XML encoding

XML documents

- can contain international characters, like Norwegian æøå, or French êèé. To avoid errors, you should specify the encoding used, or save your XML files as UTF-8. Character Encoding Character encoding defines a unique binary code for each different character used in a document. In computer terms, character encoding are also called character set, character map, code set, and code page.

- The Unicode Consortium The Unicode Consortium develops the Unicode Standard. Their goal is to replace the existing character sets with its standard Unicode Transformation Format (UTF).

- The Unicode standard is also supported in many operating systems and all modern browsers. The Unicode Consortium cooperates with the leading standards development organizations, like ISO, W3C, and ECMA.

- The Unicode Character Sets Unicode can be implemented by different character sets. The most commonly used encodings are UTF-8 and UTF-16. UTF-8 uses 1 byte (8-bits) to represent basic Latin characters, and two, three, or four bytes for the rest.

- UTF-16 uses 2 bytes (16 bits) for most characters, and four bytes for the rest. UTF-8 = The Web Standard UTF-8 is the standard character encoding on the web.

- UTF-8 is the default character encoding for HTML5, CSS, JavaScript, PHP, SQL, and XML. XML Encoding The first line in an XML document is called the prolog: <?xml version="1.0"?> The prolog is optional.

- Normally it contains the XML version number.

- It can also contain information about the encoding used in the document. This prolog specifies UTF-8 encoding: <?xml version="1.0" encoding="UTF-8"?> The XML standard states that all XML software must understand both UTF-8 and UTF-16.

- UTF-8 is the default for documents without encoding information. In addition, most XML software systems understand encodings like ISO-8859-1, Windows1252, and ASCII XML Errors Most often, XML documents are created on one computer, uploaded to a server on a second computer, and displayed by a browser on a third computer.

- If the encoding is not correctly interpreted by all the three computers, the browser might display meaningless text, or you might get an error message. For high quality XML documents, UTF-8 encoding is the best to use. UTF-8 covers international characters, and it is also the default, if no encoding is declared. Conclusion When you write an XML document: Use an XML editor that supports encoding