



An Introduction to Internationalization


Richard Ishida
W3C Internationalization Lead

Objectives


You will be able to tell your friends and colleagues:

- Why localization is not just a question of grabbing a technical guy to translate stuff
- Why you need to think about localization earlier than people typically expect
- Insights into internationalization at the W3C

Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 2



Overview




W3C's I18n Activity

- L10n or i18n?
- Content vs. presentation
- I18n overview
 - Characters
 - Document formats
 - Presentation matters
 - Practical barriers
 - Cultural differences
- Summary

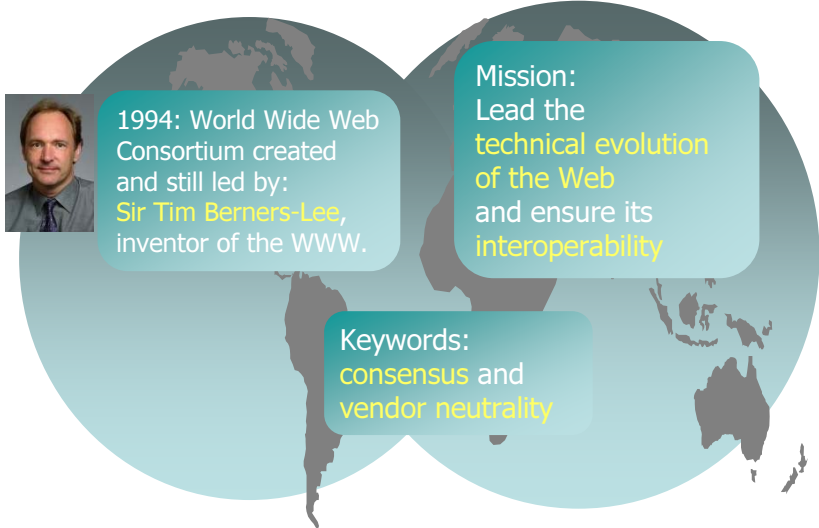
Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 3



W3C Internationalization Activity

About the W3C

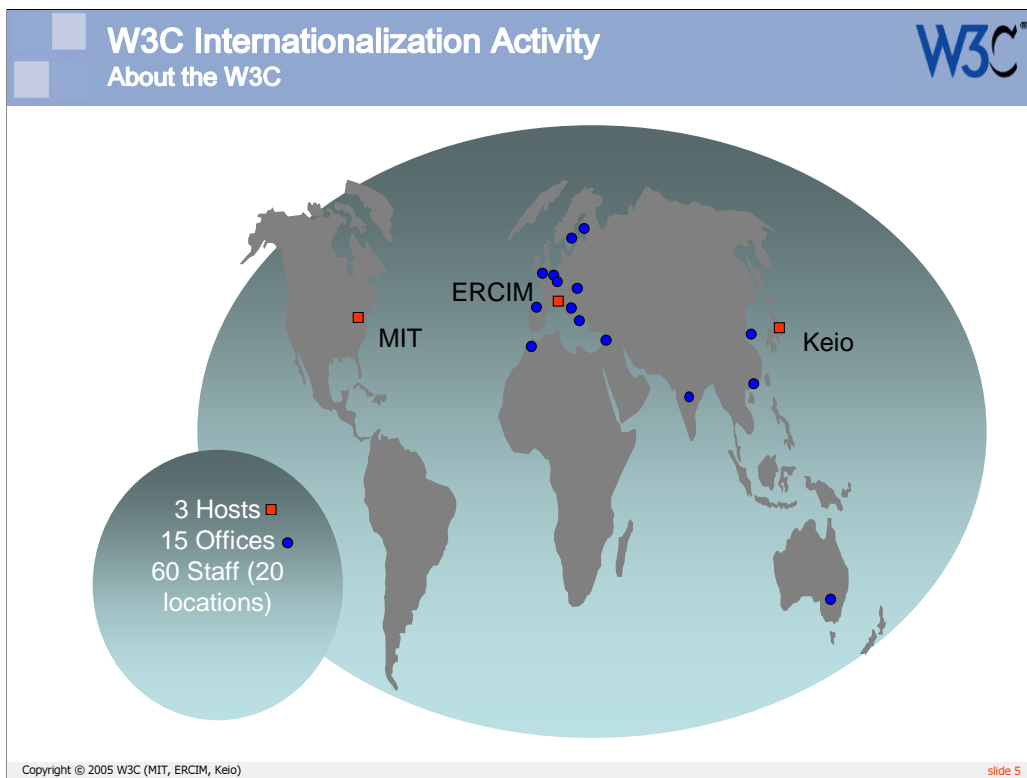


1994: World Wide Web Consortium created and still led by: **Sir Tim Berners-Lee**, inventor of the WWW.

Mission: Lead the **technical evolution of the Web** and ensure its **interoperability**

Keywords: **consensus** and **vendor neutrality**

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 4



W3C Internationalization Activity

About the W3C






Copyright © 2005 W3C (MIT, ERCIM, Keio)

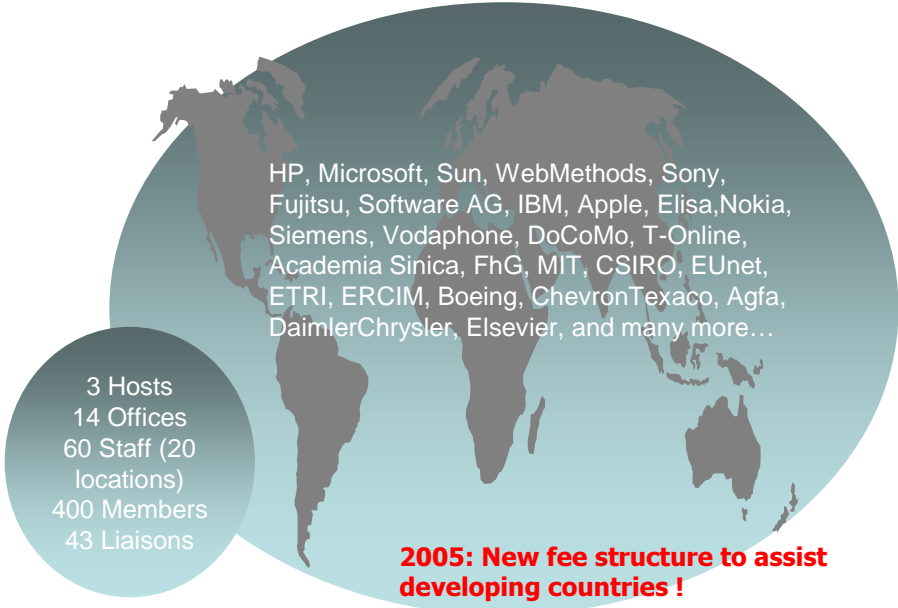
slide 6

We are real (and very nice) people!



