



Internationalizing the Web

Richard Ishida
W3C Internationalization Lead

L10n or i18n?

- Establishing a standard baseline
- Extending technology to support local needs
- Authoring with local needs in mind
- Making it happen together

Localization

The **adaptation** of a product, application or document content to meet the language, cultural and other requirements of a specific target market.

Internationalization

The **design and development** of a product, application or document content that **enables** easy localization for target audiences that vary in culture, region, or language.

<http://www.w3.org/International/questions/qa-i18n>

L10n or i18n?
Establishing a standard baseline
Extending technology to support local needs
Authoring with local needs in mind
Making it happen together

جعل شبكة الويب العالمية عالمية حقاً !

締造真正全球通行的万维网

締造真正全球通行的萬維網

የዓለም አቀፉን ድር በእውነት አለም አቀፍ ግድረግ!

Κάνοντας τον Παγκόσμιο Ιστό πραγματικά Παγκόσμιο

לוצור מהרה תשרא ללל עולמית במאמץ

वर्ल्ड वाईड वेब को सचमुच विश्वव्यापी बना रहे हों !

Сделай ΔΡΑΚΟΡΑΒΟ ΡΟΛΟΙΟΤΕΛΕΟ ὁλοκλήρως.

Making the World Wide Web world wide!

ワールド・ワイド・ウェブを世界中に広げましょう

Hogy a Világháló valóban az egész világé lehessen!

वर्ल्ड वाईड वेबलाई यथार्थमे विश्वव्यापी बनाउने !

"Дүниежүзілік торды" нағыз дүниежүзілік етеміз!

전세계의 월드 와이드 웹으로 만들기!

ਵਰਲਡ ਵਾਈਡ ਵੈਬ ਨੂੰ ਵਾਕਈ ਵਿਸ਼ਵ-ਵਿਆਪੀ ਬਣਾਉਣਾ !

Сделаем "Всемирную паутину" действительно всемирной!

การทำให้ World Wide Web แพร่หลายไปทั่วโลกอย่างแท้จริง

U ita uri Webu Nyangaredzi ya Dzhango i vhe nyangaredzi ngangoho!

W3C technology must cater for all scripts and languages – even more than India. In terms of W3C technologies, English is just another language.

Only a decade ago, a page that showed all these scripts would have been difficult to create. Now things are much easier, thanks to Unicode. The W3C adopted Unicode as the document character set for HTML and XML, and tries to ensure that all its technologies support Unicode.

European alphabetic scripts

Latin
Greek
Cyrillic
Armenian
Georgian
Runic
Ogham
Modifier letters
Combining characters

East Asian scripts

Han
Hiragana
Katakana
Hangul
Bopomofo
Yi

Middle East scripts

Hebrew
Arabic
Syriac
Thaana

Symbols

Currency symbols
Letter like symbols
Mathematic operators
Numeric forms
Technical symbols
Geometrical symbols
Miscellaneous
symbols & dingbats
Enclosed & square
Braille

South & South East Asian scripts

Devanagari
Bengali
Gurmukhi
Gujurati
Panjabi
Oriya
Tamil
Telugu
Kannada
Malayalam
Sinhala
Thai
Lao
Tibetan
Myanmar
Khmer



Additional scripts

Ethiopic
Cherokee
Canadian Aboriginal
Syllabics
Mongolian

Etc....

Unicode is a single character set that covers all the commonly used scripts of the world in one place. This allows for simple display and storage of multilingual content, and for easy transitions between localized content.

Standardizing on Unicode is also helpful because so many other Web, operating system, application, database, etc environments are also working with Unicode. It is a well-known and commonly used encoding.

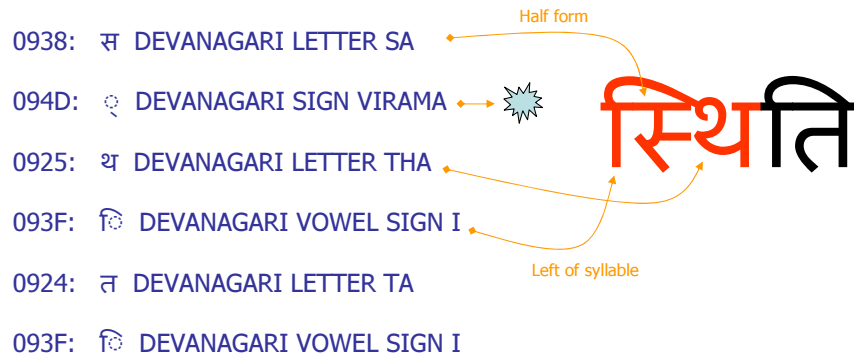
Note that you don't *have* to use Unicode to use Web technologies, but specifications are written using Unicode as the reference model, to ensure that the foundation is securely internationalized for the whole world.

