Global Design:
Characters, Language, and More

Richard Ishida
W3C Internationalization Activity Lead
Getting more information

W3C Internationalization Activity

http://www.w3.org/International/

Copyright © 2005 W3C (MIT, ERCIM, Keio)
Outline

Character encoding: What's that all about?
Characters: What do I need to do?
Characters: Using escapes
Language: Two types of declaration
Language: The new language tag values
Text size
Navigating to localized pages
Character encoding: What's that all about?
Character encoding

The Enigma

Photo by David Blaikie
Character encoding

Berber

4,000 BC
Character encoding

Tifinagh

http://www.dailymotion.com/video/x1rh6m_tifinagh_creation
Character encoding

Character set
Character encoding

Coded character set

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.</td>
<td>Φ</td>
<td>+</td>
<td>Δ</td>
</tr>
<tr>
<td>1</td>
<td>Θ</td>
<td>Ø</td>
<td>!</td>
<td>′</td>
</tr>
<tr>
<td>2</td>
<td>Φ</td>
<td>!</td>
<td>′</td>
<td>′</td>
</tr>
<tr>
<td>3</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>4</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>5</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>6</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>7</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>8</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>9</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>A</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>B</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>C</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>D</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>E</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
<tr>
<td>F</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
<td>Ω</td>
</tr>
</tbody>
</table>

Coded character set

Code point

33 (hexadecimal)

52 (decimal)
### Character encoding

#### Code pages

<table>
<thead>
<tr>
<th>Code</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>@</td>
</tr>
<tr>
<td>1</td>
<td>P</td>
</tr>
<tr>
<td>2</td>
<td>&quot;</td>
</tr>
<tr>
<td>3</td>
<td>#</td>
</tr>
<tr>
<td>4</td>
<td>$</td>
</tr>
<tr>
<td>5</td>
<td>%</td>
</tr>
<tr>
<td>6</td>
<td>&amp;</td>
</tr>
<tr>
<td>7</td>
<td>'</td>
</tr>
<tr>
<td>8</td>
<td>(</td>
</tr>
<tr>
<td>9</td>
<td>)</td>
</tr>
<tr>
<td>A</td>
<td>:</td>
</tr>
<tr>
<td>B</td>
<td>;</td>
</tr>
<tr>
<td>C</td>
<td>&lt;</td>
</tr>
<tr>
<td>D</td>
<td>=</td>
</tr>
<tr>
<td>E</td>
<td>&gt;</td>
</tr>
<tr>
<td>F</td>
<td>?</td>
</tr>
</tbody>
</table>

*ASCII*
**Character encoding**

**Code pages**

<table>
<thead>
<tr>
<th>Character</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>!</td>
<td>0x00</td>
<td>Space, Exclamation mark</td>
</tr>
<tr>
<td>1</td>
<td>0x01</td>
<td>Upper Case A, Lower Case a</td>
</tr>
<tr>
<td>2</td>
<td>0x02</td>
<td>Upper Case B, Lower Case b</td>
</tr>
<tr>
<td>#</td>
<td>0x03</td>
<td>Number sign, Hashtag</td>
</tr>
<tr>
<td>$</td>
<td>0x04</td>
<td>Dollar sign</td>
</tr>
<tr>
<td>%</td>
<td>0x05</td>
<td>Percent</td>
</tr>
<tr>
<td>&amp;</td>
<td>0x06</td>
<td>Ampersand</td>
</tr>
<tr>
<td>(</td>
<td>0x07</td>
<td>Left parenthesis</td>
</tr>
<tr>
<td>)</td>
<td>0x08</td>
<td>Right parenthesis</td>
</tr>
<tr>
<td>*</td>
<td>0x09</td>
<td>Star, Asterisk</td>
</tr>
<tr>
<td>+</td>
<td>0x0A</td>
<td>Plus, Plus sign</td>
</tr>
<tr>
<td>,</td>
<td>0x0B</td>
<td>Comma</td>
</tr>
<tr>
<td>&lt;</td>
<td>0x0C</td>
<td>Less than</td>
</tr>
<tr>
<td>=</td>
<td>0x0D</td>
<td>Equal sign</td>
</tr>
<tr>
<td>&gt;</td>
<td>0x0E</td>
<td>Greater than</td>
</tr>
<tr>
<td>?</td>
<td>0x0F</td>
<td>Question mark</td>
</tr>
<tr>
<td>A</td>
<td>0x10</td>
<td>Upper Case A</td>
</tr>
<tr>
<td>B</td>
<td>0x11</td>
<td>Upper Case B</td>
</tr>
<tr>
<td>C</td>
<td>0x12</td>
<td>Upper Case C</td>
</tr>
<tr>
<td>D</td>
<td>0x13</td>
<td>Upper Case D</td>
</tr>
<tr>
<td>E</td>
<td>0x14</td>
<td>Upper Case E</td>
</tr>
<tr>
<td>F</td>
<td>0x15</td>
<td>Upper Case F</td>
</tr>
<tr>
<td>G</td>
<td>0x16</td>
<td>Upper Case G</td>
</tr>
<tr>
<td>H</td>
<td>0x17</td>
<td>Upper Case H</td>
</tr>
<tr>
<td>I</td>
<td>0x18</td>
<td>Upper Case I</td>
</tr>
<tr>
<td>J</td>
<td>0x19</td>
<td>Upper Case J</td>
</tr>
<tr>
<td>K</td>
<td>0x1A</td>
<td>Upper Case K</td>
</tr>
<tr>
<td>L</td>
<td>0x1B</td>
<td>Upper Case L</td>
</tr>
<tr>
<td>M</td>
<td>0x1C</td>
<td>Upper Case M</td>
</tr>
<tr>
<td>N</td>
<td>0x1D</td>
<td>Upper Case N</td>
</tr>
<tr>
<td>O</td>
<td>0x1E</td>
<td>Upper Case O</td>
</tr>
<tr>
<td>P</td>
<td>0x1F</td>
<td>Upper Case P</td>
</tr>
<tr>
<td>Q</td>
<td>0x20</td>
<td>Upper Case Q</td>
</tr>
<tr>
<td>R</td>
<td>0x21</td>
<td>Upper Case R</td>
</tr>
<tr>
<td>S</td>
<td>0x22</td>
<td>Upper Case S</td>
</tr>
<tr>
<td>T</td>
<td>0x23</td>
<td>Upper Case T</td>
</tr>
<tr>
<td>U</td>
<td>0x24</td>
<td>Upper Case U</td>
</tr>
<tr>
<td>V</td>
<td>0x25</td>
<td>Upper Case V</td>
</tr>
<tr>
<td>W</td>
<td>0x26</td>
<td>Upper Case W</td>
</tr>
<tr>
<td>X</td>
<td>0x27</td>
<td>Upper Case X</td>
</tr>
<tr>
<td>Y</td>
<td>0x28</td>
<td>Upper Case Y</td>
</tr>
<tr>
<td>Z</td>
<td>0x29</td>
<td>Upper Case Z</td>
</tr>
</tbody>
</table>

