Unicode, XSL, and HTML5

Web Programming

Uta Priss
ZELL, Ostfalia University

2013

Outline

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()
```

HTML: history

- "HTML tags", Berners-Lee, first mentioned in 1991
- HTML 2, 1995 - 1997 (basic tags, forms, tables, image maps)
- HTML 3, 1995 - 1997 (browser-specific tags, browser wars)
- HTML 4, 1997 - 2000 (stylesheets instead of visual markup)
- XHTML 1, 2000 - 2001 (XML precision)
- HTML 5 and XHTML 5, 2008 -

HTML5: Deprecated features

- font, center, dir, big (to be replaced by stylesheets)
- presentational attributes: align, etc (to be replaced by stylesheets)
- frame, frameset (replaced by iframe, PHP/Ajax ?)
- applet (replaced by object)

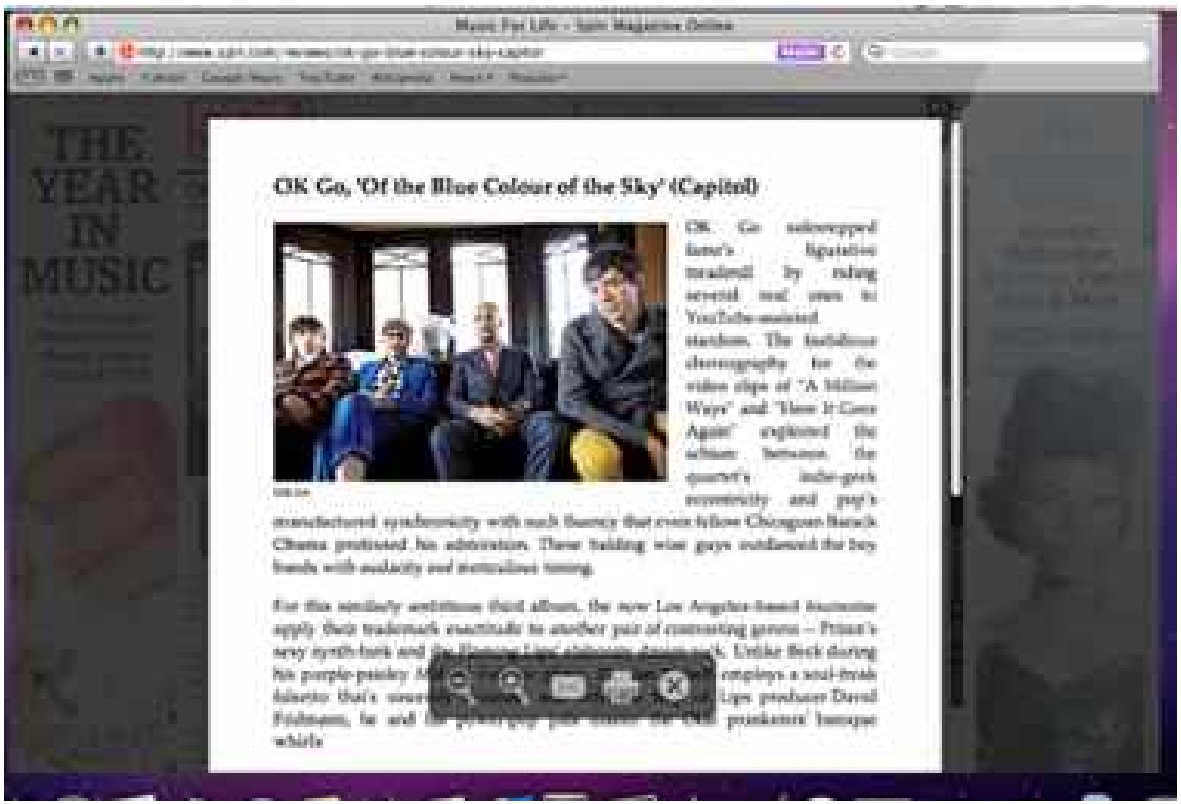
HTML5: New features

- new elements: audio, video, canvas, figure
- new form controls: date, time, url, email, search
- structuring: article, section, footer, details, summary, nav
- web storage
- DOM scripting

HTML5: New APIs

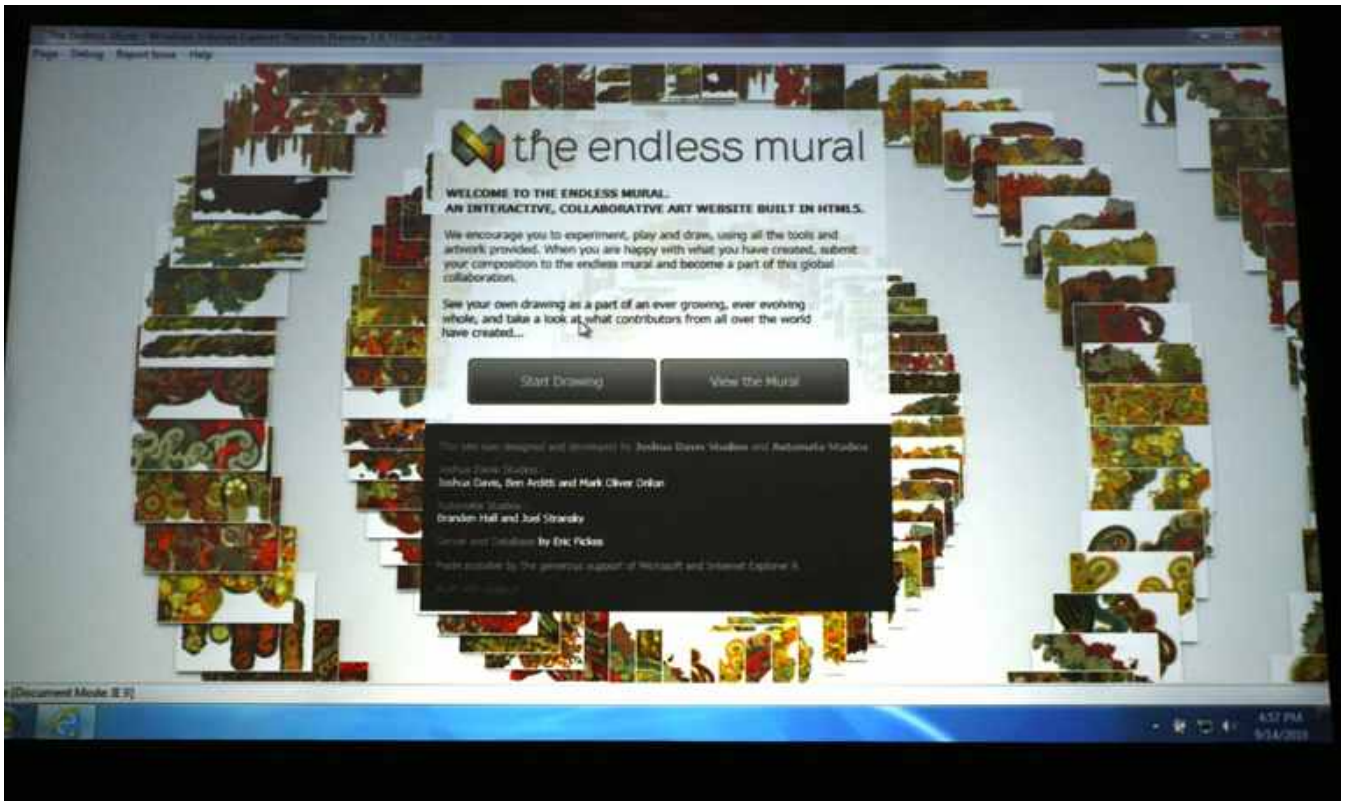
- geolocation
- playing audio and video
- offline web applications
- drag and drop
- history (reading the history, control of the back button)

Example: article tag in Safari Reader



`http://www.apple.com/safari/whats-new.html`

Example: collaborative drawing (IE9)



`http://www.microsoft.com/presspass/presskits/internetexplorer/ImageGallery.aspx`

Web Storage and DOM Storage

- persistent, client-side storage
- modern version of “cookies”
- client-side scripting (Javascript)
- local and session storage (per-page-per-window)
- associative array, hash

Web SQL Database

Similar to Web Storage, but uses an SQL database (SQLite).

Accessed via Javascript.

At the moment: not supported by all browsers.

How much control does the user have?
(e.g. turn off database storage, determine size)

SQLite