Participating stakeholders:

- W3C adopted and monitors the use of Unicode as the document character set for its specifications,
- platform developers need to provide for Unicode support, such as rendering algorithms
- application developers should ensure that expected Unicode behaviors are implemented
- content developers and managers should use Unicode whenever possible
- Unicode Consortium defines Unicode and expected character-level behaviors – W3C liaises with the Unicode Consortium

The early adoption and support of Unicode at the W3C was a vital ingredient in establishing a base architecture that would enable use of the technologies around the world.

However, the decision to use Unicode as a reference model for W3C technologies is not sufficient. There are aspects related to the use of Unicode that must be borne in mind by and enabled by various participants in the deployment of the Web. Platform developers, application developers, content developers, site managers, and others need to play their part to capitalize on the potential benefits provided by an architecture based on Unicode.



Unicode has to be used with rendering algorithms for complex scripts. These are described by the Unicode Standard, but need to be enabled or integrated by developers of operating systems, applications and fonts.

This is crucial to the international Web, but note that it is not in within the mission of the W3C.

European alphabetic scripts

Latin
Greek
Cyrillic
Armenian
Georgian
Runic
Ogham
Modifier letters
Combining characters

East Asian scripts

Han
Hiragana
Katakana
Hangul
Bopomofo
Yi

Middle East scripts

Hebrew
Arabic
Syriac
Thaana

Symbols

Currency symbols
Letter like symbols
Mathematic operators
Numeric forms
Technical symbols
Geometrical symbols
Miscellaneous
symbols & dingbats
Enclosed & square
Braille

South & South East Asian scripts

Devanagari
Bengali
Gurmukhi
Gujurati
Panjabi
Oriya
Tamil
Telugu
Kannada
Malayalam
Sinhala
Thai
Lao
Tibetan
Myanmar
Khmer



Additional scripts

Ethiopic
Cherokee
Canadian Aboriginal
Syllabics
Mongolian
Tifinagh
Etc....

Some things, however, *are* of concern at the W3C. For example, XML 1.0 is based on version 2 of the Unicode Standard. These means that the red scripts above (added to Unicode since version 2) cannot be used for element and attribute names, enumerated lists, etc. Not only that, but numerous new characters have been added to scripts that did exist in version 2 (such as the Bengali ko phala), but these cannot be used in element names, etc. (Note that the use of all these scripts *is* supported in content. We are only talking about element and attribute names and the like.)

XML 1.1 provides support for all these later additions to the Unicode Standard, and the I18n Activity is encouraging developers of W3C specifications to make them support XML 1.1.

Note: All of these scripts and characters can be used in Web content, just not for element and attribute names, and a few similar constructs.

Character	Bytes
A	41
á	C3 A1
あ	E3 81 82
不	F0 A3 8E B4

In an encoding such as UTF-8 characters can be encoded using a mixture of 1 to 4 bytes. This means that when manipulating, comparing, pointing into, wrapping, or styling data, etc., you need to know where the character boundaries are, and never separate the bytes that constitute a single character.

Specification developers in W3C WGs need to consider this, but also implementers need to bear this in mind when developing applications that handle multilingual text.

a|x あ a|x あ

61 D7 90 E3 81 82 61 D7 90 E3 81 82



This sequence of slides shows how a cursor would have to jump through the bytes in memory as you press the right cursor key. The point is that applications and specifications need to recognize and respect character boundaries, not byte boundaries.

ax|あ ax あ

61 D7 90 E3 81 82 61 D7 90 E3 81 82



axあ|axあ

61 D7 90 E3 81 82 61 D7 90 E3 81 82



Establishing a standard baseline W3C®

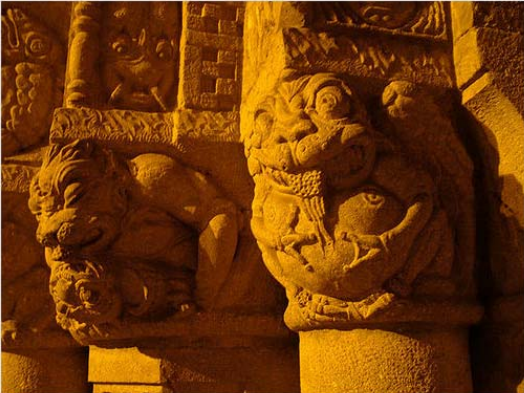
Home | Tags | Groups | People | Invite Logged in as r12a | Your Account | Help | Sign Out

