HTML5: proposed markup changes related to i18n
What I mean when I say HTML5 is…
the main markup specification

- http://dev.w3.org/html5/spec/
This presentation contains notes for most slides that can be accessed from the icon at the top left corner of the slide.
Outline

Character encoding
Language
Ruby
Bidirectional text
Date & time
Character encoding
Encoding choice

Unicode on the Web

Other

ASCII

UTF-8

Encoding choice

Encoding choice

Encoding choice

Encoding choice

Encoding choice

Encoding choice

Encoding choice

Encoding choice

Encoding choice
Character encoding

Encoding declarations

• Strong encouragement to use UTF-8.
• New meta charset declaration. Either approach will work, but check you don't have both.
• Must be completely within the first 512 bytes of the file.
Character encoding

Polyglot documents

```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 lang='en' xml:lang='en' xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <meta http-equiv=Content-Type content="text/html; charset=utf-8" />
  </head>
  ...
</html>
```

- Strong encouragement to use UTF-8.
- New meta charset declaration. Either approach will work, but check you don't have both.
- Must be completely within the first 512 bytes of the file.
- Polyglot documents use UTF-8 only, but no XML declaration.
UTF-16 documents

• Strong encouragement to use UTF-8.
• New meta charset declaration. Either approach will work, but check you don't have both.
• Must be completely within the first 512 bytes of the file.
• Polyglot documents use UTF-8 only, but no XML declaration.
• Must NOT use this for UTF-16. HTML5 will rely on the byte-order mark.

✘
Character encoding

**charset attributes**

```html
<link rel="stylesheet" charset="Windows-1251"
    href="mystyles.css" type="text/css">
```

See our `<a href="/mysite/mydoc.html" charset="ISO-8859-1">list of publications</a>.

- Not well supported by browsers.
- Hard to ensure it continues to be correct.
- There are better ways to do it.
Character encoding

**charset attributes**

```html
<link rel="stylesheet" charset="Windows-1251"
href="mystyles.css" type="text/css">

See our `<a href="/mysite/mydoc.html" charset="ISO-8859-1">list
of publications</a>.
```

- Not well supported by browsers.
- Hard to ensure it continues to be correct.
- There are better ways to do it.

- Do not use with link or a elements.
- Ok for script element.
Language
Language
definitions

Language declarations

```html
<DOCTYPE html>
<html lang=it>
<head>
<meta http-equiv=Content-Language content="en, it"/>
</head>
...
```

- Attributes indicate the language of text inside that element for text processors. Only one language value allowed.
- Meta elements indicate the language of the expected readership. Multiple languages are ok.
- Attributes override other declarations.
Language declarations

Attributes indicate the language of text inside that element for text processors. Only one language value allowed.

Meta elements indicate the language of the expected readership. Multiple languages are ok.

Attributes override other declarations.

The meta element with Content-Language is now non-conforming.
Ruby markup
Standards support

Ruby annotation

\(<ruby><rb>凝</rb><rt>ぎょう</rt></ruby>\)
\(<ruby><rb>視</rb><rt>し</rt></ruby>\)

HTML5
A vocabulary and associated APIs for HTML and XHTML.
Date and time
Standards support

Date and time

<time datetime="2004-08-08">8 สิงหาคม ๒๕๔๗</time>

Datetime picker demo

<form>
  <input type="date">
</form>
Bidirectional text
نشاط التدوير، W3C

<p dir="rtl">W3C نشاط التدوير،</p>
Inline element content often needs to be isolated – especially when data is imported into the page.
Bidirectional text

Bidi isolation

• Existing bidi markup in HTML is NOT designed to solve this problem.

<span dir=rtl>ليلית</span> - 5 reviews

TOP RATED RESTAURANTS

Aroma - 3 reviews

לייזה - 5 reviews

<span dir=rtl>לייזה</span> - 5 reviews - 5 reviews
Bidi isolation

- CSS3 added the “isolate” value to the unicode-bidi property.
- HTML5 adds a new <bdi> element, with unicode-bidi:isolate in the default stylesheet.
- The <output> element behaves the same way.

<p><bdi>לילית</bdi> - 5 reviews</p>
Bidirectional text

\texttt{dir=auto} value

- Sometimes the appropriate direction can only be determined at run-time.
Bidirectional text
dir=auto value

• Sometimes the appropriate direction can only be determined at run-time.
• HTML5 adds new “auto” value for the dir attribute.

• CSS3 adds a “plaintext” value to the unicode-bidi property to allow per-paragraph auto-direction, primarily for use on <textarea> and <pre> elements.

• dir=auto sets the unicode-bidi CSS property to “plaintext” for <textarea> and <pre> elements, to “bidi-override isolate” for <bdo> elements, and to “isolate” otherwise.

• It estimates a direction according to the UBA method.

<p>Your search - <span class=booktitle dir=auto>CSS</span> - did not match any documents.</p>
Other bidi changes

• Reporting the chosen direction of `<input>` and `<textarea>` in form submissions

• `<br>` should serve as a bidi separator

• Block elements as bidi separators

• `<title>` should support the dir attribute

• `<option>` should support the dir attribute and be displayed accordingly both in the dropdown and after being chosen

• …
Getting involved...
The Web needs your help

this is your Web – not the W3C's

we need You to make the Web worldwide

get involved

Thank you

http://www.w3.org/International/talks/1104-pisa/