**ISO 8859-1 (Latin 1)**

- **Western Europe**
  - LATIN CAPITAL LETTER ETA (E7) Ҫ

---

Copyright © 2005 W3C (MIT, ERCIM, Keio)
<table>
<thead>
<tr>
<th>Code Page</th>
<th>Character Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO 8859-7</td>
<td>Greek</td>
</tr>
</tbody>
</table>

Greek: η (E7)
## Character encoding

### Double-byte characters

<table>
<thead>
<tr>
<th>Standard</th>
<th>Country</th>
<th>No. of chars</th>
</tr>
</thead>
<tbody>
<tr>
<td>KS X 1001:1992</td>
<td>Korea</td>
<td>8,224</td>
</tr>
<tr>
<td>GB 2312-80</td>
<td>Mainland China</td>
<td>7,445</td>
</tr>
<tr>
<td>Big Five</td>
<td>Taiwan</td>
<td>13,494</td>
</tr>
<tr>
<td>CNS 11643-1992</td>
<td></td>
<td>48,711</td>
</tr>
<tr>
<td>JIS X 0208:1997</td>
<td>Japan</td>
<td>6,879</td>
</tr>
<tr>
<td>JIS X 0212-1990</td>
<td></td>
<td>6,067</td>
</tr>
</tbody>
</table>
Character encoding

Unicode

All languages
Character encoding

Tifinagh Unicode block

Coded character set

Code point

2D63 (hexadecimal)
11619 (decimal)
Character encoding

Character encodings

UTF-16

2D30 2D63 2D53 2D4D 21

UTF-8

E2 B4 B0 E2 B5 A3 E2 B5 93 E2 B5 8D 21

UTF-32

00 00 2D 30 00 00 2D 63 00 00 2D 53 00 00 2D 4D 00 00 00 21
Recap

- Character set
- Coded character set
- Code point
- Character encoding

- Charset
Character encoding: What do I need to do?
Step 1: Choose an encoding

- Supports many languages in a single character set so that
  - it's easy to architect multilingual solutions
  - you can have a mix of languages on a single page

- Reduce complexity when storing multilingual data, eg. in databases, and when receiving form data

- Provide all symbols you'll need, eg. © ® ™ etc…

- Wide support in browsers, scripts, editors, databases, operating systems, etc.