W3C Internationalization Activity

About the W3C



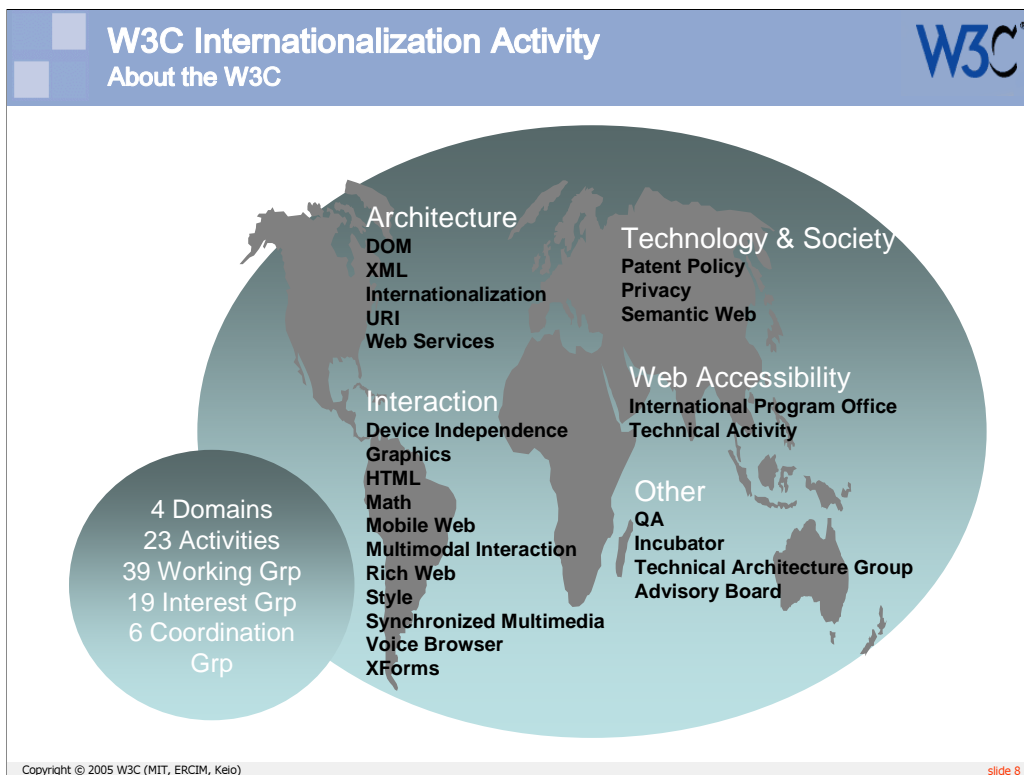
HP, Microsoft, Sun, WebMethods, Sony, Fujitsu, Software AG, IBM, Apple, Elisa, Nokia, Siemens, Vodaphone, DoCoMo, T-Online, Academia Sinica, FhG, MIT, CSIRO, EUnet, ETRI, ERCIM, Boeing, ChevronTexaco, Agfa, DaimlerChrysler, Elsevier, and many more...


3 Hosts
14 Offices
60 Staff (20 locations)
400 Members
43 Liaisons

2005: New fee structure to assist developing countries !

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 7





W3C Internationalization Activity
Internationalization Activity

Internationalization Activity


Core Working Group
Reviews, advice, and internationalization specifications

ITS (Internationalization Tag Set) Working Group
Elements and attributes for schema developers


GEO (Guidelines, Education & Outreach) Working Group
Making internationalization aspects of W3C technology better understood and more widely and consistently used

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 9



Overview



W3C's I18n Activity

L10n or i18n?

Content vs. presentation


I18n overview

- Characters
- Document formats
- Presentation matters
- Practical barriers
- Cultural differences


Summary

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 10



L10n or i18n?



- "Visitors linger twice as long as they do at English-only URL's.
- Business users are 3 times more likely to buy when addressed in their language.
- Customer service costs drop when instructions are displayed in the user's native language."

'Strategies for Global Sites'
Donald DePalma
Forrester Research Inc.

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 11



L10n or i18n?



"One large IT company discovered that a significant percentage of inquiries were coming from South Korea - they created a Korean website and revenues rose by 8 percent."

'Global eCommerce'
Donald J. Plumley
Bowne Global Solutions

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 12

L10n or i18n?

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>

Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 13

Localization without internationalization can be very hard. This presentation will use examples to make that point, and stress the value of considering internationalization as an integral part of the design and development activity – not an afterthought left to the 'localization folks'.



Overview


W3C's I18n Activity
L10n or i18n?
Content vs. presentation
I18n overview

- Characters
- Document formats
- Presentation matters
- Practical barriers
- Cultural differences

Summary

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 14

Separating content & presentation


Content (XHTML)

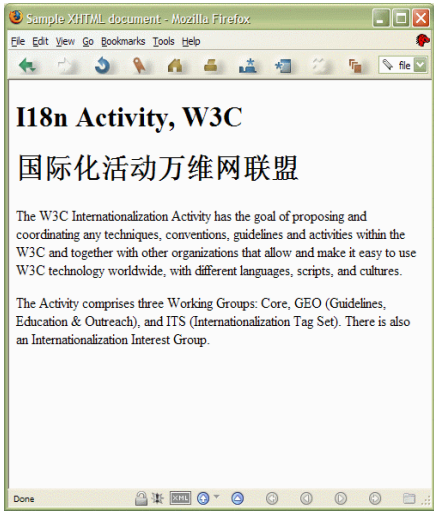
```

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<head>
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
  <title>About the W3C I18n Activity</title>
  <style type="text/css" src="mystyling.css" />
</head>

<body>
  <h1>I18n Activity, W3C</h1>
  <div class="international-text" xml:lang="zh-Hans"
        lang="zh-Hans">国际化活动万维网联盟</div>
  <div class="description">
    <p>The W3C Internationalization Activity has the goal of proposing
    and coordinating any techniques, conventions, guidelines and
    activities within the W3C and together with other organizations
    that allow and make it easy to use W3C technology worldwide,
    with different languages, scripts, and cultures.</p>
    <p>The Activity comprises three Working Groups: Core, GEO
    (Guidelines, Education & Outreach), and ITS (Internationalization
    Tag Set). There is also an Internationalization Interest Group.</p>
  </div>
</body>
</html>

```



Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 15

The HTML is shown on the left. There is no presentational information in the HTML – which is as it should be.

To the right is some CSS code that applies styling to the HTML.