```
:)sqlite3 search.sqlite
SQLite version 3.6.22
Enter ".help" for instructions
sqlite> .tables
engine_data
sqlite> select * from engine_data;
8|[profile]/sourceforge.xml|order|8
10|[app]/google.xml|order|1
11|[app]/yahoo.xml|order|2
12|[app]/amazondotcom.xml|order|3
13|[app]/answers.xml|order|4
14|[app]/creativecommons.xml|order|5
15|[app]/eBay.xml|order|6
16|[app]/wikipedia.xml|order|7
17|[app]/google.xml|used|0
sqlite> .quit
```

Firefox: Indexed Database API

SQLite used for cookies, downloads, permissions, search, etc

JSON used for bookmarkbackups, search, etc

RDF used for localstore

Kamkar: “Evercookie”

Creating a cookie that cannot be deleted.

Using flash cookies, Silverlight cookies,
and three different types of HTML5 storage.

Kamkar: "Evercookie"

Creating a cookie that cannot be deleted.

Using flash cookies, Silverlight cookies,
and three different types of HTML5 storage.

To delete this cookie in Safari:

Reset, restart + script to delete from folders,
in the databases and in LocalStorage.

In iPhone this requires a jailbreak.

New commercial uses: DoubleClick++

RLDGUID: Ring Leader Digital Globally Unique ID

- made by a mobile advertising company
- globally unique ID!
- this ID cannot (easily) be deleted
- opt out is possibly but cannot be verified

If one of RLDG's customers knows your name
$\Longrightarrow$ all participating websites know
who you are when you visit their site!

From the user's viewpoint

- More functionality (Web 2.0)
- More transparency needed
  for example: "private mode browsing" and LocalStorage?
- Browser extensions may give users more control