The answer is:

Unicode
Step 3: Declaring the encoding

1. In the HTTP header

2. In the document
HTTP/1.1 200 OK
Date: Wed, 05 Nov 2003 10:46:04 GMT
Server: Apache/1.3.28 (Unix) PHP/4.2.3
Content-Location: CSS2-REC.en.html
Vary: negotiate,accept-language,accept-charset
TCN: choice
P3P: policyref=http://www.w3.org/2001/05/P3P/p3p.xml
Cache-Control: max-age=21600
Expires: Wed, 05 Nov 2003 16:46:04 GMT
Last-Modified: Tue, 12 May 1998 22:18:49 GMT
ETag: "3558cac9;36f99e2b"
Accept-Ranges: bytes
Content-Length: 10734
Connection: close
Content-Type: text/html; charset=utf-8
Content-Language: en
Set using

- Server defaults or overrides

Apache directives

- AddCharset UTF-8 .php
- AddType 'text/html; charset=UTF-8' html
- <Files ~ "events\.html">
  ForceType 'text/html; charset=UTF-8'
  </Files>
Working with characters

HTTP header

Set using

- Server defaults or overrides
- Scripting directives

**PHP**
- `header( 'Content-type: text/html; charset=utf-8' );`

**JSP**
- `<%@ page contentType="text/html; charset=UTF-8" %>`

**ASP**
- `<%Response.charset="utf-8"%>`

**Perl**
- `print "Content-Type: text/html; charset=utf-8\n\n";`
HTTP header

- user agents can easily find the information
- highest priority in case of conflict – good for transcoding

- possibly difficult for content authors to change the setting - especially with an ISP
- server settings may get out of synch with the document
- doesn't cater for documents saved to or read from CD or hard disk
- may not facilitate processing or translation
Always declare the encoding inside the document, too!
Working with characters

In-document declarations: HTML

```
<meta http-equiv="Content-type" content="text/html; charset=UTF-8">
```

- In the head element, as near as possible to the top of the file
- Use preferred names from the IANA registry
- Avoid unregistered names (ie. x-… )
Document can be edited or read from CD or hard disk, etc.

Dynamic documents can be saved to disk

Helps developers, testers, or translation production managers perform a visual check

The only in-document alternative
In-document declarations: XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
```

- Required if not UTF-8 or UTF-16!
- First thing in the file
- Useful even with UTF-8 or UTF-16
- Use preferred names from the IANA registry
- Avoid unregistered names (ie. x-… )
Working with characters

In-document declarations: XML

- Useful when processing as XML, e.g. using XSLT
- Document can be edited or read from CD or hard disk, etc.
- Dynamic documents can be saved to disk
- Helps developers, testers, or translation production managers perform a visual check
- The only in-document alternative
Working with characters

XHTML served as text/html

<table>
<thead>
<tr>
<th>Language:</th>
<th>Served with MIME type:</th>
<th>Treated as:</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML</td>
<td>text/html</td>
<td>HTML</td>
</tr>
<tr>
<td>XHTML</td>
<td>application/xhtml+xml</td>
<td>XML</td>
</tr>
<tr>
<td></td>
<td>application/xml</td>
<td></td>
</tr>
<tr>
<td></td>
<td>text/xml</td>
<td></td>
</tr>
</tbody>
</table>
<?xml version="1.0" encoding="UTF-8"?>

and?

or?

<meta http-equiv="Content-type"
content="text/html; charset=UTF-8"/>
<?xml version="1.0" encoding="UTF-8"?>

Use both unless … you need your page to conform to Standards mode on IE6.

<meta http-equiv="Content-type" content="text/html; charset=UTF-8">
Working with characters

**XHTML served as text/html**

---

**Standards**

```
A div with CSS width:170px, margin:20px, padding:50px and border:6px.
```

Text in a `p` element.

```
Text in a table.
```

---

**Quirks**

```
A div with CSS width:170px, margin:20px, padding:50px and border:6px.
```

Text in a `p` element.

```
Text in a table.
```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xml:lang="en" lang="en">
<head>
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
  <title>XHTML document</title>
  <style type="text/css">
    body { background: white; color: black; font-family: arial, sans-serif; font-size: 12px; }
    p { font-size: 100%; }
    h1 { font-size: 16px; }
    div { margin: 20px; width: 170px; padding: 50px; border: 6px solid teal; }
    table { border: 1px solid teal; }
  </style>
</head>
<body>
  <h1>Test file for Standards/Quirks</h1>
  <div>
    A div with CSS width:170px, margin:20px, padding:50px and border:6px.
  </div>
  <p>Text in a p element.</p>
  <table>
    <tr><td>Text in a table.</td></tr>
  </table>
</body>
</html>
Working with characters

**XHTML served as text/html**

```xml
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
 "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xml:lang="en" lang="en">
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
    <title>XHTML document</title>
    <style type="text/css">
      body { background: white; color: black; font-family: arial, sans-serif; font-size: 12px; }
      p { font-size: 100%; }
      h1 { font-size: 16px; }
      div { margin: 20px; width: 170px; padding: 50px; border: 6px solid teal; }
      table { border: 1px solid teal; }
    </style>
  </head>
  
