XML

This worksheet on document analysis accompanies the third presentation in the XML collection. It gives a summary of the DTD language syntax.

Linked to presentation: working with well-formed documents
http://humbox.ac.uk/3116

DTD Language Syntax Summary

Declaration

Internal
<!DOCTYPE *name of document type* [*definition of elements attributes and entities goes here*]>

System - External file on local system, network or intranet or the Internet
<!DOCTYPE *name of document type* SYSTEM "*path to document definition fil*"">
Path can be just the file name if its in the same directory, must commence with the word 'file' if it points to a network path or http:// for intranet of Intranet.

Public – This is a direction to a Formal Public Identifier (FPI) following ISO 9070 Rules
<!DOCTYPE *name of document type* PUBLIC "*FPI path" "physical path usually http://2*>
Used for public standard DTDs eg XHTML
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">

Elements

Declaration
<!ELEMENT *elementname* (*content of element*)>

Data element
<!ELEMENT *elementname* (#PCDATA)>
PCDATA is Parsed Character data – this means that XML processors will try to pares it so if you want to include eg '>20' you have to use the entity for '>' ie &gt:

Spaces (aka Whitespace) allowed.

Specials

Empty element
<!ELEMENT *elementname* EMPTY>

Any other element
<!ELEMENT *elementname* ANY>
NB Avoid this, it's too loose!

One or more elements
<!ELEMENT *elementname* (*element1* , *element2* , *element3*)>
The order is significant

Notations

Sequence
The comma separator ',' means each element MUST appear once only in the order specified

Alternatives
The pipe separator '|' means ONLY one element MUST appear once only

Combining Sequence and Alternatives
<!ELEMENT *elementname* (*elementA* | (*element1* , *element2* , *element3*))>
Means EITHER elementA OR element1 and element2 and element3

Occurrence
How many times an element must appear. Default is – MUST appear ONCE ONLY.

? – element MUST appear once OR not at all
+ – element MUST appear one or more times
* – element MAY appear zero or many times

Mixed content
Data and elements mixed eg <p> in XHTML

<!ELEMENT *elementname* (#PCDATA | *element2* | *element3*)*>
Must use the choice notation |
Must use the * indicator
#PCDATA comes first
No inner content model

Attributes

Declaration
A list of attributes applicable to a particular element
<!ATTLIST *elementname attributename attributype attributevaluedeclaration*>

Attribute types

CDATA	Any text (the default type) this will not be parsed so can include ANY character which will appear literally including eg < > etc, and Whitespace
ID	A unique identifier for the element
IDREF	A unique identifier referencing another uniquely identified element
IDREFS	A list of IDREFS
ENTITY	A reference to an external unparsed entity eg an image or an mp3 file – the ENTITY must be declared – see below
ENTITIES	A list of ENTITIES
NMTOKEN	A name – this means as PC data but no Whitespace and you can't use entities.
NMTOKENS	A list of NMTOKENS
Enumerated list	A list of possible values within brackets() separated by | Must be NMTOKEN type values. <!ATTLIST *elementname attributename* (*value1* | *value2* | *value3*)>

Attribute Value Declarations

Default Value
<!ATTLIST *elementname attributename* (*value1* | *value2* | *value3*) *value2*">
value2 is default

Fixed Value
<!ATTLIST *elementname attributename* CDATA #FIXED *valueZ*>
valueZ is the fixed value for this attribute

Required Value
<!ATTLIST *elementname attributename* CDATA #REQURED>
This attribute is mandatory for the element.

Implied Value
<!ATTLIST *elementname attributename* CDATA #IMPLIED>
This attribute may or may not be used with the element.

Entities

Declaration
<!ENTITY *entityname* "*entitydefinition*">

Default entities
As these are defaults they can be used without defining them in a DTD

&amp;	&
&lt:	<

&gt;	>
&apos;	′
&quot;	″

Character entities
Direct references to Unicode characters are not defined in the DTD
Typing these codes in PCDATA:
&x#00A3; produces £
&x#00A9; produces ©

General entities
These are defined in the DTD and used for simple replacement text
<!ENTITY UCL “University College London”> in the DTD allows
&UCL; types in the document to produce ‘University College London’

Parameter entities
These are used ONLY within DTDs. They create replacement text like general entities but are used to re-use content models within the DTD. They are used extensively in the creation of large-scale DTDs

Declaration
<!ENTITY % *entityname* "*entitycontent*">

Referencing
%*entityname*;