



## The i18n Activity



## Activity Groups


- Internationalization Working Group
- Internationalization Interest Group
- Task Forces

## Community Groups

- Character Description Language Community Group
- Best Practices for Multilingual Linked Open Data
- Chinese Digital Publishing
- Mobile Web in Indian Languages
- ...

There are currently two groups in the Internationalization Activity. Task forces can be created under each of these groups.

In addition, anyone who can attract enough interest can set up a Community Group at the W3C. These groups are self-organizing. There are already some dealing with internationalization topics.

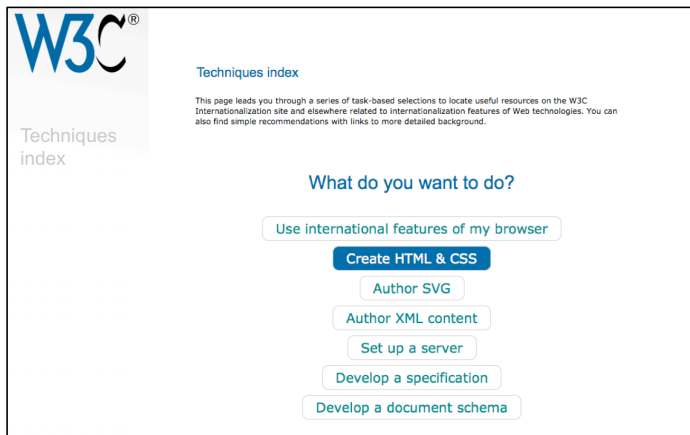


## Cross-organization links

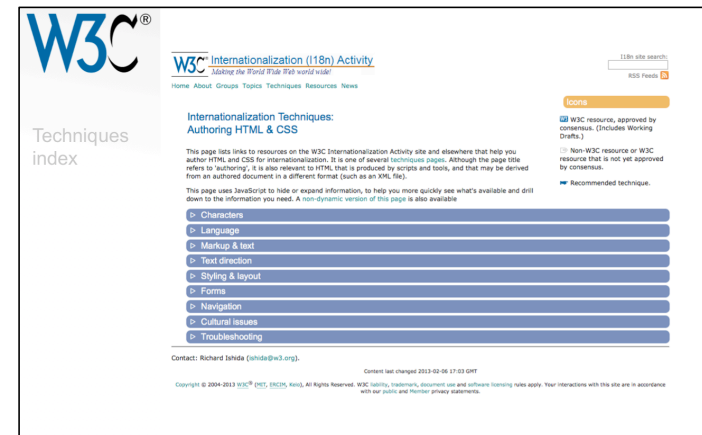
- Unicode (bidi, encodings, script features, ... )
- IETF (language tags, IDN & IRI, ...)
- EcmaScript (internationalization features)
- Language industry (ITS, MultilingualWeb, ...)

The Internationalization Working Group contains members from and liaises with many key organizations involved in ensuring that the Web remains multilingual, including Unicode, the IETF, EcmaScript, various language industry initiatives, etc.

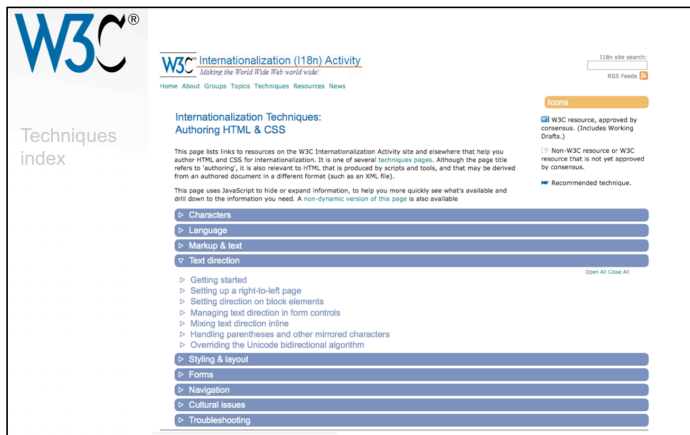




The techniques index allows you to discover how to incorporate internationalization needs into your work. It is organized by task, and helps you narrow in quickly on the information you need.