Separating content & presentation
W3C[®]

Content (XHTML)

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
    <title>About the W3C I18n Activity</title>
    <style type="text/css" src="mystyling.css" />
  </head>

  <body>
    <h1>I18n Activity, W3C</h1>
    <div class="international-text" xml:lang="zh-Hans"
        lang="zh-Hans">国际化活动万维网联盟</div>
    <div class="description">
      <p>The W3C Internationalization Activity has the goal of proposing
        and coordinating any techniques, conventions, guidelines and
        activities within the W3C and together with other organizations
        that allow and make it easy to use W3C technology worldwide,
        with different languages, scripts, and cultures.</p>
      <p>The Activity comprises three Working Groups: Core, GEO
        (Guidelines, Education & Outreach), and ITS (Internationalization
        Tag Set). There is also an Internationalization Interest Group.</p>
    </div>
  </body>
</html>
```

Presentation (CSS)

```
body {
    background: white;
    color: black;
    font-family: serif;
    font-size: 1em;
}

h1 {
    font-size: 240%;
}

div.international-text {
    font-family: MingLiu, sans-serif;
    font-size: 240%;
}

p {
    margin-top: 1em;
}
```

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 16

The HTML is shown on the left. There is no presentational information in the HTML – which is as it should be.

To the right is some CSS code that applies styling to the HTML.



Each of these windows shows EXACTLY the same HTML file. The changes made to the CSS file produced three very different presentations of that basic content.

This is particularly useful for changing the presentational aspects of a site or group of pages. You typically only need to edit a single CSS file, rather than editing all the code of each HTML file.

This can also be beneficial for localization, since typographic approaches, colors, etc, may need to be changed for different locales. Making such changes in the CSS is much easier than adapting the HTML.

W3C®

Separating content & presentation




Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 18

Remember, also, that the Mobile Web is becoming increasingly important these days – and may be especially so in developing countries in the future. This means that content needs to be adapted to fit on handheld devices with smaller screens.

Again, this would ideally be achieved by styling the content, rather than writing a completely separate Web.

You should not make assumptions, when creating content, that you know what it will look like when finally displayed. These days, it may well be displayed in a number of different formats.

W3C[®]

Separating content & presentation

International issues


- ◆ problems of resolution to support bold and italics in small CJK characters on-screen
- ◆ different ways of emphasizing text in Japanese (wakiten & amikake)

これは日本語です。

これは日本語です。


Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 19

Here are some ways in which typographic differences may appear between language versions of the same content.



Separating content & presentation

International issues



- ◆ problems of resolution to support bold and italics in small CJK characters on-screen
- ◆ different ways of emphasizing text in Japanese (wakiten & amikake)
- ◆ no upper- vs. lower-case distinction in most non-Latin scripts
- ◆ no convention of distinguishing between proportional and mono-spaced fonts for some scripts

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 20

W3C[®]

Separating content & presentation
Practical implications


Making the World Wide Web *worldwide*.

<p>Making the World Wide Web <i>worldwide</i></p>

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 21


You should try to remove all presentational constructs from your content. For example, use of `<i>` tags shows that you are assuming that the text will be italicized. Because ideographic text doesn't support italicizations well in small font sizes, you could be causing problems for localization.

W3C[®]

Separating content & presentation
Practical implications

Making the World Wide Web *worldwide*.


```
<p>Making the World Wide Web <em>worldwide</em></p>
```



Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 22

Not only is it better for localization to express the idea or semantics in the content, and leave the presentation to the style sheet, it will also improve your original text by making you more aware of what you are actually doing.



Separating content & presentation

Practical implications


See the **System Administrator Guide** for an example of re-use.

`<p>See the System Administrator Guide`
`for an example of re-use.</p>`

Copyright © 2005 W3C (MIT, ERCIM, Keio)


slide 23

The same applies to document conventions such as representation of referenced resources. When using class annotations or microformats, don't describe the expected presentational rendering, describe the function of the text.



Separating content & presentation


Practical implications




See the **System Administrator Guide** for an example of re-use.

`<p>See the System Administrator Guide for an example of re-use.</p>`


doctitle
chaptertitle
inputsequence
etc.



Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 24



Overview



W3C's I18n Activity
L10n or i18n?
Content vs. presentation
I18n overview
 Characters
 Document formats
 Presentation matters
 Practical barriers
 Cultural differences
Summary

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 25

118n Overview: Characters

Character sets & encodings

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 ማድረግ!

Κάνοντας τον Παγκόσμιο Ιστό πραγματικά Παγκόσμιο

ליצור מהרשת כלל עולמית באמת

वर्ल्ड वाईड वेब को सचमुच विश्वव्यापी बना रहें हैं !

Сделав ΔΡΑCΠΑC CΛCΓC CΛCΓC.

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!

전 세계의 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!

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 26

English is just another language.

This kind of multilingual text on a single page was very rare only 10 years ago.



I18n Overview: Characters

Character sets & encodings

	0	1	2	3	4	5	6	7
0				0	@	P	`	p
1			!	1	A	Q	a	q
2			"	2	B	R	b	r
3			#	3	C	S	c	s
4			\$	4	D	T	d	t
5			%	5	E	U	e	u
6			&	6	F	V	f	v
7			'	7	G	W	g	w
8			(8	H	X	h	x
9)	9	I	Y	i	y
A			*	:	J	Z	j	z
B			+	;	K	[k	{
C			,	<	L	\	l	
D			-	=	M]	m	}
E			.	>	N	^	n	~
F			/	?	O	_	o	

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 27

Early character sets based on 7-bit bytes, gave 2^7 (ie. 128) possible characters.

I18n Overview: Characters

Character sets & encodings



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0			0	@	P	`	p				°	À	Ð	à	ð	
1		!	1	A	Q	a	q			i	±	Á	Ñ	á	ñ	
2		"	2	B	R	b	r			¢	²	Â	Ò	â	ò	
3		#	3	C	S	c	s			£	³	Ã	Ó	ã	ó	
4		\$	4	D	T	d	t			¤	´	Ä	Ö	ä	ö	
5		%	5	E	U	e	u			¥	µ	Å	Ø	å	ø	
6		&	6	F	V	f	v			¦	¶	Æ	Ö	æ	ö	
7		'	7	G	W	g	w			§	·	Ç	×	ç	÷	
8		(8	H	X	h	x			"	,	È	Ø	è	ø	
9)	9	I	Y	i	y			©	¹	É	Ù	é	ù	
A		*	:	J	Z	j	z			ª	º	Ê	Ú	ê	ú	
B		+	;	K	[k	{			«	»	Ë	Û	ë	û	
C		,	<	L	\	l				¬	¼	Ì	Ü	ì	ü	
D		-	=	M]	m	}			-	½	Í	Ý	í	ý	
E		.	>	N	^	n	~			®	¾	Î	Þ	î	þ	
F		/	?	O	_	o				™	¿	Ï	ß	ï	ÿ	

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 28

Adding an 8th bit gave a total of 256 possible characters. Still this was not enough for all European needs.

I18n Overview: Characters

Character sets & encodings




	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0			0	@	P	`	p				°	ı	ı	ı	ı	ı
1		!	1	A	Q	a	q			ˆ	±	A	P	α	ρ	
2		"	2	B	R	b	r			ˆ	²	B	□	β	ç	
3		#	3	C	S	c	s			£	³	Γ	Σ	γ	σ	
4		\$	4	D	T	d	t			¤	´	Δ	T	δ	τ	
5		%	5	E	U	e	u			¥	μ	E	Y	ε	v	
6		&	6	F	V	f	v			¦	¶	Z	Φ	ξ	φ	
7		'	7	G	W	g	w			§	·	H	X	η	χ	
8		(8	H	X	h	x			¨	ˆ	Θ	Ψ	θ	ψ	
9)	9	I	Y	i	y			©	ˆ	I	Ω	ι	ω	
A		*	:	J	Z	j	z			□	ˆ	K	ı	κ	ı	
B		+	;	K	[k	{			«	»	Λ	Υ	λ	υ	
C		,	<	L	\	ı				¬	ˆ	Ο	Μ	α	μ	ό
D		-	=	M]	m	}			–	½	N	Ξ	ν	ύ	
E		.	>	N	^	n	~			®	ˆ	Υ	Ξ	ή	ξ	ώ
F		/	?	O	_	o				–	Ω	Ο	ı	ο	□	

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 29

The code page mechanism, where the meaning of the upper cells was changed according to context helped a little, but was very messy.

It still didn't come close, however, to addressing the needs of the Far East, where the character sets had to incorporate thousands of ideographic characters at a time.



I18n Overview: Characters

Character sets & encodings

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....



Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 30

Unicode solves this problem.

It 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.


I18n Overview: Characters					W3C®	
Character sets & encodings						
	A	κ	好	不		
Code point	41	5D0	597D	233B4		
Encodings	UTF-8	41	D7 90	E5 A5 BD	F0 A3 8E B4	
	UTF-16	00 41	05 D0	59 7D	D8 4C DF B4	
	UTF-32	00 00 00 41	00 00 05 D0	00 00 59 7D	00 02 33 B4	

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 31


An 'encoding' refers to the way that characters are mapped from the character set to bytes in the computer. Different encodings yield different byte sequences.

To emphasize that character sets and encodings are different things, note how Unicode has three possible encodings, even though the actual character set is just defined once. In order to correctly interpret byte sequences and convert them into the right characters, you need to know what encoding was used.



I18n Overview: Characters

Working with characters