  <body>
    <h1>Test file for Standards/Quirks</h1>
    <div>
      A div with CSS width:170px, margin:20px, padding:50px and border:6px.
    </div>
    <p>Text in a p element.</p>
    <table>
      <tr><td>Text in a table.</td></tr>
    </table>
  </body>
</html>
```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
  <head>
    ......
  </head>
  
  <ul>
    <li>In IE6, prevents application of standards mode</li>
    <li>Choices:</li>
    <ul>
      <li>If no affected features, no worries</li>
      <li>If IE6 irrelevant, no worries</li>
      <li>Use workarounds in CSS</li>
      <li>Remove the XML declaration</li>
    </ul>
    <li>If you omit, use UTF-8 or UTF-16</li>
  </ul>
Step 2: Save your data in the right encoding

It's not enough to label your data with an encoding:

you need the content to be in the right sequence of bytes!
Step 2: Save your data in the right encoding
Step 2: Save your data in the right encoding
Working with characters

Step 2: Save your data in the right encoding

Default encoding: Unicode (UTF-8)

Use when opening existing files that don't specify encoding.
Step 2: Save your data in the right encoding

Unicode Normalization Form: C (Canonical Decomposition, followed by )

Include Unicode Signature (BOM)
Working with characters

Step 2: Save your data in the right encoding

 serão BOM

2D30 2D63 2D53 2D4D 21

FE FF 2D 30 2D 63 2D 53 2D 4D 00 21

FF FE 30 2D 63 2D 53 2D 4D 2D 21 00

{i} in UTF8
Working with characters

Step 2: Save your data in the right encoding

- The image shows a screenshot of a text editor with options for setting encoding and character sets.
- The text editor is configured to use Unicode (UTF-8) encoding by default.
- The preference for including the Unicode Signature (BOM) is enabled in the settings.
Step 2: Save your data in the right encoding

Working with characters
Step 4: Secure the backend

Working with characters

Château de La Napoule

Comments

r12a says:
Just so you know, I have raised a bug about the incorrect character in the title. I left this one as is for the bug to have a reference.

Posted 36 months ago. (permalink | delete | edit)
Some lessons from this section

- Choose an encoding – consider UTF-8 or UTF-16 (Unicode)
- Figure out how to save your data in your chosen encoding.
- Add character encoding declarations to your content.
- Use the meta tag for HTML and the XML declaration for XML.
- Use both for XHTML served as text/html, unless you want to avoid quirks mode on IE6 (in which case I suggest to omit the declaration and use UTF-8)
- Try to avoid getting a Byte-Order Mark (BOM) at the beginning of your UTF-8 content.
- Ensure that your back-end scripts, databases, etc can handle the characters you need in a seamless fashion.

http://www.w3.org/International/tutorials/tutorial-char-enc
Character encoding: Using escapes
What is an escape?

- \xE1; hex NCR
- \#225; decimal NCR
- &aacute; character entity
- \E1 CSS escape

NCR = Numeric Character Reference
Jako efektivnější se nám jeví pořádání tzv. Road Show prostřednictvím našich autorizovaných dealerů v Čechách a na Moravě, které proběhnou v průběhu září a října.
Jako efektivně se nám jeví pořádání tzv. Road Show prostřednictvím našich autorizovaných dealerů v Čechách a na Moravě, které proběhnou v průběhu září a října.
When should I use escapes?

- **Syntax-related characters**
  
  \&lt; (<) \&gt; (>) \&amp; (&) \&quot; ("")

- **Characters not supported by the document encoding**
  It may be better to change the encoding!

- **Characters not supported by the input tools**
  Not a long term solution!

- **Characters that are invisible or ambiguous**
  
  eg. \&rlm; / \&#x200F;
  
  eg. \&nbsp; / \&#xA0;
By the way…

- € is `&#x20AC;` not `&#x80;`
  - Use Unicode code points, no matter what encoding you used!

- U+00E1
  - Use Hex values rather than Decimal

- &aacute;
  - Character entities may cause problems for XML processing

- `&#x12345` not `&#xD808; '&#xDF45;`
  - Use a single value for supplementary characters, not two
Some lessons from this section

• Save data in an encoding that means you won't need to use escapes (eg. Unicode).