Photos: [Yours](#) | [Upload](#) | [Organize](#) | [Your Contacts](#) | [Explore](#) **flickr** BETA

Chateau de La Napoule

Uploaded on June 9, 2005 by r12a

ADD NOTE | SEND TO GROUP | ADD TO SET | BLOG THIS | ALL SIZES | ORDER PRINTS | ROTATE | DELETE



r12a's photostream
1177 photos
View as slideshow

Tags
0506-cote-dazur [x]
Mandelieu-La Napoule [x]
Alpes-Maritimes [x]
France [x]

[Add a tag](#)

Additional Information
All rights reserved (change)
Taken with a Fujifilm FinePix S7000.
More properties
Taken on June 6, 2005 (edit)
See different sizes
Viewed 92 times. (Not including you)
[Edit](#) title, description, and tags
[NEW](#) [Replace](#) this photo

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 14

Developers also need to ensure that the applications, databases and scripting environments they are dealing with – especially any back-end scripting – can appropriately deal with text, and through the entire process.

This slide shows a photo uploaded to Flickr with XMP meta data in UTF-8. The Flickr user interface, *which supports UTF-8*, has taken the title of the photo from the XMP data, but some backend process has mangled the encoding. You can guess at the meaning of this title, but text in, say, Chinese, would be completely unreadable.

Be careful that the functions you use in languages such as PHP and Python can handle multibyte characters correctly, and that encoding information is recognized and appropriately dealt with.

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

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

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

	HTTP	<?xml ..	<meta ..
HTML	(✓)	x	✓
XHTML (text/html)	(✓)	(✓)	✓
XHTML (XML)	(✓)	✓	x

<http://www.w3.org/International/tutorials/tutorial-char-enc/>

Content authors must declare the encoding of content somewhere, so that it can always be discovered by any application that wants to interpret the text. Otherwise the text can be mangled before it reaches the eyes of the user.

There are a number of ways of doing this. For more information see <http://www.w3.org/International/tutorials/tutorial-char-enc/> .

Note that you must also *save* your data in the appropriate encoding – labeling alone is not sufficient (see <http://www.w3.org/International/questions/qa-changing-encoding>).

L10n or i18n?
Establishing a standard baseline
[Extending technology to support local needs](#)
Authoring with local needs in mind
Making it happen together

The last section used the example of Unicode to illustrate how the W3C has adopted a basic approach that supports international deployment.

We also noted how the W3C is just one player in the game of internationalizing your deliverables. We need to liaise with other standards bodies, and platform, application and content developers need to play their part also.

The next section will illustrate how we do also look out for ways of extending our technologies to help support international use.

Characters as ordered in memory:

The title says "תוליע פּוּא נִיבּהּ , W3C" in Hebrew.



The title says "W3C ,פּעִילוֹת הַבִּינְאוֹם" in Hebrew.

Certain types of markup that are needed to support non-Latin scripts. One important example is markup to support bidirectional text in languages based on Arabic or Hebrew scripts.

If you develop content for these languages, you must become familiar with their use (see for example <http://www.w3.org/International/articles/inline-bidi-markup/>). If you develop schemas, you should ensure that you provide such constructs for others to use.

The ITS (International Tag Set) Working Group at the W3C is currently specifying markup of this kind that can be used by any schema developer to support international use of documents, and also efficient localization of documents.

For example, this slide shows what you would expect to see displayed, given the sequence of characters in memory shown near the top of the slide.

Characters as ordered in memory:

The title says "פ ת ו ל י ע פ ם ו א נ י ב ה , W3C" in Hebrew.


The title says "W3C ,פעילות הבינאום" in Hebrew.


Using the bidi algorithm only
The title says "פעילות הבינאום" , W3C" in Hebrew.

If we rely solely on the Unicode bidirectional algorithm we will not get what we expected.

Characters as ordered in memory:

The title says "פועליות הבינאום, W3C" in Hebrew.

The title says "W3C, פעילות הבינאום" in Hebrew.



Using the bidi algorithm only

The title says "פעילות הבינאום, W3C" in Hebrew.