```
<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	(✓)	✗	✓
XHTML (text/html)	(✓)	(✓)	✓
XHTML (XML)	(✓)	✓	✗

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

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 32

You must declare the encoding of your content somewhere, so that it can always be discovered by any application that wants to interpret the text.

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 – labelling alone is not sufficient (see <http://www.w3.org/International/questions/qa-changing-encoding>).

I18n Overview: Characters	
Working with characters	
	W3C®
Character	Bytes
A	41
á	C3 A1
あ	E3 81 82
不	F0 A3 8E B4
Copyright © 2005 W3C (MIT, ERCIM, Keio)	
slide 33	

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.

I18n Overview: Characters

Working with characters

W3C®

a|kあa|kあ


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

↑

Copyright © 2005 W3C (MIT, ERCIM, Keio)


slide 34


This sequence of slides shows how a cursor would have to jump through the bytes in memory as you press the right cursor key.



I18n Overview: Characters

Working with characters





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

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 35



I18n Overview: Characters

Working with characters



ax あ | ax あ

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

↑

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 36

I18n Overview: Characters

Working with characters

[Home](#) | [Tags](#) | [Groups](#) | [People](#) | [Invite](#)

Logged in as [r12a](#) | [Your Account](#) | [Help](#) | [Sign Out](#)

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

Château de La Napoule

Uploaded on June 9, 2005
by [r12a](#)

[r12a's photostream](#)

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 37

You need to ensure that the applications you are dealing with – especially any back-end scripting – can appropriately deal with text.

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.

W3C[®]

I18n Overview: Characters
 Working with characters

NFC

NFD


Ízelítőül

Iózelítóóüö

Ha a világ beszélni akarna, Unicode-ul szólalna meg.
 Regisztráljon már most a Tizedik Nemzetközi Unicode
 Konferenciára, melyet 1997. március 10-12-én rendeznek
 Mainz-ban, Németországban. Ezen a konferencián az iparág
 több neves szakértője is részt vesz. Ízelítőül a témákból: a
 világháló és a Unicode nemzetköziesítése és lokalizálása, a
 Unicode alkalmazása működő rendszerekben és
 alkalmazásokban, szövegelrendezésnél, és többnyelvű
 számítógépeken.

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 38

If you are running processes on text, you may also want to normalize the text beforehand to make it easier to collate character sequences in Unicode that are different but canonically equivalent.



I18n Overview: Characters

Multi-script Web addresses

<http://raksmorgas.josefsson.org/mal/franzen.html>

<http://räksmörgås.josefsson.org/mål/franzén.html>

Easier to create

- ... memorize
- ... transcribe
- ... interpret
- ... guess / find things
- ... relate to (branding)

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 39

There is a lot of demand for people to be able to use non-ASCII characters in Web addresses.

I18n Overview: Characters

Multi-script Web addresses

W3C

http://raksmorgas.josefsson.org/mal/franzen.html

http://räksmörgås.josefsson.org/mål/franzén.html

domain name path


http://rksmrgrs-5wao lo.josefsson.org/m%C3%A5l/franz%C3%A9n.html

- Phishing (www.paypal.com)


Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 40

New standards have come out of the IETF recently that make this possible. The W3C personnel contributed to the development of these standards.

There are still some hurdles to overcome with regard to security and deployment, but it is possible to use these now. For more information see <http://www.w3.org/International/articles/idn-and-iri/> .



Overview



W3C's I18n Activity
L10n or i18n?
Content vs. presentation
I18n overview
 Characters
 Document formats
 Presentation matters
 Practical barriers
 Cultural differences
Summary

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 41

W3C[®]

I18n Overview: Document formats

Declaring the language of text

HTTP Content-Language header

Language attribute on html tag

Content-Language meta tag

Language attribute on embedded element

```

HTTP/1.1 200 OK
Date: Wed, 05 Nov 2003 10:46:04 GMT
Server: Apache/1.3.28 (Unix) PHP/4.2.3
...
Content-Type: text/html; charset=utf-8
Content-Language: en

<html lang="en">
<head>
...
<meta http-equiv="Content-Language" content="en" />
...
</head>
<body>
<p>The French word for <em>cat</em> is
<em lang="fr">chat</em>.</p>
...
</body>
</html>
                    
```


Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 42

Applications exist that can use natural language information about content to deliver to users the most relevant information or styling according to their language preferences. The more content is tagged and tagged correctly, the more useful and pervasive such applications will become.

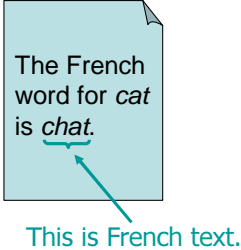
There are a number of possible ways to declare language information in HTML, but the effectiveness and the rules that apply to each approach vary. For more information see <http://www.w3.org/TR/i18n-html-tech-lang/>.


I18n Overview: Document formats

Declaring the language of text



- ◆ Text-processing language
 - ◆ the language of a specific range of text
 - ◆ used for processing such as text-to-speech, styling, etc.
 - ◆ can indicate only ONE language at a time
- ◆ Primary language metadata
 - ◆ describes the language(s) of the document as a whole
 - ◆ not a list of all languages used in the document
 - ◆ could be more than one language





Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 43

In particular, it is important to recognize that there are two different types of language declaration. Different mechanisms (shown on the previous page) naturally fall into one or other of the different types.

I18n Overview: Document formats

Script-specific markup

W3C

Characters as ordered in memory:

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



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

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 44

In addition to language declarations, there are other 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 that can be used to support international use of documents, and also efficient localization of documents.