The lowest level of the index provides links to useful information, and also a set of do's and don'ts, linked to explanations and examples.

W3C<sup>®</sup>

W3C Internationalization Checker (Prototype only)  
Is your Web site internationalized?

Address:

Results: 2 8 1

<http://validator.w3.org/i18n-checker/>

Help users & content authors benefit from the new features

Character encoding	Code
HTTP Content-Type: No charset found	Content-Type: text/html
Byte order mark (BOM): <span style="color: red;">✗</span>	
xml declaration: None found	
meta charset element: <span style="color: red;">✗</span>	meta: http-equiv="Content-Type" content="text/html; charset="iso-8859-1" />
HTML5 meta charset element: None found	
Language	Code
<html lang="ja" <span style="color: green;">✓</span>	<html lang="ja" xml:lang="ja" />
<html xml:lang="ja" <span style="color: green;">✓</span>	<html lang="ja" xml:lang="ja" />
HTTP Content-Language: <span style="color: green;">✓</span>	Content-Language: ja, tk
meta content-language element: <span style="color: green;">✓</span>	meta: http-equiv="Content-Language" content="ja, tk" />
Text direction	Code
Default direction: <span style="color: green;">✓</span>	<html lang="ja" xml:lang="ja" />
Class & id names	Code
Non-ascii class or id names: <span style="color: red;">✗</span>	

You can run your pages through the online Internationalization Checker to spot issues and find out how to fix them.

W3C<sup>®</sup>

MultilingualWeb workshops

Developers  
Creators  
Localizers  
Machines  
Users  
Policy

**MultilingualWeb**

<http://multilingualweb.eu/>

For some years we have been running very successful workshops in Europe, attracting 100-150 people, with the stated intent to bridge between disciplines and bring together people who don't typically mix, and yet who are all working on making the World Wide Web world wide.



The Internationalization Working Group also develops tests for internationalization features, and adds many of those tests to the HTML and CSS test suites. These tests are useful to content developers, to give an idea of what is and is not currently supported, but they are also often used by browser implementers too, to test feature support.



Aim:  
To start the flow of ideas for the experts in the room about important topics to discuss.

To familiarize non-experts in the room with some typical aspects of internationalization.

Nowhere near exhaustive!



## Current preoccupations


- Empowering local communities to participate and develop requirements
- Ensure that newly specified features make it into browsers
- Expand work with the localization & language technology communities
- Make spec developers more aware of internationalization issues and needs
- Help users & content authors benefit from the new features


$$f(x) = \left\{ \begin{array}{ll} \frac{1}{x} & x > 0 \\ 0 & x = 0 \\ x & x < 0 \end{array} \right\} = (x, \infty) \cup (-\infty, x)$$

- Counter styles
- Arabic Mathematics
- Ruby Use Cases
- eBooks Workshop

2 Lepcha

† Greek	໓ Lao	໒ Thai
໓ Gujarati	໙ Latin	໑ Tibetan



Bidirectional scripts

**TOP RATED RESTAURANTS**

---

Aroma - 3 reviews  
 ⭐⭐⭐⭐☆  
 מִינָה סְנוּרָה - 5 reviews  
 ⭐⭐⭐⭐⭐  
 רומא מִינָה סְנוּרָה  
 ⭐⭐⭐⭐⭐

### "Additional Requirements for Bidi in HTML5 & CSS"

- bidi isolation
- handling plain text line ends
- form submission
- passing text to the browser chrome or scripts
- ...

**TOP RATED RESTAURANTS**

---

Aroma - 3 reviews  
 ⭐⭐⭐⭐☆  
 מִינָה סְנוּרָה - 5 reviews  
 ⭐⭐⭐⭐⭐  
 רומא מִינָה סְנוּרָה - 3 reviews  
 ⭐⭐⭐⭐☆

<http://www.w3.org/International/tutorials/bidi-xhtml/>

Recently a lot of work has been done to improve support for Arabic, Hebrew, and other right-to-left scripts in HTML and CSS. In particular, the ability to isolate items has significantly improved handling of text that is added to a page from an external source, as well as many normal content authoring situations.

