HTML5: proposed markup changes related to i18n

Richard Ishida
W3C Internationalization Activity

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

جعل شبكة الويب العالميّة عالميّة حقًّا!

وب جهانی را به درستی جهانی سازیم!

عالمگیر ویب کو حقیقی طور پر عالمگیر بنانا

Ꮢ ɑʍɑ2ʈuɑnhɑyïu gɑugu hnnᏢ hɑʍɑ2ʈuɑnhɑyïu nɑnál

ĊᒼᒷL ᐃᑭᐊᖦᑭᐸᖦ ᔭᒪᖦᐺᐊᑕᒍ� ᓵᓄᑎᐸᐺ.

"Дүниежүзілік торды" нағыз дүниежүзілік етеміз!

वर्ल्ड वाईड वेबलाई यथार्थमै विश्वव्यापी बनाउने !

የዓለም አቀፉን ድር በእውነት አለም አቀፍ ማድረግ!

Κάνοντας τον Παγκόσμιο Ιστό πραγματικά Παγκόσμιο

ਵਰਡ ਵਾਈਡ ਵੈਬ ਨੂੰ ਵਾਕਈ ਵਿਸ਼ਵ-ਵਿਆਪੀ ਬਨਾਉਣਾ !

缔造真正全球通行的万维网

ליצור מהרשת רשת כלל עולמית באמת!

ˈmeɪkɪŋ ðə wɜːld waɪd wɛb ˈtruːlɪ ˈwɜːldˈwaɪd

ワールド・ ワイド・ ウェッブを世界中に広げましょう

ធ្វើឱ្យវេឡាំយវែបមានទូទាំងពិភពលោកពិប្រាកដមែន!

전세계의 월드 와이드 웹으로 만들기!

Gwneud y we fyd-eang yn wirioneddol fyd-eang!

การทำให้ World Wide Web แพร่หลายไปทั่วโลกอย่างแท้จริง

འཛམ་གླིང་ཡོངས་འབྲེལ་འདི་ རྫོ་མ་འདད་རང་ འཛམ་གླིང་ཡོངས་ལ་ཁྱབ་ཚུགས་པ་བཟོ་བ།

Other
ASCII
UTF-8

Unicode on the Web

Encoding declarations

```
<!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>

...
```

```
<DOCTYPE html>

<html>

<head>

<meta  charset=utf-8>

</head>

...
```

- 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

```
<?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>

...
```
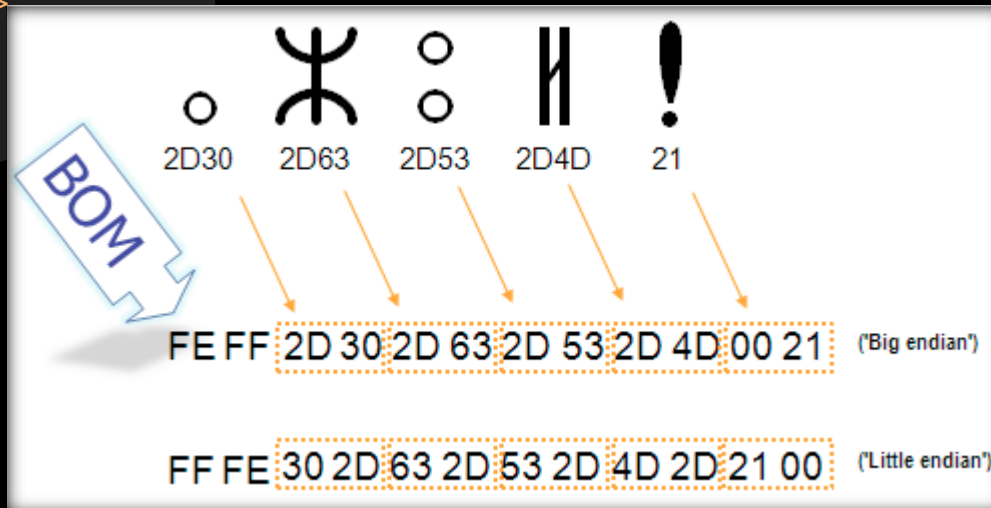
- Strong encouragement to use UTF-8.