W3C[®]

I18n Overview: Document formats
 Script-specific markup

Characters as ordered in memory:

The title says "פ ת ו ל י ע פ , W3C" in Hebrew.



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



Using the bidi algorithm only

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

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 45

In addition to language declarations, there are other 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 that can be used to support international use of documents, and also efficient localization of documents.

W3C[®]

I18n Overview: Document formats
 Script-specific markup

Characters as ordered in memory:

The title says "תוליעפ , W3C" in Hebrew.



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

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 46

In addition to language declarations, there are other 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 that can be used to support international use of documents, and also efficient localization of documents.

W3C®

I18n Overview: Document formats

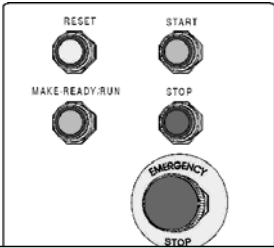
Markup to support localization

At the VISTA console, submit a job to print. (Refer to "Submitting a Job" in Chapter 5.)

At the operator control panel, make sure the printing system is in Make-Ready mode. The MAKE-READY/RUN indicator should not be lit.

➔ Press the START button to sound the horn. The MAKE READY / RUN indicator flashes.

At the third beep, press the START button again. The START indicator remains lit and beeps.



```
<para>
  Press the
  <span translate="no">START</span>
  button to sound the horn. The
  <span translate="no">MAKE-READY/ RUN</span>
  indicator flashes.
</para>
```

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 47

An example of markup that can help make translation more efficient is the provision of a flag to indicate whether or not text should be translated. This can be used by translation tools to screen text from translators or machine translation systems where necessary.

In this example of product documentation, 'START' and 'MAKE-READY/RUN' appear on a hard panel that will not be translated. The markup can be used to indicate that. In actuality, the ITS group will come up with a number of ways of implementing a translate flag. In some cases these may be used by content authors, in other cases they may be applied via rules. For more detail, follow the development of the working draft at <http://www.w3.org/TR/its/>.

I18n Overview: Document formats

Markup to support localization



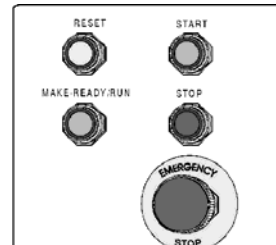
Von der VISTA-Konsole aus einem Druckauftrag übermitteln. (Siehe hierzu "Auftrag übergeben" in Kapitel 5.)

Am Steuerpult prüfen, ob der Make-Ready-Modus aktiv ist. (Die Anzeige MAKE-READY/RUN darf nicht leuchten).



START drücken, so dass die Hupe ertönt und die Anzeige MAKE READY / RUN blinkt.

Beim dritten Ton erneut START drücken. Die Anzeige START leuchtet konstant und der



```
<para>
  Press the
  <span translate="no">START</span>
  button to sound the horn. The
  <span translate="no">MAKE-READY/ RUN</span>
  indicator flashes.
</para>
```


I18n Overview: Document formats

Avoid text in attributes, and other such useful advice



Volcanic eruptions have literally devastated large inhabited areas. During the 1914 eruption of Sakurajima in Kyushu, 687 houses in Kurokami were buried in hot ash. What remained of this shrine gate, previously five meters tall, was left as a reminder.



*Kurokami maibutsu gate
(腹五社神社黒神埋没鳥居),
Sakurajima Island.*

```
<image src="kk-torii.jpg" height="180" width="240"
caption="Kurokami maibutsu gate (腹五社神社黒神埋
没鳥居), Sakurajima Island."/>
```

Can't mark up for language, bidirectional markup, abbreviation, styling, etc.

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 49


In some cases, an approach to schema design is important, rather than specific tags. For example, the Japanese text in an attribute value shown here cannot be marked up for language, directionality, abbreviation, styling, etc, since it is part of the attribute text.

W3C[®]

I18n Overview: Document formats

Avoid text in attributes, and other such useful advice

Volcanic eruptions have literally devastated large inhabited areas. During the 1914 eruption of Sakurajima in Kyushu, 687 houses in Kurokami were buried in hot ash. What remained of this shrine gate, previously five meters tall, was left as a reminder.



Kurokami maibutsu gate
(腹五社神社黒神埋没鳥居),
Sakurajima Island.

```


<image src="kk-torii.jpg" height="180" width="240">
  <caption>Kurokami maibutsu gate
  (<span xml:lang="ja">腹五社神社黒神埋没鳥居</span>),
  Sakurajima Island.</caption>
</image>
```

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 50


It would have made more sense to use an element for the caption.

The ITS Working Group will also provide advice of this kind to schema developers.

The I18n Core Working Group has also discussed concepts such as this with other W3C working groups. For example, XHTML 2 will hopefully address a number of situations in HTML where text cannot be marked up appropriately.



Overview



W3C's I18n Activity
L10n or i18n?
Content vs. presentation
[I18n overview](#)
 Characters
 Document formats
 [Presentation matters](#)
 Practical barriers
 Cultural differences
Summary

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 51

I18n Overview: Presentation matters

Character glyph rendering

W3C®

Character

vs.

Glyph

a *a*

雪 雪

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 52

The term 'glyph' refers to the visual representation of an underlying character.

There is only one character code for the Latin letter *a*, but the visual representation can have many different forms, depending on font, emphasis, etc.

I18n Overview: Presentation matters

Character glyph rendering





عندما يريد العالم أن يتكلم، فهو يتحدث بلغة يونيكود. تسجل الآن لحضور المؤتمر الدولي العاشر ليونيكود (Unicode Conference)، الذي سيعقد في 10-12 آذار 1997 بمدينة مانتس، ألمانيا. وسيجع المؤتمر بين خبراء من كافة قطاعات الصناعة على الشبكة العالمية إنترنت ويونيكود، حيث ستم، على الصعيدين الدولي والمحلي على حد سواء مناقشة سبل استخدام يونيكود في النظم القائمة وفيما يخص التطبيقات الحاسوبية، الخطوط، تصميم النصوص والحوسبة متعددة اللغات.

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 53

Arabic letters join up with adjacent letters, so there can be up to four different forms for each letter when displayed. You use the same character code, however, for a given letter of the alphabet, and the different glyph shapes are chosen from the font according to the position of the letter at run time.

I18n Overview: Presentation matters

Character glyph rendering

W3C®

ह + ि + न + ् + द + ी

हिन्दी

Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 54

This slide shows how characters can be visually represented in apparently different orders and with different shapes in Hindi, also.

Unicode is making a similar distinction between content and presentation as we saw earlier. This enables easier text analysis and comparison, given that the underlying codes represent semantic rather than presentational information.



I18n Overview: Presentation matters

Script-specific typography


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。	

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 55

CSS3 holds the promise of a number of typographic approaches that are needed for non-Latin scripts, such as Chinese and Japanese.

