Advanced Topics: Unicode and XSL

SET09103 Advanced Web Technologies

School of Computing
Napier University, Edinburgh, UK
Module Leader: Uta Priss

2008

Outline

Unicode

Unicode Programming

XSL

What is Unicode

Unicode is an international standard for representing characters independently of

- the platform
- the program
- the language

It provides a unique number (code point) for each character.

Non ASCII Characters:

$\alpha\beta\gamma$...

My name: Uta Priß

Café, København

Characters from other languages: Chinese, Arab, Hebrew, ...

Other graphic characters: , . ? ! % ♣♢♡♠

Character versus renderings

The same Unicode character in different fonts or renderings:

Characters are represented abstractly as a number (code point):

U+0041 U+4E94

Unicode facts

- 65536 chars in Basic Multilingual Plane 0000-FFFF.
- First 256 code points are identical to ISO 8859-1 (this includes the 128 ASCII characters).
- For compatibility reasons: some duplicate characters exist.
- Unicode is widely, internationally accepted, but there are some cultural issues and difficulties.
- Unicode covers basically all modern scripts and also many historical ones. More will be added.

Unicode and XML/HTML

- XML supports Unicode.
- HTML/XML:
  - Either use bytes according to document encoding, i.e. binary representation of U+0041, U+4E94 etc.
  - Or numeric character references (in ASCII), i.e. A 五
- in URIs: non-ASCII characters must be percent-encoded.

Numeric character references can be used in XML documents if there is any doubt about whether the tools that are used with the documents support Unicode.

Unicode Transformation Format (UTF)

UTF maps code points to code values.

- UTF-8: 8 bits per code value
  - 1 byte for all ASCII characters (using the same code points)
  - up to 4 bytes for other characters
  - used internally in Unix
- UTF-16: 16 bits per code value
  - variable-width encoding
  - used internally in Windows, Mac, KDE and Java

Recommended use for email is UTF-8 or Base64

Programming with Unicode

"Modern programming languages support Unicode."

- That means: Unicode can be used in strings.
- Although some programming languages may allow the use of Unicode in variable names etc, it is probably NOT a good idea to do so at this point in time because of compatibility issues.

String processing

- What is a new line (\n, \r, \r\n)?
- What are word characters?
- What is a word boundary?
  (For example: foreign language punctuation marks: ○, ¡ , ¿)
- Search engines:
  does searching for Kobenhavn retrieve København?
- What is an "alphabetical ordering"?

Using Unicode in MySQL

MySQL supports Unicode since version 4.1.

```
CREATE TABLE example (
      id INTEGER PRIMARY KEY,
      unicodeText VARCHAR(50));
CHARACTER SET utf8 COLLATE utf8_general_ci;
SET NAMES 'utf8';
```

COLLATE determines how the data is sorted (for ORDER BY).

Using Unicode in PHP

PHP 5 supports UTF-8 natively.

At the beginning of the file:

```
<?php echo '<?xml version="1.0" encoding="utf-8"?>';
?>
```

Convert from HTML entity to UTF-8:

```
print html_entity_decode("&#9730;", ENT_QUOTES, 'UTF-8')
```

Convert from UTF-8 to HTML entity:

```
print htmlentities("'", ENT_QUOTES, 'UTF-8');
```

Note: general Unicode conversion is available in PHP 6 using unicode_decode() and unicode_encode()

Using Unicode in Perl

Perl 5.8+ has comprehensive support for Unicode.

Opening a file:

```
open (FILE, "<:utf8","$filename");
```

Convert from UTF-8 to numeric character references:

```
use Encode;
$line = encode("ascii", $line, Encode::FB_XMLCREF);
```

Unicode and script security

User submitted Unicode may require special security checks.

- Unicode characters can be control characters, etc.
- One security strategy is to convert non-ASCII characters into numeric character references. But these contain semicolons.
- In case of database and shell access, semicolons pose a security risk. Thus more checks and tests are required.

Extensible Stylesheet Language (XSL)

A family of transformation languages:

- XSL Transformations (XSLT):
  for transforming XML documents.
- XSL Formatting Objects (XSL-FO):
  for specifying visual formatting.
- XML Path Language (XPath):
  a non-XML language for selecting nodes.
  Part of XSLT.

XSLT

- Declarative language
  similar to functional languages or text processing languages (e.g. awk).
- For example, for converting between different XML schemas or between XML, HTML, and XHTML.
- XML source document + XSLT stylesheet $\Longrightarrow$
  output document.
- XSLT is Turing complete,
  i.e., equivalent to other programming languages.

XSLT example

Part of an XML document:

```
<food><ingredient amount='5'>eggs</ingredient></food>
```

XSL document:

```
<?xml version='1.0' ?>
<xsl:stylesheet version='1.0'
    xmlns:xsl='http://www.w3.org/1999/XSL/Transform'>

    <xsl:template match='ingredient'>
      <xsl:value-of select='@amount'/>
    </xsl:template>
</xsl:stylesheet>
```

XSLT example

Part of an XML document:

```
<food><ingredient amount='5'>eggs</ingredient></food>
```

XSL document:

```
<?xml version='1.0' ?>
<xsl:stylesheet version='1.0'
    xmlns:xsl='http://www.w3.org/1999/XSL/Transform'>

    <xsl:template match='ingredient'>
      <xsl:value-of select='@amount'/>
    </xsl:template>
</xsl:stylesheet>
```

Output: 5

XSL-FO (Formatting Objects)

- XSLT is used to translate XML into XSL-FO.
- FO processor then generates PDF (or PS, RTF, etc.) from XSL-FO.
- Possibilities for automatic generation of table of contents, linked references, index, etc.
- For printed, page-based media
  (in contrast to HTML which is for screen media).

XPath

- Language for selecting nodes.
  (XPath is for XML what SQL is for databases.)
- Navigates the XML tree structure.
- The abbreviated syntax is similar to Unix path notation.

XPath examples

Part of an XML document:

```
<product>
    <food>
        <ingredient amount='5'>eggs</ingredient>
    </food>
</product>
```

XPath expressions:

```
/product/food/ingredient
/product/food/ingredient/@amount
/product//ingredient
/product/food/ingredient[@amount='5']/text()
```