- 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

```
<DOCTYPE html>

<html>

<head>

<meta charset=utf-16>

</head>

…
```


UTF-8.

Either
you don't

- 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.

charset attributes

```
<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.

charset attributes

```
<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 declarations

```
<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

```
<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.

- The meta element with Content-Language is now non-conforming.

Ruby markup

Ruby annotation

鬼門の方角を凝視する
きもん ほうがく ぎょうし

```
<ruby<rb>凝</rb><rt>ぎょう</rt></ruby>
  <ruby><rb>視</rb><rt>し</rt></ruby>
```

```
<ruby>凝<rt>ぎょう</rt>視<rt>し</rt></ruby>
```

```
<ruby>
  <rbc><rb>凝</rb><rb>視</rb></rbc>
  <rtc><rt>ぎょう</rt><rt>し</rt></rtc>
  </ruby>
```

東南の方角
とうなん たつみ

HTML5

A vocabulary and associated APIs for HTML and XHTML

http://dev.w3.org/html5/spec/Overview.html
Previous Versions:
http://www.w3.org/TR/2009/WD-html5-20090825/
http://www.w3.org/TR/2009/WD-html5-20090423/
http://www.w3.org/TR/2009/WD-html5-20090212/
http://www.w3.org/TR/2008/WD-html5-20080610/
http://www.w3.org/TR/2008/WD-html5-20080122/

Status of This document

This section describes the status of this document at the time of its publication. Other documents may supersede this document. A list of current W3C publications and the most recently formally published revision of this technical report can be found in the W3C technical reports index at http://www.w3.org/TR/.

If you wish to make comments regarding this document, please send them to public-html-comments@w3.org

Date and time

Date and time

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

Datetime picker demo

```
<form>
        <input type="date">
</form>
```

Bidirectional text

Bidirectional text support

نشاط التدويلW3C ، ❌

✔ W3C ، نشاط التدويل

`<p dir="rtl">`W3C ، نشاط التدويل`</p>`

Bidi isolation

Inline element content often needs to be isolated – especially when data is imported into the page.

