Encodings or "Why do I see strange characters?"

When special characters such as "Umlaute" or french accents don't look right, you almost certainly are having character encoding problems.

What is an encoding?

In the end, computers are dealing with bits and bytes, i.e. binary coded numbers only. In order to store, process, or display text, there needs to be some convention on how to map each character (letter, digit, special character) to a number and vice versa. Such a number <-> character mapping is called an encoding. So, for a computer, a text file is just a row of numbers that by accident represents text in some specific encoding.

Why is it so complicated?

In the early computer times, there existed just one single mapping between numerical values in the range of 0..127 and letters (small and capital), digits, and some special characters, the well-known ASCII encoding. As it was defined in the USA, the set of all 128 representable ASCII characters does not contain European characters such as "Umlaute" or French accents, let alone Chinese characters. Therefore, it was necessary to extend the ASCII standard or replace it by new encodings to be able to represent all important characters of some specific area or even the whole world. Unfortunately, there exist many different encodings even to represent the same set of characters. They have names like ISO-8859-1, UTF-8, or Windows-1252. At least, most of the encodings use the good old ASCII mapping for the 128 ASCII characters.
What makes things worse: in a simple file, even if it is known to represent text, there is no information on the encoding used. So this has to be known in advance or guessed in some way...

How about XML files?

XML files are regular text files as well that can be stored in different encodings. Fortunately, the XML designers introduced the means for identifying the encoding used. It may be specified in the first line using the `encoding="..."` entry. If this entry is missing, UTF-8 is assumed as the default. Note that the first line itself always needs to be UTF-8 encoded. XML tools look at the given encoding specification to decide how the file content should be interpreted. Therefore, to get things correct, the specified encoding has to match the encoding that was effectively used to write the file.

So, how do I create and check XML files?

This topic is unfortunately very much tool and operating system dependent. The following tips might help you:

- Checking XML file encodings is quite simple if you have a modern web browser such as Internet Explorer 6 or Mozilla Firefox. These browsers are able to understand and display XML files in their specified encoding. So just open your declaration in the browser and check if the special characters are ok.
- Many non XML-aware editors such as MS Notepad or Wordpad try to guess the encoding from the file to be opened. Even though this method often works for normal text files, it makes no sense for correct XML files and even masks encoding problems of incorrect ones. So don't rely on this feature.
- The best way to create correct XML files are XML editors like Morphon, XML Spy or Stylus Studio. They handle encodings correctly.
- Even with simple editors you can create correctly encoded XML files. You just have to know which encoding is used by the tool to save the file and set the same encoding in the XML header. In Notepad, you can choose between UTF-8 and "ANSI" (which in this case means Windows-1252). So if you plan to save the document in UTF-8, e.g., enter the entry `encoding="UTF-8"` in the XML header. Since the default text file encoding in Windows is Windows-1252, you can try out this one if your Windows editor does not offer any choices for file encodings. On Unix systems, ISO-8859-1 is the most likely text file encoding.

- If you create XML files programmatically, you have to keep in mind that each operation that deals with the file and its contents potentially has to cope with the encoding issue. A single operation (e.g. the final save) might inadvertedly temper with the encoding and result in a corrupt result.
- Be careful with console tools (such as vi on Unix). In addition to the editor, the terminal needs to interpret the encodings correctly.

Links

- Encoding in XML
- List with websites about encodings
- Encoding tutorial
- Wikipedia: Introduction to UTF-8

Dominik Auf der Maur, July 2005