

## Summary of XML EBNF notation

Each rule in the grammar defines a symbol in the format: **symbol** ::= expression

With “::=” meaning “Defined as” and symbols being used in successive definitions to create full definitions.

Whilst expressions consist of various combinations of basic terms, operators and previously defined symbols. Basic terms define a pattern for a string of one or more characters and consist of:

- Literal quoted strings – enclosed by either a single or double quote character (eg: “XXX” or ‘XXX’).
- Unicode character numeric references (ISO 10646) - of the form #xH where H is a hexadecimal integer or #N where N is a base 10 integer (eg: #xF6, #x1F2B2, #67, #65421).
- A character class – enclosed between “[“ and “]” characters and defining one or many possible values for a single character. This may include either one or more literal characters or one or more Unicode character numeric references and may specify ranges by using a “-” character between two characters or character references. In addition a character class may specify any character except for those given by starting the class with “[^”. Examples of character classes are:
  - #x6D or #155 or [m] or “m” - all match the single character “m”
  - [#x61-#x7A] or [#141-#172] or [a-z] – all match the lowercase letters “a” to “z”
  - [A-Za-z] or [#141-#172#x41-#x5A] or [a-z] | [A-Z] – all match all letters in the ASCII range
  - [^a-z] or [^#x61-#x7A] – matches any character that is not in the range “a” to “z”

Operators which may be used in combination with basic terms or expressions in order of precedence are:

- Brackets – enclosing expressions in opening and closing brackets “(“ and “)” alters the order of expression evaluation or can be used to make this order clear.
- Occurrence operators – these are postfix operators which follow an expression or term and define how many times this may occur. They are “?” for a single occurrence which is optional, “\*” for zero or many occurrences and “+” for one or more. If no occurrence operator is used then the expression or term must occur once.
- But Not operator – this is the infix operator “-” and defines a result which matches the left hand side expression but does not match the right hand side expression.

- Concatenation – this is an infix operator implied by space(s) between expressions meaning expressions applied in sequence.
- Or operator – this is the infix operator “|” meaning a result which matches either expression.

Examples of expressions are:

- ( [ñ-~] | [0-9] )+ – matches one or more characters which are Thai characters or ASCII digits.
- ( “w” | “t” ) “alk” ( “ing” | “ed” | “s” ) ? - matches any of the words walk, walking, walked, walks, talk, talking, talked or talks (The trailing “?” operator make the alternative endings optional).
- [#x21-#xFE] – [0-9] – matches any printable ASCII character except digits.

Refer to [Extensible Markup Language \(XML\) 1.1 – Notation](#)