Bidi isolation

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

`<span dir=rtl>לילית</span> - 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.
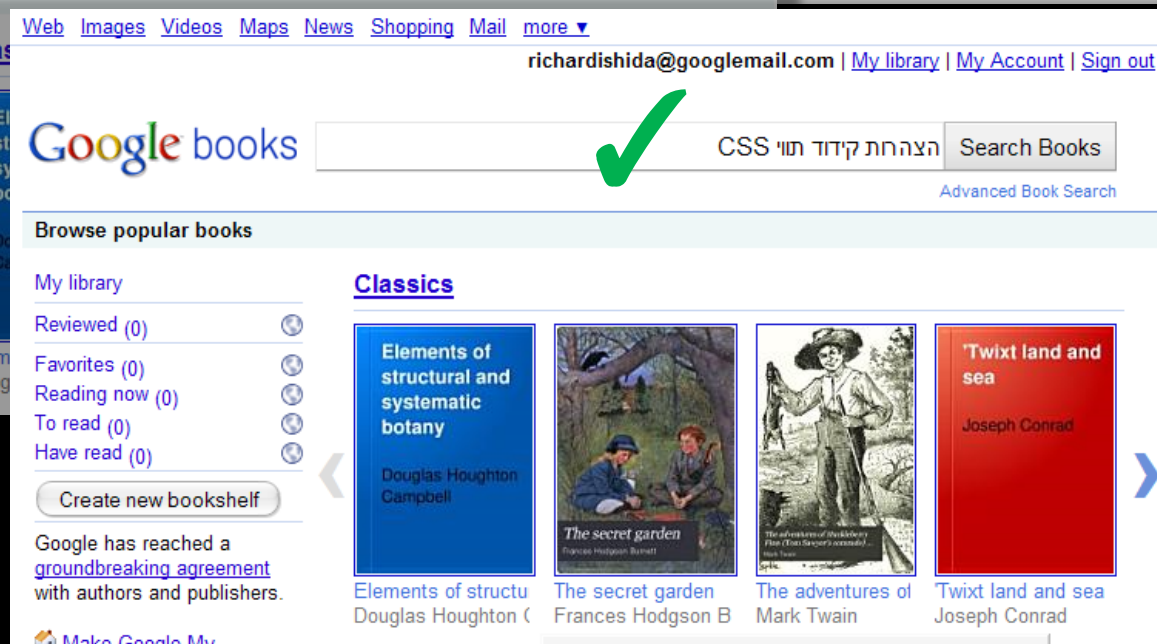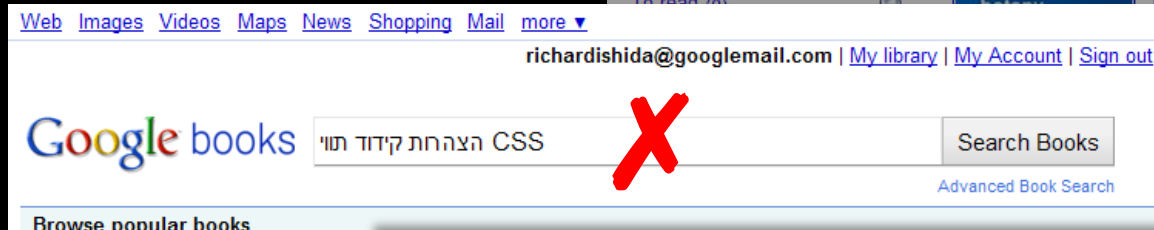
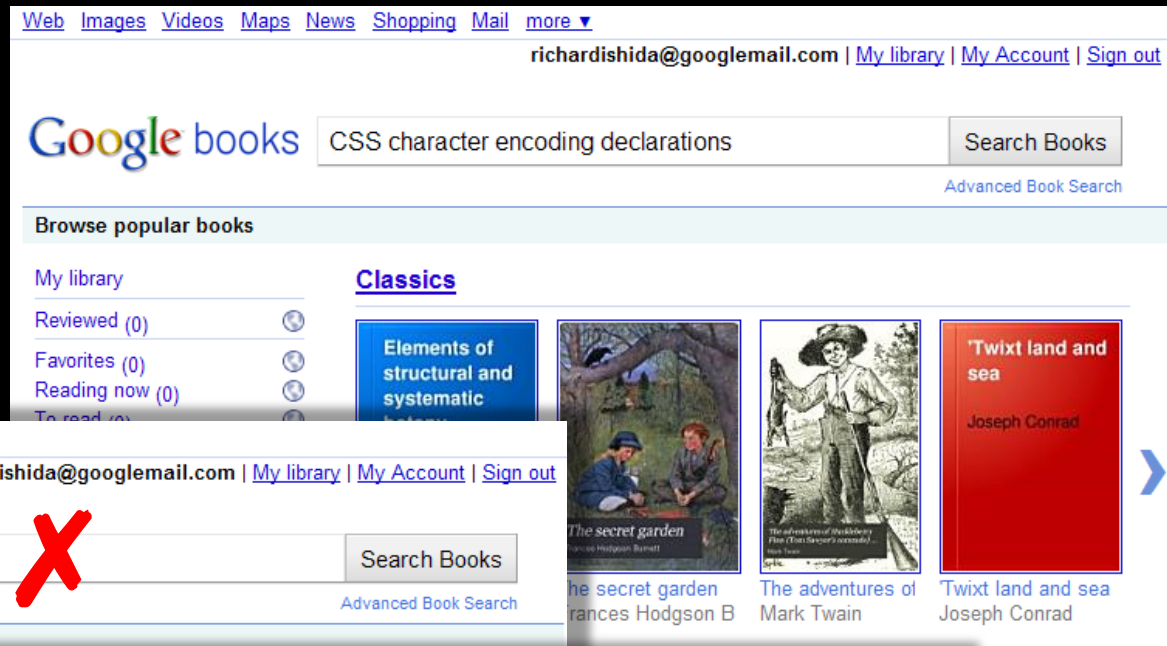- The <output> element behaves the same way.
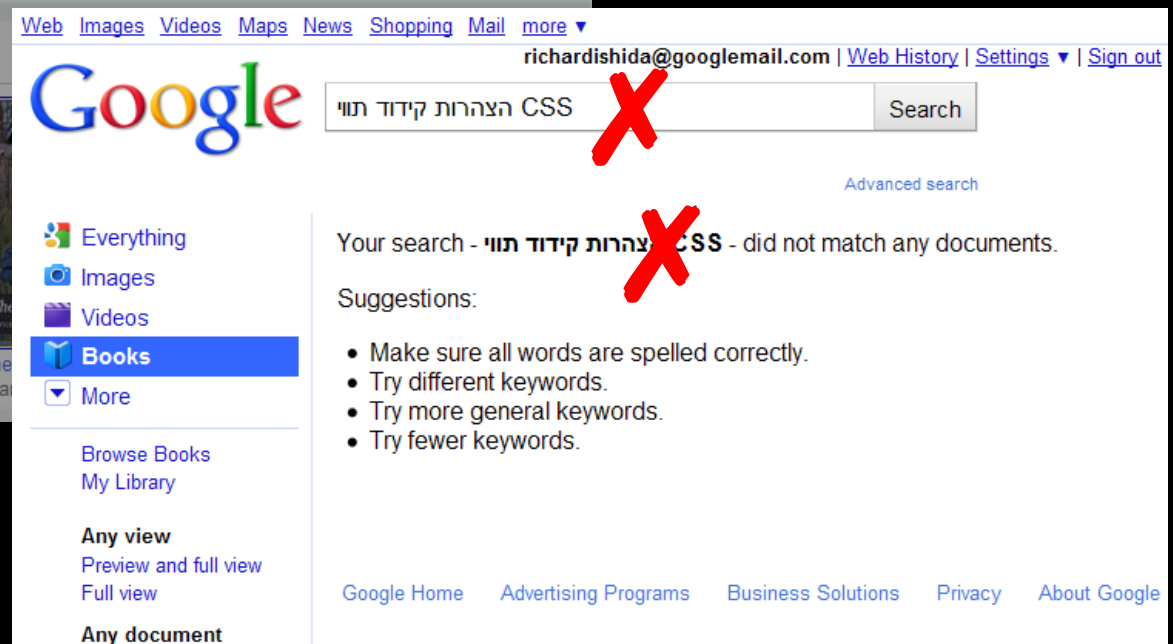