• Use escapes for characters that are invisible or ambiguous.
• Use Unicode code points for escapes, regardless of the character encoding you are using for your content.
• Be careful about using character entities (like &aacute;) if you are going to process the content later as XML.

http://www.w3.org/International/questions/qa-escapes
Language tagging:
Two types of declaration
Language tagging

Why?

- Accessibility
- Authoring tools
- Translation tools
- Font selection
- Page rendering
- Search
- Scripting
Language tagging

Why?

- Accessibility
- Authoring tools
- Translation tools
- Font selection
- Page rendering
- Search
- Scripting

Use now to increase potential applications!
Two things to be labelled

- **Text-processing language**
  - the language of a specific range of text
  - used for processing such as text-to-speech, styling, etc.
  - can indicate only ONE language at a time

- **Metadata about the intended audience**
  - describes the language(s) of the document as a whole
  - not a list of all languages used in the document
  - could be more than one language

The French word for *cat* is *chat*.

This is French text.

The French word for *cat* is *chat*.

This is an English document.
Two approaches

Japanese phrase book

hello 今日は
thank you ありがとう
goodbye さようなら

French

Bonjour cher lecteur.
Voici des choses que vous pouvez lire. Nous les avons traduits pour que vous les compreniez; malgré le fait que la navigation reste en anglais.

MyOrganization

Go here
Go there
Go elsewhere
Go back
Language tagging

Two approaches

Greetings

This Web page
contains an
everous
amount of
useful stuff.

We hope you
will enjoy

Bienvenue

Cette page
contient une
quantité
énorme de
 choses utiles.

Nous
espère que
ceci vous
plaira.
Language tagging

Two approaches

Japanese phrase book

Hello 今日は
Thank you ありがとう
Goodbye さようなら

English

MyOrganization

Go here

Bonjour cher lecteur.

Go there

Voici des choses que vous pouvez lire. Nous les avons traduits pour que vous les compreniez; malgré le fait que la navigation reste en anglais.

Go elsewhere

Go back

English & French

Bienvenue

Cette page contient une quantité énorme de choses utiles. Nous les avons traduits pour que vous les compreniez; malgré le fait que la navigation reste en anglais.
Language tagging
Ways of declaring language

HTTP/1.1 200 OK
Date: Wed, 05 Nov 2003 10:46:04 GMT
Server: Apache/1.3.28 (Unix) PHP/4.2.3

Content-Type: text/html; charset=utf-8
Content-Language: en

HTTP Content-Language header

Language attribute on html tag
<html lang="en">
<head>
  ...
  <meta http-equiv="Content-Language" content="en" />
  ...
</head>
<body>
  <p>The French word for <em>cat</em> is <em lang="fr">chat</em>.</p>
  ...
</body>
<html>

Content-Language meta tag

Language attribute on embedded element
Language tagging

Using HTTP

- Set on the server
- Can indicate more than one language
  - metadata ✓
  - text-processing ✗
- Lower precedence for text-processing than language attribute on html tag
- Use for text-processing in mainstream browsers is patchy

HTTP/1.1 200 OK
Date: Wed, 05 Nov 2003 10:46:04 GMT
Server: Apache/1.3.28 (Unix) PHP/4.2.3

Content-Type: text/html; charset=utf-8
Content-Language: en, fr, sp
Using attributes

- Information is inherited, so html element sets the default for the whole document
- Overrides any conflicting HTTP declaration
- Can indicate only a single language
  - metadata ✗ text-processing ✓
- Most user agents recognize it for text-processing
Language tagging
Using attributes

- **HTML:**
  
  `<html lang="ja">`

- **XHTML as text/html:**
  
  `<html lang="ja" xml:lang="ja"
  xmlns="http://www.w3.org/1999/xhtml">`

- **XHTML as XML:**
  
  `<html xml:lang="ja"
  xmlns="http://www.w3.org/1999/xhtml">`
<meta http-equiv="Content-Language" content="en,fr,sp" />

- Not mentioned in HTML specification

- Can indicate more than one language
  - metadata ✓
  - text-processing ✗

- Although informal guidance on the Web commonly points to this, and some authoring tools use it, user agents and applications don't actually recognize it
Language tagging

Embedded language changes

- **element:**

