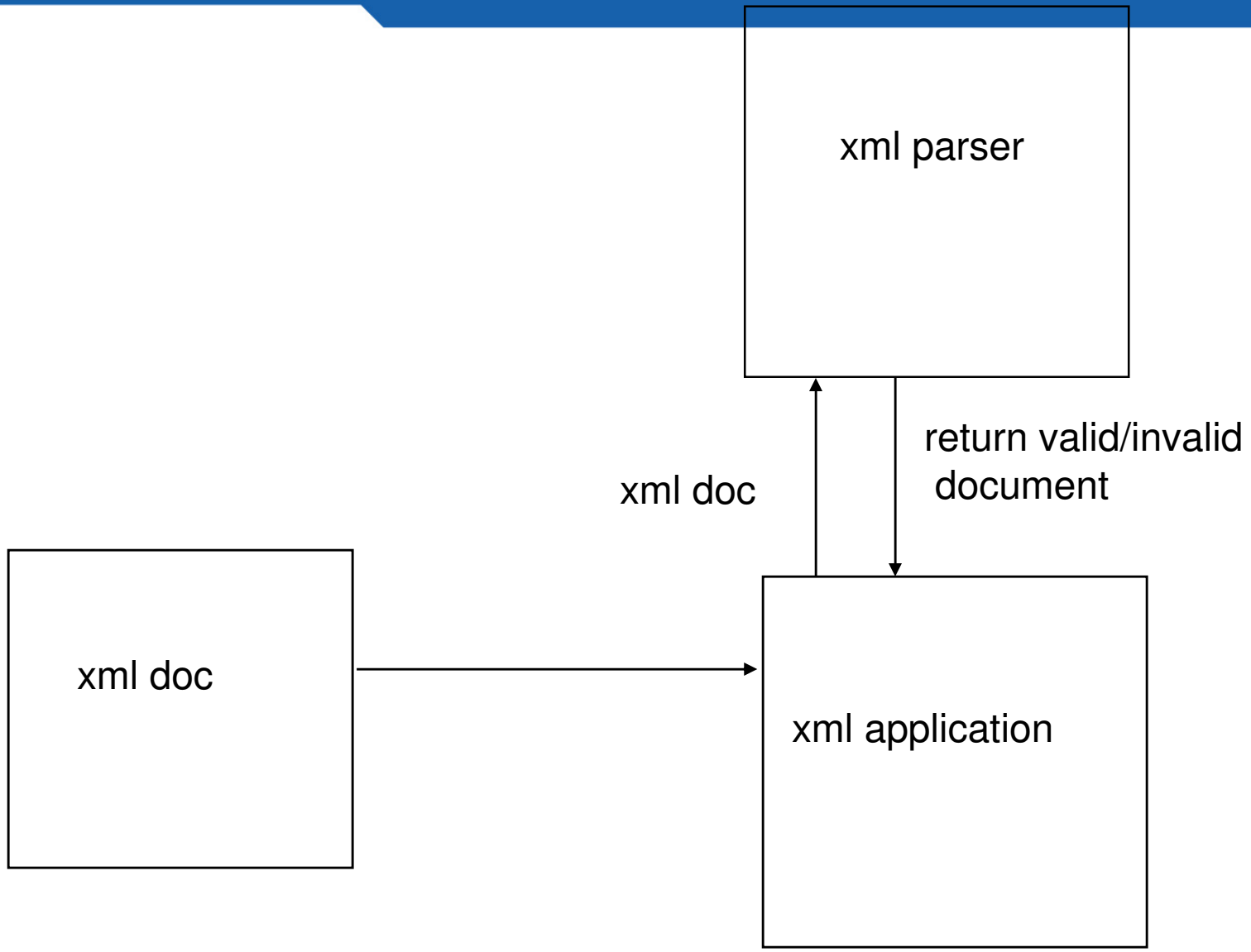
```
<? xml-stylesheet href="mystyle.css" type="text/css" ?>
```

HCL

XML Parser

- XML parsers/processors check if the XML document is well-formed or valid
- Non-validating parser: ensure that the XML document is well-formed.
- Validating parser: ensure that the XML document is
  - Well-formed
  - Valid
  - Resolves external resources

HCL


xml parser
return valid/invalid
document
xml doc
xml doc
xml application

XML Parser available

- Apache
  - Xerces-C(C++)
  - Xerces-J(Java)
- IBM
  - IBM 4C(C++)
  - IBM4J (Java)
- Microsoft
  - MSXML
  - IE
- Oracle
  - XML Parser for Java
  - XML parser for C and C++
- Sun
  - JAXP and JAXB API

HCL