

W3C India Workshop

Best Practices for Internationalization

RKVS Raman
CDAC Electronics City
Bangalore
raman@cdacbangalore.in

Terminologies First

- Internationalization (I18N)
 - Process of creating an application so it can be adapted to different languages and regions with minimal or no coding changes

Terminologies...

- Character Set
 - A character set is a collection of letters and symbols used in a writing system,
 - eg. the ASCII character set covers letters and symbols for English text, ISO-8859-6 covers letters and symbols needed for many languages based on the Arabic script, and Unicode contains characters for most of the living languages and scripts in the world.

Terminologies...

- Font/Typeface
 - A typeface is a coordinated set of glyphs designed with stylistic unity.
 - A font is a set of glyphs (images) representing the characters from a particular character set in a particular typeface.

Terminologies...

- Encoding
 - A character encoding is the key that maps a particular byte or sequence of bytes to particular characters that the font renders as text.

Terminologies...

- Text Flow Direction
 - The General direction in which the flow of characters occur for a language.
 - Left to Right
 - Right to Left
 - Vertical
 - Bi-Directional

Standard Resources

www.unicode.org

- Online Standard
- Technical Reports
- FAQs
- General Information
- Discussion Forums, Conferences

Programming Resources

- System APIs:
 - Linux, Java, Unix, Windows, Oracle ...
- Languages
 - Java, JavaScript, C#, Perl, Python, PHP, .Net
- Cross-platform libraries:
 - ICU, Pango ...

Best Practices - Encoding

- Default recommended encoding for content on web is UTF-8
- Denoted by HTTP Header
 - Content-Type: text/html; charset=utf-8
- Can also be informed in the document using the meta-tag or as XML directive in XHTML
 - `<meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>`
 - `<?xml version="1.0" encoding="UTF-8"?>`
 - `@charset "UTF-8";`

Best Practices - Language

- Information about the language in use on a page is important for
 - accessibility
 - styling
 - searching:

Best Practices – Specifying Language

- HTTP Header
 - Content-Language: en, fr, sp
- In document as
 - `<html lang="en">`
 - `<meta http-equiv="Content-Language" content="en,fr,sp" />`
 - `<p>The French word for cat is <em lang="fr">chat.`

Best Practices - Text Direction

- Text Direction can be controlled by specifying the direction attribute for an element's style

- CSS

```
div  
{  
direction:rtl;  
}
```

- Javascript DOM

- `document.getElementById("p1").style.direction="rtl";`

User Preferred Language

- Multi-lingual sites may need to serve the content in a language preferred by user.
- User can set their preferred language in the browsers which is sent to the web site with the request as a header.
- Following header can be investigated by server for user preferred language.
 - Accept-Language: kok, mr;q=0.8, hi;q=0.7

Localization

- Process of adapting software *for a specific language/region* by adding locale-specific components and text translations

Localization Aspects

- Language
- Region
- Country
- Culture
- Date and Currency formats
- Legal Issues

Planning a localized application

- Different development platforms offer varied solutions
- Some common guidelines can be drawn.
- Adoption of certain frameworks assist in creating localized web applications
- Example: Zend, Django, Ruby on Rails, GWT
- Content Management Systems like Drupal and Plone are also internationalized ground up.

Localization Model and Tools

- Text translation
 - Localization formats
 - HTML with template library
 - W3C Internationalization Tag Set (tool support?)
 - GNU gettext/PO
 - XLIFF - XML Localization Interchange File Format
 - Localization tools
 - OmegaT
 - Open Language Tools (Sun)
 - The WordForge Project: Pootle
 - PO Edit

References

- World Wide Web (<http://www.w3c.org>)
- W3 Schools (<http://www.w3schools.com>)
- Unicode (<http://www.unicode.org>)

Thank you

Questions?