  ```html
  <p>The French for <em lang="fr" xml:lang="fr">Cat</em> is 
  <em lang="fr" xml:lang="fr">chat</em>.</p>
  ```

- **span:**

  ```html
  <p>The title in Chinese is <span lang="zh-Hans" xml:lang="zh-Hans">中国科学院文献情报中心</span>.</p>
  ```
Language tagging
Some lessons from this section

- Don't munge together declarations of text-processing language and metadata about the intended audience.
- Use HTTP for metadata about the intended audience.
- Use attributes for labeling text-processing language.
- Always include language information in the html tag.
- Avoid use of the meta element for language information you want to be used for text processing.

http://www.w3.org/International/TR/i18n-html-tech-lang
Language:
The new language tag values
Before: RFC 3066

language – region

IANA registered tag

- ISO 639 language codes
- ISO 3166 country codes

en
en-GB
en-scouse
Language tags

BCP 47 (includes RFC 4646)

Now: BCP 47 (includes RFC 4646)

language – script – region – variant – extension – private_use

♦ only language subtag required
♦ codes available only from new IANA registry (based on ISO and UN codes)
Type: language
Subtag: cs
Description: Czech
Added: 2005-10-16
Suppress-Script: Latn

Type: language
Subtag: cu
Description: Church Slavic
Description: Old Slavonic
Description: Church Slavonic
Description: Old Bulgarian
Description: Old Church Slavonic
Added: 2005-10-16

Type: language
Subtag: cv
Description: Chuvash
Added: 2005-10-16

http://www.iana.org/assignments/language-subtag-registry
IANA Language Subtag Registry Search

Select what you want to see. Click on a coloured bar for more information. Since the registry is in English, search in English.

Search descriptions: norw
Search for subtag(s):
Display a list: Languages

language codes

nb  Norwegian Bokmål
nn  Norwegian Nynorsk
no  Norwegian

geographic regions

NO  Norway

grandfathered

no-bok  Norwegian Bokmål [Do not use!]
no-nyn  Norwegian Nynorsk [Do not use!]

related links

About language subtags
PHP source code
IANA subtag registry

notes

Descriptions: You can search on any part of a description (e.g. 'eng' gives 3 results, including Bengali, English and Tengwar).

Subtags: Only exact matches of subtags are returned, i.e. 'ak' doesn't match 'akb'. If you enter a single subtag, the tool will search for any matches. If you enter a hyphen-separated list, the tool will look up potential matches for each subtag.

Validity and well-formedness: This is not reported for hyphen-separated tags. If the results are in a different order than the hyphenated list, or there are duplicate subtags in one category, or there are subtags missing from the results (except for a private use subtag, which always results in an empty list), then it is likely that something is wrong with the search.

http://rishida.net/utils/subtags/
Language tags

BCP 47 (includes RFC 4646)

- **language** – script – region – variant – extension – private_use

- always required
- codes available only from new IANA registry
- two-letter codes provided if available, 3-letter if not
- coming soon: ISO 639-6 codes
Language tags

Save your data in the encoding

language – script – region – variant – extension – private_use

- ISO 15924 code
- only one, directly after language
- 4 letters long
- use only if needed
- ' Suppress-Script' labels in registry,
- eg. en-Latn

zh-Hans
az-Cyrl
"… avoid script tags except where they add useful distinguishing information."

"for virtually any content that does not use a script tag today, it remains the best practice not to use one in the future."

Addison Phillips

zh-Hans
az-Cyrl

language – script – region – variant – extension – private_use
Save your data in the encoding

language – script – region – variant – extension – private_use

- ISO 3166-1 code or UN M.49 region code
- 2 letter alpha or 3 digit codes
- only one, following language and any script codes
- script code not required

- en-GB
- es-419
- zh-Hant-HK
Language tags

Some lessons from this section

• Use language attribute values from the IANA Language Subtag Registry.

• Follow the rules in BCP 47 – not RFC 3066.

• Keep the tags short – only use script subtags where really needed.

http://www.w3.org/International/articles/language-tags/
Text size
### Text size in translation

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean</td>
<td>조회</td>
<td>0.8</td>
</tr>
<tr>
<td>English</td>
<td>views</td>
<td>1</td>
</tr>
<tr>
<td>Chinese</td>
<td>次検視</td>
<td>1.2</td>
</tr>
<tr>
<td>Portuguese</td>
<td>visualizações</td>
<td>2.6</td>
</tr>
<tr>
<td>French</td>
<td>consultations</td>
<td>2.6</td>
</tr>
<tr>
<td>German</td>
<td>-mal angesehen</td>
<td>2.8</td>
</tr>
<tr>
<td>Italian</td>
<td>visualizzazioni</td>
<td>3</td>
</tr>
</tbody>
</table>

Around 3,100 m high, in Bhutan.

953 views | 3 comments

© This photo is public. Change?

Uploaded on Jan 28, 2007 | Map | Delete
Compound nouns

Input processing features

Eingabeverarbeitungsfunktionen
Seals and sealions at Pier 39
Seals and sealions at Pier 39

Seals and sealions at Pier 39 Encargar impresión
## Form field label positioning

### Global settings

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface language</td>
<td>English</td>
</tr>
<tr>
<td>Search language</td>
<td>English</td>
</tr>
<tr>
<td>Number of results</td>
<td>10</td>
</tr>
</tbody>
</table>

[Save preferences]
Form field label positioning

Global settings
- Interface language: English
- Search language: English
- Number of results: 10
- Save preferences

Acuan Umum
- Bahasa Pengantar di Antar Muka: Ingriss
- Bahasa Pengantar untuk Penelusuran: Ingriss
- Jumlah Hasil Penelusuran: 10
- Simpan Acuan

Allgemeine Voreinstellungen
- Sprache der Benutzeroberfläche: Englisch
- Suchsprache Anzahl der Ergebnisse: 10
- Einstellungen speichern
Text size

Form field label positioning

Global settings
- Interface language: English
- Search language: English
- Number of results: 10

Save preferences

Acuan Umum
- Bahasa Pengantar di Antar Muka: Ingriss
- Bahasa Pengantar untuk Penelusuran: Ingriss
- Jumlah Hasil Penelusuran: 10

Simpan Acuan

Allgemeine Voreinstellungen
- Sprache der Benutzeroberfläche: Englisch
- Suchsprache: Englisch
- Anzahl der Ergebnisse: 10

Einstellungen speichern
Graphics in fixed width boxes

Idioma de la interfaz

- arabe
- inglés
- finlandés
- francés
- italiano
- japonés
- ruso
- español
- sueco
Graphics in fixed width boxes

Bahasa Pengantar
untuk Penelusuran

- Arab
- Inggris
- Finlandia
- Perancis
- Italia
- Jepang
- Rusia
- Spanyol
- Swedia
Graphics in fixed width boxes