This slide shows markup that can be used to indicate the desired directional behavior over and above what comes with the Unicode bidirectional algorithm.

かみしばい
これは紙芝居です。

```
<ruby>  
  <rb>紙芝居</rb>  
  <rt>かみしばい</rt>  
</ruby>
```

```
<p>これは<ruby><rb>紙芝居</rb><rt>かみしばい</rt></ruby>です。</p>
```

The Ruby Annotation specification is another example of markup extensions to support international use. Ruby markup allows, for example, Japanese authors to add phonetic annotations to educational or obscure kanji texts in XHTML 1.1.

這一晚會如常舉行

這一|晚會|如常|舉行

This banquet is held as usual.

這一|晚會|如|常|舉行

If this banquet is held frequently.

這一晚|會|如常|舉行

(An event) will be held tonight as usual.

Workshops in Beijing and Crete explored international requirements for markup to support speech synthesis.

Since there are no spaces between words in Chinese, the sentence above can be read in a number of different ways. Markup to show word boundaries, when needed for disambiguation, was one of the requested enhancements at the Beijing workshop.

$$f(x) = \begin{cases} \sum_{i=1}^s x^i & \text{if } x < 0 \\ \int_1^s x^i dx & \text{if } x \in S \\ \tan \pi & \text{otherwise (with } \pi \simeq 3.141) \end{cases}$$

$$\left. \begin{array}{l} \text{مجا س ب إذا كان س > 0} \\ \text{ب = 1} \\ \text{آ س ب ء س إذا كان س \ni م} \\ \text{ظا \pi غير ذلك (مع \pi \simeq 3.141)} \end{array} \right\} = (س)$$

This slide provides some examples of differences between English and Arabic approaches to mathematical presentation. The W3C has recently produced a note about this, with a view to enabling the various Arabic approaches in the future.

We are always looking out for other requirements, related to non-Latin typography. If you are aware of things that the Web should support, please let us know.

punctuation trim

经验分
（万维

经验分
（万维

auto-space

第10回のUnicode会議

第 10 回 の Unicode 会議

emphasis

これは日本語の文章です。

これは日本語の文章です。

CSS3 holds the promise of a number of typographic approaches that are needed for non-Latin scripts, such as Chinese and Japanese, but also others. We need people familiar with these scripts and their usage to help us ensure that we haven't missed any important typographic requirements for their culture.



This slide shows a picture of vertical text on an Indian doorway that I came across recently. We will need to check that the vertical text properties in CSS take into account that the text proceeds downwards syllable by syllable, not letter by letter.

When these new typographic features are available and supported in user agents, developers and content authors will need to familiarize themselves with the numerous properties that are available.

Note also that it is not sufficient to specify these extensions at the W3C – if you want to ensure that these things are available for use, you should push developers of user agents to include support for them, and ensure that you use the features when they become available.

ಯೆ ಹೋವನು ಇಡೀ ಲೋಕದ
ಭಾಷೆಯನ್ನು ತಾರು ಮಾರು
ಮಾಡಿದ್ದು ಆ ಸ್ಥಳದಲ್ಲೇ. ಆದ್ದರಿಂದ
ಆ ಸ್ಥಳಕ್ಕೆ ಬಾಬಿಲೋನ್ ಎಂದು
ಹೆಸರಾತು. ಹೀಗೆಯೆ ಹೋವನು
ಜನರನ್ನು ಆ ಸ್ಥಳದಿಂದ ಭೂಮಿಯ
ಲ್ಲೆಲ್ಲಾ ಚದರಿಸಿಬಿಟ್ಟನು.

OCAF: ಯ KANNADA
LETTER YA

OCC6: ೆ KANNADA
VOWEL SIGN E

We are still exploring certain issues, such as first-letter styling (useful for styling the first letter of a paragraph, as above, without adding markup). In the Kannada example above, the 'first letter' is composed of more than one character – and in conjunct consonants with vowel signs, could be several characters.

CDAC is currently looking at the proposals for the next version of the CSS spec and checking them against requirements for the Indian scripts.

والدهان والمخرج والتاجر
والضابط والحي والميت، ويومئ
الينا فو از حديد عوسم ٤٩ اذ
يهيئ حسن فكرت عشر
مسرحيات، ثم يومئ اذ تراود

As this and the next slide show, Arabic justification stretches words rather than spaces. Another example of script-differentiated behavior.