`<p><bdi>לילית</bdi> - 5 reviews</p>`

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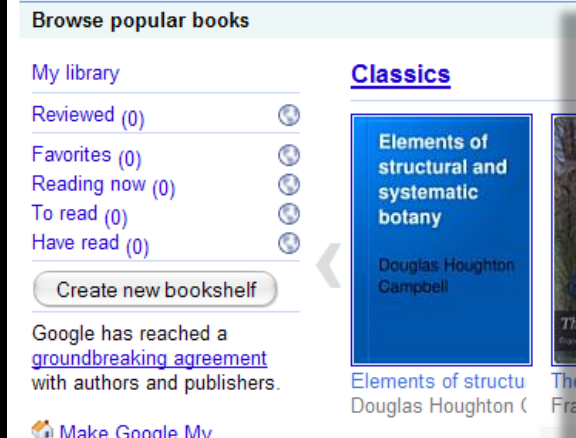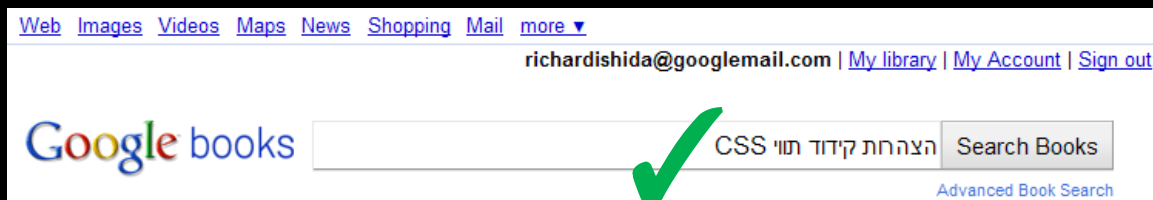
- Sometimes the appropriate direction can only be determined at run-time.

dir=auto value

- Sometimes the appropriate direction can only be determined at run-time.

dir=auto value

- 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>הצהרות קידוד תווי יוות דודיק CSS</span
> - did not match any
documents.</p>
```

Other bidi changes

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

-
 should 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/