I18n Overview: Presentation matters

Script-specific typography



国際サッカー連盟（F
I F A）本部での理事
会を前に記念撮影にお
さまる理事ら。前列右
から2人目がヨハンソ
ン（Johansson） 欧州
サッカー連盟会長。

国際サッカー連盟（F
I F A）本部での理事
会を前に記念撮影にお
さまる理事ら。前列右
から2人目がヨハンソ
ン（Johansson） 欧州
サッカー連盟会長。

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 56

A mechanism for support of vertical text is being proposed. Discussions are currently centered on mixing text of different directionalities on the same line.

This and the following slide indicate some of the changes that are applied to text as its orientation is changed. Note that this should just be presentational sugar. There should be no need to alter the content, just the styling.

I18n Overview: Presentation matters
Script-specific typography

W3C®




Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 57

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.

Before that, if you use a non-Latin script, you should check that your requirements have been taken into account. 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.

I18n Overview: Presentation matters

Script-specific typography



Richard Ishida Consultant, Design of International User Interfaces

ريتشارد ايشيدا - مستشار, (456-123) تصميم عالمي لواجهات (global design) المستعمل

ריצ'רד אישידה - יועץ, (123-456) תכנון מישקי למשתמש 123-456 בין-לאומי

理查德·伊喜达、国际化软件界面设计顾问。国际化软件界面、设计顾问「国际化软件界面设计顾问」

石田リチャード ユーザーインターフェース「国際化対応設計」コンサルタント。

리차드 이시다 각국어 인터페이스 기술컨설턴트

ริชาร์ด อิชิดะ ริชาร์ด อิชิดะ ที่ปรึกษา งานออกแบบการใช้งานเครื่องใช้และระบบสากล (UI)

Ρίτσαρντ Ισιντα Εὐμβουλος, Εχεδιασμός διεθνῶν διασυνδέσεων χρήστη

Ричард Ишида Консультант по интернационализации интерфейса пользователя

रिचड ईशिदा परामशरदाता, डिज़इन ऑफ इंटरनेशनल यूजर इंटरफेज़


<http://people.w3.org/rishida/scripts/samples/wrapping.html>

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 58

This and the following slide illustrate how different scripts exhibit different wrapping behavior at the end of a line.

It is important to ensure that user agents perform such wrapping correctly. It is also important to ensure that all the user parameters that are needed to control wrapping are available to the styling mechanism (eg. CSS).



I18n Overview: Presentation matters

Script-specific typography

Richard Ishida Consultant, Design of International User Interfaces

ريشارد ايشيدا - مستشار, (456-123) تصميم عالمي لواجهات (global design) المستعمل

ריצ'רד אישידה - יועץ, (123-456) תכנון מישקי למשתמש -123
456 בין-לאומי

理查德·伊喜达、国际化软件界面设计顾问。国际化软件界面、设计顾问「国际化软件界面设计顾问」

石田リチャード ユーザーインターフェース「国際化対応設計」コンサルタント。

리차드 이시다 각국어 인터페이스 기술컨설턴트

ริชาร์ด อิชิดะริชาร์ด อิชิดะ ที่ปรึกษา งานออกแบบหน้าจอการใช้เครื่องระบบสากล (UI)

Ρίτσαρντ Ισιδα Εύμβουλος, Εχεδιασμός διεθνών διασυνδέσεων χρήστη


Ричард Ишида Консультант по интернационализации интерфейса пользователя

रिचड ईशिदा परामशरदाता, डिज़इन और इंटरनेशनल यूजर इंटरफ़ेज़

<http://people.w3.org/ishida/scripts/samples/wrapping.html>

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 59



I18n Overview: Presentation matters
Script-specific typography

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

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 60

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

I18n Overview: Presentation matters

Script-specific typography



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

I18n Overview: Presentation matters

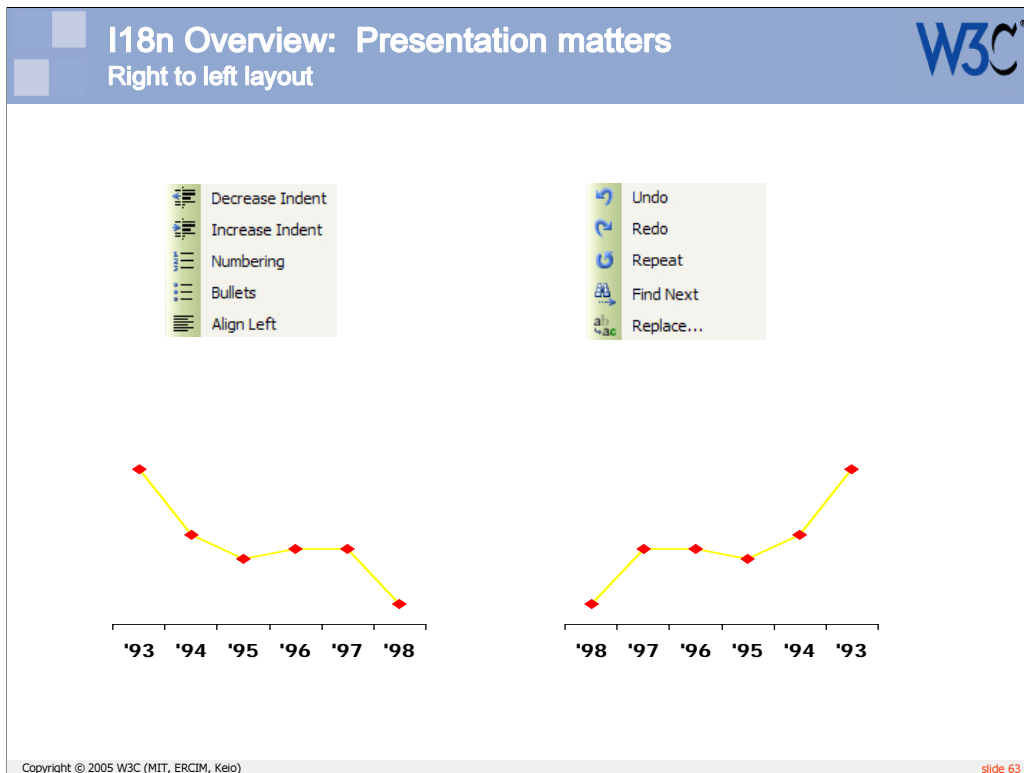
Right to left layout







Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 62

Because the predominant reading order for Arabic and Hebrew script-based text is right-to-left, the layout of pages also tends to be a mirror-image of a page in a European language.



Icons may also have directional bias. In some cases the icon can simply be flipped horizontally, in others a new icon will need to be created.

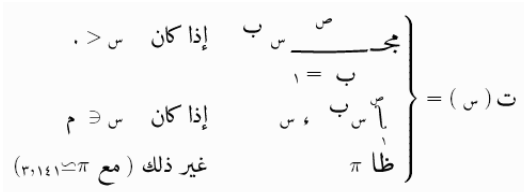
Tables, spreadsheets, charts, collated pictures, etc. also tend to flow from right to left where a right to left script is used.



I18n Overview: Presentation matters

MathML

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



Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 64

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.