والدهان والمخرج والتاجر
والضابط والحي والميت، ويومئ
الينا فو از حديد عوسم ٤٩ اذ
يهيئ حسن فكرت عشر
مسرقيات، ثم يومئ اذ تراود

L10n or i18n?
Establishing a standard baseline
Extending technology to support local needs
[Authoring with local needs in mind](#)
Making it happen together

This section sets out to remind you that the development of specifications is not the end of the issue. Those specifications need to be used sensibly by content developers.

You are speaking to her from my new house.

Están hablándole desde mi casa nueva.

私の新しい家から彼女と話しています。

تكلّمونها من بيتي الجديد

This slide shows the same idea expressed in multiple languages. Within each translation of the sentence, the number of words is different, and the order of those words changes.

There were %d spelling mistakes in file: %s.

Datei %s enthält %d Rechtschreibfehler.

```
printf( "There were %d spelling mistakes in file %s.",
currentpage, totalpages)
```



```
printf( "There were %1$d spelling mistakes in file %2$s .",
currentpage, totalpages)
```




```
printf( "Datei %2$s enthält %1$d Rechtschreibfehler.",
currentpage, totalpages)
```

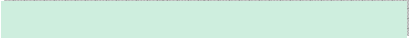
This is an example of syntax differences affecting development techniques.


The order of variables needs to be different between English and German versions. Unless you are using slightly more advanced techniques in PHP (shown in the lowest two lines), you will prevent this possibility and seriously affect translatability.

It's a question of the content developer using the technology wisely.

Interface Language 

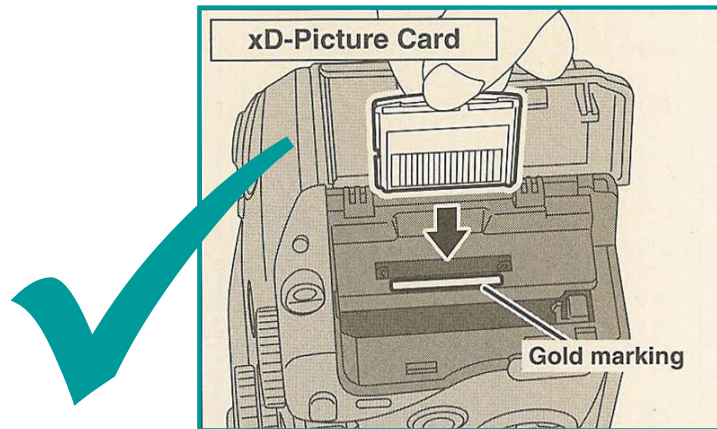
Sprache der Benutzeroberfläche 

Interface Language 

Sprache der Benutzeroberfläche 

English and Chinese text usually expand when translated. You should consider the potential impact of this on page design, and either allow text to flow into larger areas, or leave expansion space.

For example, putting labels beside form fields is often likely to cause expansion space problems. This issue can often be avoided by allowing text to expand above the field, instead.



Symbolism can differ from place to place. For example the check mark means *incorrect* in some places around the world.

Ensure that you do not give the wrong message through your use of colors, symbolism, examples, etc.



Here, in Japan, the circles mean the same as the check mark – they are not zeros!

Website Directory Sites organised by subject.

- Arts & Humanities**
Literature, History, Photography...
- Business & Economy**
B2B, Shopping, Investments, Property...
- Computers & Internet**
Internet, Reviews, Software, Games...
- Education**
UK, Ireland, Universities...
- Entertainment**
Humour, Movies, Music, Actors...
- Government**
UK, Ireland, Politics, Law...
- Health**
Medicine, Drugs, Diseases, Fitness...
- News & Media**
Newspapers, Weather, TV...
- Recreation & Sport**
Sport, Hobbies, Travel, Motoring...
- Reference**
Maps, Dictionaries, Phone Numbers...
- Regional**
UK, Ireland, Countries...
- Science**
Animals, Geography, Engineering...
- Social Science**
Economics, Languages, Psychology...
- Society & Culture**
People, Food & Drink, Environment, Sexuality...

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 34

This and the following slides show how Yahoo adapts its categorizations to reflect the preoccupations of various different countries. So it doesn't only translate the text, it also adapts content for the local audience.

Guide Web - Classement thématique de sites [Suggérer un site](#)

Actualités et médias
[Journaux](#), [Télévision](#), [Météo](#)...

Commerce et économie
[B2B](#), [Shopping](#), [Emploi](#),
[Immobilier](#)...