You can learn more about how to use these techniques at <http://www.w3.org/International/tutorials/bidi-xhtml/>.

W3C®

Japanese

vertical writing mode

行組版方法についての解説

horizontal writing mode

日本語組版処理の要件

日本語組版処理の要件


行組版方法についての解説

character class tables

“Requirements for Japanese Text Layout”


- page layout
- vertical text
- headers & footers
- justification
- ruby
- warichu
- positioning tables/illustrations
- character class tables

The Japanese layout requirements document showed how useful such a document can be. It has been used to guide development of several major technologies of the Open Web Platform. It was made available in Japanese and English, and published also in book form.




Basic process

- Task force established.They work in their language, meeting regularly and use a mailing list in their language.
- They provide regular updates and translations of their document in English.The i18n WG reviews and assists.
- Production of near final version in English and in W3C format.
- Wide review of the document, promoted by i18n WG.
- Incorporation of review feedback and publication as WG Note.
- Second version produced.
- Nominal additional step: Create 'gap' document to prioritise and promote changes to technologies, such as CSS or HTML.



Korean



**“Requirements for Hangul Text Layout & Typography”**

- page layout
- vertical text
- headers & footers
- paragraph adjustment
- positioning tables/ illustrations
- character class tables
- kerning

Work has also started on a similar document for Korean requirements.

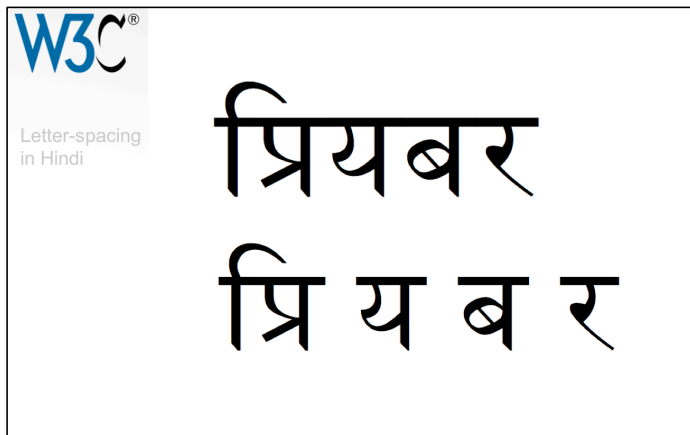
[illegible]

## More work is needed!

And another document is looking at requirements for languages of India.

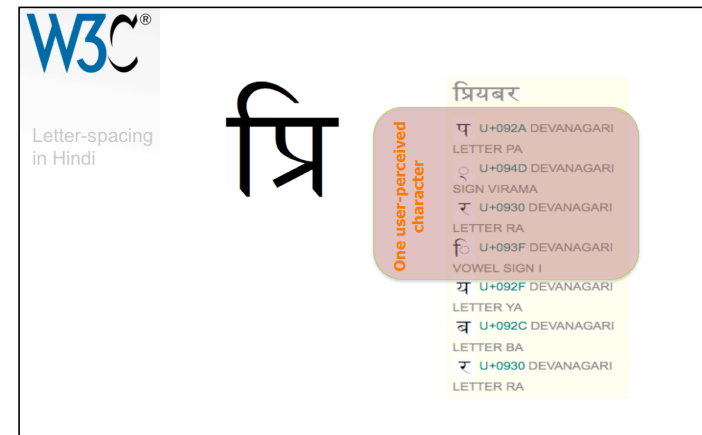







This slide shows Hindi text. Hindi can be letter-spaced, in which case you would expect it to look like the text on the second line.

However, the representation of the first syllable as a single unit depends on the availability of glyphs in the font. Change the font for a less sophisticated one and you may actually see the text as displayed on the bottom line of the slide.



The devanagari script, which is used to write Hindi, is based on syllabic structures. The list to the right shows the characters used to write the word, but note how the first four characters are combined visually to make one user-perceived unit. The characters in this unit should not be separated by letter-spacing (nor by first-letter styling, justification, line-wrapping, vertical text layout, etc).