This section on presentation invites you to:

- find out and use features that are currently available
- design your applications in an extensible way, so that these features can be incorporated when needed for international content
- push for new features to be implemented by user agents – getting support in the W3C standards is not sufficient, the user agent developers must also be convinced that they should support them – this means both pushing for feature to be supported, and using them when they are made available.



I18n Overview: Presentation matters
Speech synthesis


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.


Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 65

A recent workshop in Beijing explored international requirements for markup to support speech synthesis. There are plans to organize another workshop soon.

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 results of the Beijing workshop.




Overview



W3C's I18n Activity
L10n or i18n?
Content vs. presentation
[I18n overview](#)
 Characters
 Document formats
 Presentation matters
 [Practical barriers](#)
 Cultural differences
Summary

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 66



I18n Overview: Practical barriers

Text fragmentation & re-use

Tokyo hotel:

- ◆ Cooles and heates. If you want the condition of cool in your room, please control yourself.

Paris hotel:

- ◆ Guests are expected to complain at the office between the hours of 9am and 10am each day.

Tokyo hotel:

- ◆ You are invited to take advantage of the chambermaid.

Vienna hotel:

- ◆ In case of fire, do your utmost to alarm the porter.

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 67

This slide helps make the point that translation is not simply word substitution, but that there can be complex and subtle factors to take into account to achieve a good rendering of the meaning.



I18n Overview: Practical barriers
Text fragmentation & re-use

They 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

Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 68

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.