Informatique et Internet
[Internet](#), [Logiciels](#), [Fonds d'écran](#)...

Santé
[Diététique](#), [Médecine](#), [Thalasso](#)...

Enseignement et formation
[Primaire](#), [Secondaire](#), [Supérieur](#)...

Institutions et politique
[Ministères](#), [Droit](#), [Politique](#)...

Sciences et technologies
[Animaux](#), [Astronomie](#), [Physique](#)...

Sports et loisirs
[Foot](#), [Tourisme](#), [Auto/Moto](#), [Jeux](#)...

Art et culture
[Littérature](#), [Cinéma](#), [Musique](#), [BD](#)...

Divertissement
[Tests/Quiz](#), [Loteries](#), [Humour](#),
[Sorties](#)...

Classement géographique
[Pays](#), [Europe](#), [France](#), [Paris](#)...

Références et annuaires
[Dictionnaires](#), [Annuaire](#),
[Cartes/Atlas](#)...

Société
[Enfants](#), [Gastronomie](#),
[Rencontres](#)...

Sciences humaines
[Archéologie](#), [Histoire](#), [Psychologie](#)...

[Parviva](#)
Barcelone 2004?

[Yahoo! Mail](#) - FREE!
1' 56M KB

[in deadly Iraq](#)
[r more addresses](#)
0 in Riyadh
[man assaulted in](#)
[r auction](#)
1 of public-service
on sex in
July - March
24% Nasdaq 1000
inance - Weather

id's mobilie

List

Jobs - Property

on? Head over to
not get the hottest
loads to help you.

news
La 15% online
to get a quote

ions League
s in letters. Read
love to say about

es - Games - TV

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 35

The screenshot shows the Yahoo! Japan homepage with the following categories and sub-items:

Yahoo!カテゴリ	サイトの推薦
エンターテインメント 映画, 音楽, 芸能人, コミック, 占い ...	メディアとニュース テレビ, ラジオ, 新聞, 雑誌 ...
趣味とスポーツ アウトドア, ゲーム, 車, スポーツ, 旅 ...	ビジネスと経済 ショッピング, B2B, 雇用, 金融 ...
芸術と人文 写真, 建築, 美術館, 歴史, 文学 ...	各種資料と情報源 図書館, 辞書, 郵便, 電話番号 ...
生活と文化 子ども, 環境, グルメ, 障害者 ...	コンピュータとインターネット ハードウェア, ソフトウェア, WWW ...
教育 大学, 専門学校, 小中高, 資格 ...	政治 政治, 行政, 国会, 法 ...
健康と医学 病院, 病気, ダイエット ...	自然科学と技術 動物, エコロジー, 地球, 天文, 工学 ...
社会科学 経済学, 社会学, 言語, 政治学 ...	地域情報 日本の地方, 世界の国 ...

An orange arrow points to the '趣味とスポーツ' category. The footer contains the text: Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 36

L10n or i18n?
Establishing a standard baseline
Extending technology to support local needs
Authoring with local needs in mind
Making it happen together

- i18n is more than just translation – it relates to enabling technologies to support local needs
- different groups of people are involved in helping achieve Web internationalization
 - W3C drives fundamental Web-related technologies, but also works with other standards organizations
 - application and platform developers have to be encouraged to implement standards but also to enable other aspects of internationalization support
 - content developers should use standards and observe best practices for internationalization, but also consider necessary local adaptations

- Help Working Groups understand international requirements as early as possible
- Check specifications in Working Drafts, especially at Last Call, for internationalization issues
- Define, or work with other Working Groups to define, behavior needed for support of international requirements
- Evangelize the need to consider multiple languages and scripts when developing Web technologies of any kind
- Helping users of Web technology understand what's available to them and how to use it

- Provide resources to collaborate in developing the world wide aspects of the Web
 - not just the i18n WG: represent India in WGs dealing with Web Services, Semantic Web, CSS, XSL, HTML, SVG, MathML, Voice, Accessibility, etc...
- Study the specifications and raise issues / requirements specific to Indian languages
 - esp. style related: eg. first-letter, vertical text, etc...
- Encourage users to do the right thing
 - help create articles and advice, spread the word, translate, etc...
- Follow and influence groups external to the W3C
 - IDN and IRIs, language tags, Unicode, etc...

- ◆ this is your Web – not the W3C's – if something isn't right, get involved to fix it
- ◆ you need to ensure that your concerns are being dealt with
- ◆ don't leave it too late!

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