Letter-spacing  
in Hindi

प्रि

Two grapheme clusters

प्रियबर

प	U+092A DEVANAGARI
LETTER PA	
्	U+094D DEVANAGARI
SIGN VIRAMA	
र	U+0930 DEVANAGARI
LETTER RA	
ि	U+093F DEVANAGARI
VOWEL SIGN I	
य	U+092F DEVANAGARI
LETTER YA	
ब	U+092C DEVANAGARI
LETTER BA	
र	U+0930 DEVANAGARI
LETTER RA	

The current wording in the CSS3 Text spec says that you should base unit boundaries for letter-spacing on 'grapheme clusters'. This is a concept defined by the Unicode Standard, which typically associates base characters with combining characters that follow it. Unfortunately, the current definition of grapheme-clusters leads to a segmentation of the Hindi word as shown on the bottom line of the slide and does not allow for the combination of all four initial characters into a single unit.

Unfortunately, it is not so straightforward to fix this by extending the definition of a grapheme cluster because the representation of the first syllable as a single unit depends on the availability of glyphs in the font. It is not clear how to automatically detect what kind of font is being used for display (ie. does it support the larger syllable or not) and thereby where to break the word during letter-spacing.



White-space  
processing

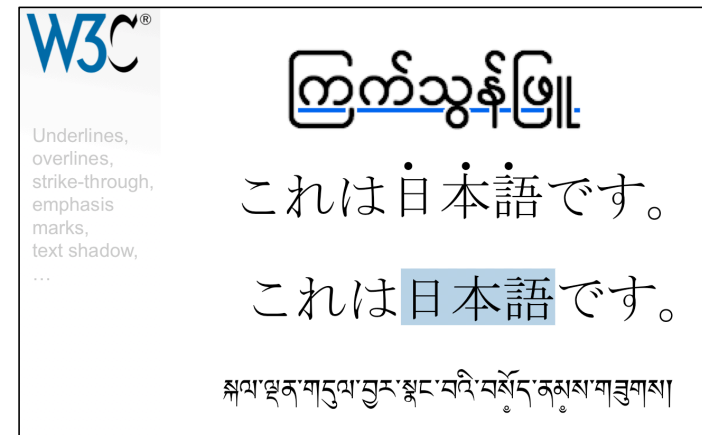
<h2>กิจกรรมระหว่างป  
ระเทศของ W3C</h2>

should the browser display a space here  
or not?

Thai is a script which doesn't use spaces to separate words. If there is content in the source code like that shown in the slide, should the two parts have a space between them when the text is displayed as part of the page? What if a line is too long to fit in the width of the editing window, but you don't want to introduce spaces?




Webkit has just this week produced support for more sophisticated drop-caps, which adds improvements for positioning but also uses grapheme-clusters as a basic unit for detecting what should be enlarged. However, use of grapheme clusters doesn't address the more complicated indic syllable which we saw earlier. What should be added? And are the positioning rules adequate for other scripts?



Has CSS correctly captured the needs for underlining of scripts, especially those where an uninterrupted line would obscure important elements of the text? What about underlining in Tibetan. Similar issues for overlines and strike-through.


What are the rules for emphasis, and how do they differ in horizontal and vertical texts? Here we see two alternative methods for Japanese, but the top one is specifically for horizontal text – there are different glyphs for vertical text. The Tibetan text at the bottom uses small symbols below a syllable to create emphasis (or sometimes as part of an annotation for commentaries). Placement is not always as straightforward as shown here, since the signs need to be centred on the syllable as a whole, rather than attached to a particular letter.



Special features:  
kumimoji,  
warichu,  
...

くみもじくみもじ  
割注(これはわりちゆです)です。

Some scripts or collections of scripts have unique typographic features, such as kumimoji and warichu in Japanese. How important is support for these? And what are the rules for use?