```css
.box h3 {
  margin: 0;
  padding: 6px 8px 4px 10px;
  font-size: 130%;
  color: #333;
  background: url(img/h3-bg.gif) no-repeat top left;
  border-bottom: 1px solid #E0CFAB;
}
```

Bulletproof Web Design, Dan Cederholm
Sliding Doors of CSS, Doug Bowman
http://alistapart.com/articles/slidingdoors/
The value of internationalization

Internationalization significantly affects the ease of the product's localization. Retrofitting a linguistically- and culturally-centered deliverable for a global market is obviously much more difficult and time-consuming than designing a deliverable with the intent of presenting it globally. (Think back to the Y2K effort and trying to "undo" two-character year fields that were built on the assumption of "19x".)

So ideally, internationalization occurs as a fundamental step in the design and development process, rather than as an afterthought that can often involve awkward and expensive re-engineering.

further reading

- W3C i18n resources relating to internationalization

Author: Richard Ishida, W3C, Susan K. Miller, Boeing

Content first published 2005-12-05. Last substantive update 2005-12-05 GMT. This version 2005-10-19 19:25 GMT

For the history of document changes, search for qa-i18n in the i18n blog.

Copyright © 2005-2006 W3C (MIT, ERCIM, Keio). All Rights Reserved. W3C liability, trademark, document use and software licensing rules apply. Your interactions with this site are in accordance with our public and Member privacy statements.
Text size

Round cornered boxes

Tell us what you think (English).
- Send us a comment
- Subscribe to an RSS feed
- New resources
- Home page news

Powiedz nam co myślisz (по angielsku)
- Wyślij komentarz
- Prenumeruj kanał RSS
- Nowe źródła
- Wiadomość ze strony głównej

قِل لنا رأيك (بالإنجليزية).
- ابعث تعليق.
- الانضمام إلى خدمة RSS
- موارد جديدة
- الصفحة الرئيسية للأخبار

Выскажите своё мнение (по-английски).
- Напишите нам
- Подпишитесь на ленту новостей RSS
- Новые публикации
- Новости главной страницы

告诉我们您的想法（用英语）。
- 将您的评论发送给我们
- 订阅 RSS 提要
- 新的资源
- 主页新闻
Round cornered boxes
Round cornered boxes

English:
Input processing features

German:
Eingabeverarbeitungsfunktionen
**Text size**

**Round cornered boxes**

**English:**

Input processing features

**German:**

Eingabeverarbeitungsfunktionen
Some lessons from this section

- Think about how to cope with text that expands in translation
- Avoid placing labels to the side of form entry fields, icons, etc, if you can
- For text backgrounds use large images and ‘sliding door’ effects to allow coverage of text whatever the size
- Leave expansion space at the end of lines of tabs
- Consider how to deal with long compound nouns: don’t make the column too narrow; warn translators
- Look out for and lobby the W3C CSS Working Group to publish rounded borders and multiple backgrounds parts of CSS3
- Lobby Mozilla to support the soft-hyphen character

http://www.w3.org/International/questions/qa-resizing-backgrounds
Navigating to localized pages
Navigating to localized pages

Language links

<table>
<thead>
<tr>
<th>Global Preferences</th>
<th>Changes apply to all Google services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface Language</td>
<td>Display Google tips and messages in: English</td>
</tr>
<tr>
<td></td>
<td>If you do not find your native language in the pull down above, you can help Google create it through our Google in Your Language program.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Search Language</th>
<th>Search for pages written in any language (Recommended)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Search only for pages written in these languages:</td>
</tr>
<tr>
<td>Arabic</td>
<td>English</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>Estonian</td>
</tr>
<tr>
<td>Catalan</td>
<td>Finnish</td>
</tr>
<tr>
<td>Chinese (Simplified)</td>
<td>French</td>
</tr>
<tr>
<td>Chinese (Traditional)</td>
<td>German</td>
</tr>
<tr>
<td>Croatian</td>
<td>Greek</td>
</tr>
<tr>
<td>Czech</td>
<td>Hebrew</td>
</tr>
<tr>
<td>Danish</td>
<td>Hungarian</td>
</tr>
<tr>
<td>Dutch</td>
<td>Icelandic</td>
</tr>
<tr>
<td></td>
<td>Indonesian</td>
</tr>
<tr>
<td></td>
<td>Italian</td>
</tr>
<tr>
<td></td>
<td>Japanese</td>
</tr>
<tr>
<td></td>
<td>Korean</td>
</tr>
<tr>
<td></td>
<td>Latvian</td>
</tr>
<tr>
<td></td>
<td>Lithuanian</td>
</tr>
<tr>
<td></td>
<td>Norwegian</td>
</tr>
<tr>
<td></td>
<td>Polish</td>
</tr>
<tr>
<td></td>
<td>Portuguese</td>
</tr>
<tr>
<td></td>
<td>Romanian</td>
</tr>
<tr>
<td></td>
<td>Russian</td>
</tr>
<tr>
<td></td>
<td>Serbian</td>
</tr>
<tr>
<td></td>
<td>Slovak</td>
</tr>
<tr>
<td></td>
<td>Slovenian</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
</tr>
<tr>
<td></td>
<td>Swedish</td>
</tr>
<tr>
<td></td>
<td>Swedish</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SafeSearch Filtering</th>