Ruby positioning and distribution

3 4 5 6 7 8


F1 F2

4 5 6 7


F1 F2 F3

表ㄅㄧㄠˇ 現ㄒㄩㄢˋ  
biǎo xiàn


Ruby markup support has been making good progress in browsers, but now the styling implementations must follow, and the rules need to be clarified. What are the various styles of alignment that the user will need to control? How will browsers implement bopomofo ruby, and what are the exact rules for positioning of the bopomofo and its tone marks?



Arabic styling differences within cursive runs



The logo on the left shows a colour change between two joining characters. There was recently a discussion about how to handle this in CSS. In particular, questions centered on whether style changes including font changes should be allowed. Or what to do if the CSS changes two blocks of text to inline runs.



Korean line-breaking rules, & justification strategies for embedded hanja

창녕 조씨는 신라 진평왕(眞平王)의 사위로 창성부원군(昌城府院君)에 봉군된 조계룡을 비조로 하고, 고려 태조(太祖)의 사위로 대락승(大樂丞)에 오른 조겸을 1세조로 하여 왔다. 하지만 구보(舊譜)의 대수(代數)가 같지 않아 13세(世) 소감(小監)

There are slight differences listed in the CSS Text spec for line-breaking in Japanese and Chinese. Korean shares some of the line breaking rules, but are there specific Korean differences to take into account?

Arabic justification

الدهان والمخرج والتاجر والضابط والحج والميت ويؤيئنا فوازيد عوسم اذ يهني عشر حيات ثم

conflicting opinions about right approach

والدهان والمخرج والتاجر والضابط والحج والميت ويؤيئنا فوازيد عوسم اذ يهني عشر حيات ثم

rules specific to font-styles!

There are different opinions among experts about the right way to justify Arabic text – do you stretch spaces, or do you stretch baselines? Or both? Many of the current views are based on fixed-sized pages, but we need a strategy that will survive as users stretch and shrink windows.

If cursive elongation is proposed (and it is certainly used in other contexts) what are the rules for application? And how do you handle differences between font-styles such as naskh, nastaliq and ruq'ah, when the text can switch between these depending on what font and rendering is available on their system?

[illegible]

Does CSS need to allow users to apply tsek padding at the end of lines for Tibetan justification? If so, what are the detailed rules for it's application, since its appropriateness varies, depending on things such as what ends the line?

## What next?

Aim:

To start the flow of ideas for the experts in the room about important topics to discuss.

To familiarize non-experts in the room with some typical aspects of internationalization.

Nowhere near exhaustive!

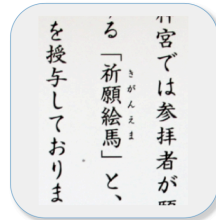


- We need to get data from other countries
  - China (ongoing)
  - Tibetan, Mongolian, Uighur (hoping)
  - SE Asia, Middle East, etc. (no movement yet)
- Organizational issues
  - Language of group, location, etc.
  - Authoritativeness of the group
- The 'magic box' syndrome



We need to ensure that newly specified features make it into browsers

- Create a gap document
- Need community lobbying
- Need script-aware developers to add support
- Communities need to use the features
- Need good tests



## You can help!

- Follow the discussions on the i18n mailing lists (eg. [www-international@w3.org](mailto:www-international@w3.org)), and track other Open Web Platform technologies for internationally relevant topics. Follow our RSS feeds and twitter channels (@webi18n)
- Read and review specifications (<http://www.w3.org/TR/tr-technology-drafts>) and send comments to the i18n list or direct to the Working Group.
- Participate in the development of current layout requirements documents: review, write, suggest, tweet about it, etc., or get your friends involved.
- Propose an expert team to work on a new set of requirements.
- Help map the requirements to the Open Web Technologies.
- Lobby the browser implementers and use the new features.



we need You to help  
make the Web  
worldwide  
get involved



Thank you  
<http://www.w3.org/International/>

Always remember that community involvement is crucial to development of W3C specifications. The W3C does not simply decide in an ivory tower to develop specifications and impose them on the public. The process only starts when we have support from the W3C member companies, experts and industry participants who will compose the Working Group. If you feel that this work is valuable, please consider participating in the Working Group.