<th>Google’s SafeSearch blocks web pages containing explicit sexual content from appearing in search results.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Use strict filtering (Filter both explicit text and explicit images)</td>
</tr>
<tr>
<td></td>
<td>Use moderate filtering (Filter explicit images only - default behavior)</td>
</tr>
<tr>
<td></td>
<td>Do not filter my search results.</td>
</tr>
</tbody>
</table>

| Number of Results | Google’s default (10 results) provides the fastest results. Display | 10 results per page. |
Navigating to localized pages

Language links
Navigating to localized pages

Pull down lists
Navigating to localized pages

Pull down lists
### Navigating to Localized Pages

#### Selection Pages

Selecting a localized page allows users to view content in their preferred language. This feature is particularly useful for international audiences, as it makes the website more accessible and user-friendly. The selection process typically involves navigating through a menu or thumbnail gallery that displays images of the website in different languages.

#### Example:

- **Bahasa Indonesia**
- **Česky**
- **Dansk**
- **Deutsch**
- **Diné bizaad**
- **Español**
- **Eesti keele**
- **Français**
- **Gagana Samoa**
- **Hiligaynon**
- **Hmoob**
- **Hrvatski**
- **Ilokano**
- **Italiano**
- **Kiribati**
- **Kreyòl Ayisyen**
- **Lea Fakatonga**
- **Magyar**
- **Nederlands**
- **Norsk**
- **Polski**
- **Português**
- **Reo Tahiti**
- **Română**
- **Slovenčina**
- **Slovenščina**
- **Suomi**
- **Suomi**
- **Tagalog**
- **Tiếng Việt**
- **Vosa vakaviti**
- **Ελληνικά**
- **Български**
- **Монгол**
- **Русский**
- **Українська**
- **ภาษาไทย**
- **ภาษาไทย**
- **ภาษาไทย**
- **中国語**
- **日本語**
- **한국말**
Navigating to localized pages

Using alternative selection lists
Country Sites

Either the webapp context your region JSP URI refers to doesn't exist or the calling web application does not have permission to access another web application context. If the web application exists in your application server, consult the Trouble Shooting Guide for details on how to configure your application server to support cross context forwarding.

<table>
<thead>
<tr>
<th>Country</th>
<th>Language 1</th>
<th>Language 2</th>
<th>Language 3</th>
<th>Language 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Samoa</td>
<td>English</td>
<td>English</td>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>Estonia</td>
<td>Estonian</td>
<td>Spanish</td>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>Argentina</td>
<td>Spanish</td>
<td>English</td>
<td>Dutch</td>
<td>Slovak</td>
</tr>
<tr>
<td>Fiji</td>
<td>English</td>
<td>Dutch</td>
<td>Slovak</td>
<td>Slovenian</td>
</tr>
<tr>
<td>Australia</td>
<td>English</td>
<td>Finnish</td>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>Austria</td>
<td>German</td>
<td>French</td>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>Belgium</td>
<td>Dutch</td>
<td>French Polynesia</td>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Niue</td>
<td></td>
<td>English</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singapore</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovak Republic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovenia</td>
<td>Slovenian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Africa</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Korea</td>
<td>Korean</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Navigating to localized pages issues

Identifying the list
Navigating to localized pages

Using maps

Xerox Websites

View a list of Worldwide Contacts

North America
- Canada/English
- Canada/Français
- United States of America

Latin America and Caribbean
- Argentina
- Bermuda
- Brasil
- Cayman Islands
- Chile
- Colombia
- Ecuador
- Mexico
- Nicaragua
- Paraguay
- Peru
- Trinidad & Tobago
- Uruguay
- Venezuela
Some lessons from this section

- For links to alternative language or country sites, translate link text into the language of the content pointed to.
- Make it easy to find the current page when changing language or site.
- Consider whether to use graphics or text, given that fonts may not be available on the user’s platform for all options.
- If you use text, you’ll need the page encoding to be UTF-8.
- Remember that pull-down lists don’t support graphics, so consider alternatives.
- Find a way to draw attention to a list that is recognizable regardless of language, and put it in a visible location.
- Consider maps as an alternative selection mechanism.

http://www.w3.org/International/questions/qa-navigation-select
Concluding remarks
In conclusion

• We haven't been talking about things that the localizers do, we're talking about the way content is created!

• Good internationalization practice is often just good content design.

• This is a part of Web Accessibility.

• Take a look at the material on the I18n Activity site.

• Pick up your quicktips cards here.
Thank you

http://www.w3.org/International/