I18n Overview: Practical barriers


Text fragmentation & re-use

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)
```

Copyright © 2005 W3C (MIT, ERCIM, Keio)

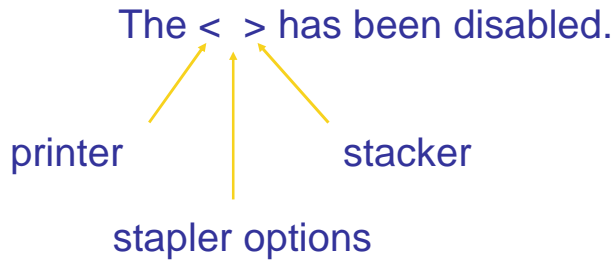
slide 69

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 advance techniques in PHP, you will prevent this possibility and seriously affect translatability.



I18n Overview: Practical barriers
Text fragmentation & re-use



The < > has been disabled.

printer stacker

 stapler options

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 70

In this example, the developer has tried to save memory by re-using part of a common sentence. Unfortunately, because of the effects of rules about agreement between gender and number in many languages, this becomes an untranslatable phrase. The developer needs to be aware of the likely impact on translatability of such things.

I18n Overview: Practical barriers

Screen usage

W3C®

Interface Language

Sprache der Benutzeroberfläche

Interface Language


Sprache der Benutzeroberfläche

Copyright © 2005 W3C (MIT, ERCIM, Keio)


slide 71

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.



Overview



W3C's I18n Activity
L10n or i18n?
Content vs. presentation
[I18n overview](#)
 Characters
 Document formats
 Presentation matters
 Practical barriers
 [Cultural differences](#)
Summary

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 72

I18n Overview: Cultural differences

Data formats

W3C®

Россия
г. Пермь 614055
ул. Крупской 93-82
Селивановой Юлии

Country: United States

First name:

Last name:

Address:

City:

State: AZ

Zip code:

Telephone: ()

Application date:

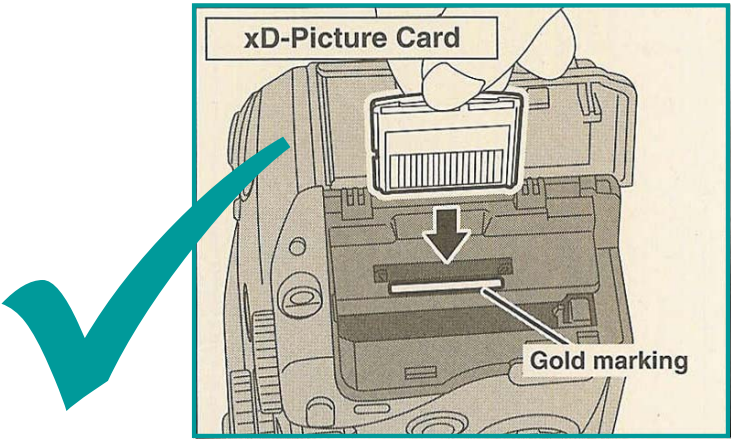
Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 73

Be careful about assuming what others' name and address formats will be. Also think about how you will store the names and addresses in the database. For example, do you really need to split out street number? How will you generate a Russian or Japanese address that goes from general to specific from top to bottom?

I18n Overview: Cultural differences
Symbolism, color, graphics...

W3C®



The diagram shows a top-down view of an xD-Picture Card being inserted into a camera's memory slot. A large red checkmark is superimposed over the card, indicating it is the correct orientation. A label 'xD-Picture Card' points to the card itself. Another label 'Gold marking' points to the gold contacts on the card. A black arrow points down into the slot, indicating the insertion direction.

Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 74

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.

I18n Overview: Cultural differences

Symbolism, color, graphics...



Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 75

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

I18n Overview: Cultural differences
Symbolism, color, graphics...


W3C®



Copyright © 2005 W3C (MIT, ERCIM, Keio)


slide 76

Graphics may need to be changed if they don't reflect the local culture of certain places.



I18n Overview: Cultural differences

Symbolism, color, graphics...



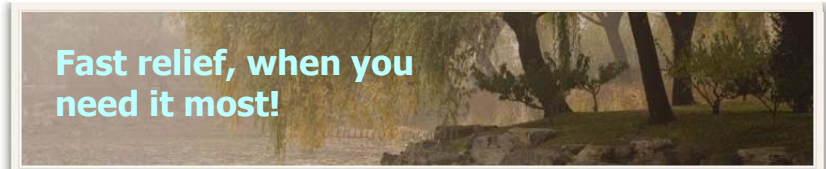
Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 77

Body language and gestures are particularly dangerous. Each of these symbols can give offense in one part of the world or another.

I18n Overview: Cultural differences

Symbolism, color, graphics...

W3C®



**Fast relief, when you
need it most!**

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 78

When dealing with graphics, consider how to deal with text. Ideally the text will be overlaid on a graphic, rather than embedded in it. If the text is within the graphic, try to ensure that you develop it in layers, with text on a separate layer, so that when it comes to translation the text can be easily removed and replaced over complicated backgrounds.

I18n Overview: Cultural differences

Symbolism, color, graphics...



Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 79

Be wary of humor. It doesn't travel well.

Also, try to ensure that examples used in text are understandable by the audience of the translated version. Use examples from a variety of cultures where possible.

I18n Overview: Cultural differences

Symbolism, color, graphics...





Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 80

Color also has different connotations in different parts of the world.

I18n Overview: Cultural differences

Symbolism, color, graphics...

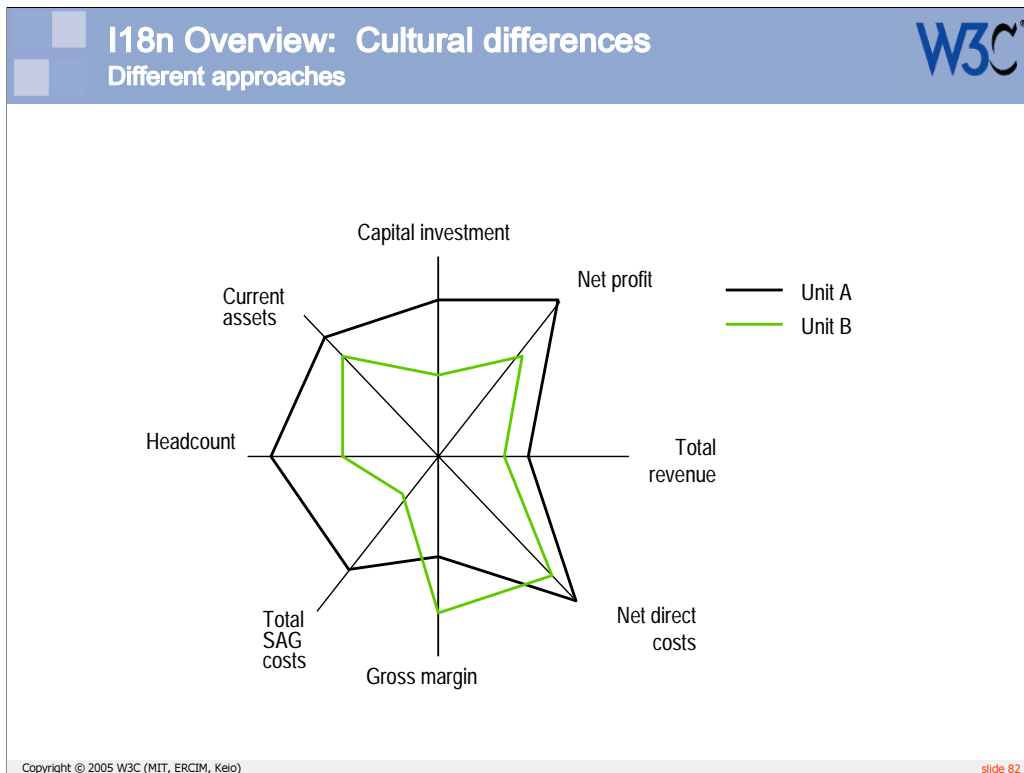





Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 81

For example, a black wedding kimono is not as strange in Japan as it may seem to a European.



Then you need to be aware that people in different parts of the world may do things in different ways. For example, the radar chart was such a common way of representing comparative data in Japan that, when Lotus 1-2-3 was launched in that area they had to reengineer it to add that.



I18n Overview: Cultural differences

Different approaches

"... one Latin American teacher recently complained to me that the US-manufactured and well-translated educational software currently being used in his country's primary schools presupposed 'solitary problem solvers', whereas his culture stressed collective problem-solving."

Kenneth Keniston,
Language International, May 1996

Copyright © 2005 W3C (MIT, ERCIM, Keio)

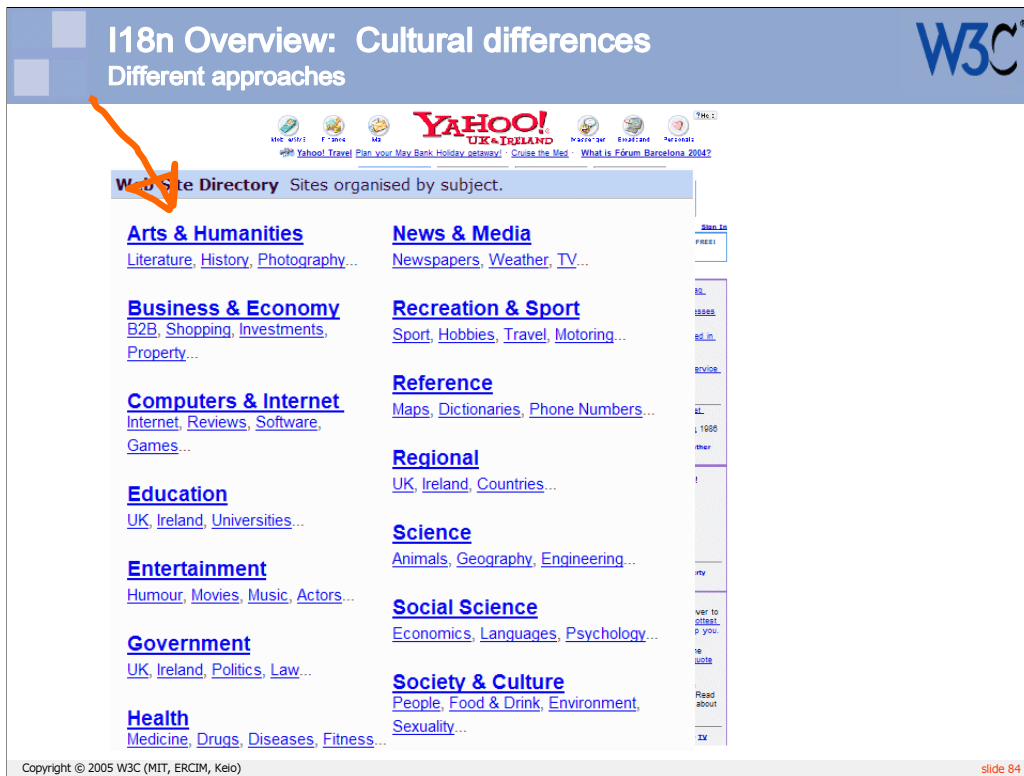
slide 83

Considerations of this kind require you to make big decisions at the very start of the development phase about how to proceed. Otherwise you could waste a lot of time and energy producing something that doesn't meet your customer's needs.

W3C®

I18n Overview: Cultural differences

Different approaches




The screenshot shows the Yahoo! UK & Ireland homepage. At the top, there's a navigation bar with links like 'Yahoo! Travel', 'Plan your Mav Bank Holiday getaway', 'Cruise the Med', and 'What is Fórum Barcelona 2004?'. Below this is the 'Web Site Directory' section, which is titled 'Sites organised by subject.' and lists various categories such as Arts & Humanities, Business & Economy, Computers & Internet, Education, Entertainment, Government, Health, News & Media, Recreation & Sport, Reference, Regional, Science, Social Science, and Society & Culture. An orange arrow points to the 'Web Site Directory' header.

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 84

This and the following slides show how Yahoo adapts its categorizations to reflect the preoccupations of various different countries.

I18n Overview: Cultural differences

Different approaches



Guide Web - Classement thématique de sites Web

[Suggérer un site](#)

<p><u>Actualités et médias</u> Journaux, Télévision, Météo...</p> <p><u>Commerce et économie</u> B2B, Shopping, Emploi, Immobilier...</p> <p><u>Informatique et Internet</u> Internet, Logiciels, Fonds d'écran...</p> <p><u>Santé</u> Diététique, Médecine, Thalasso...</p> <p><u>Enseignement et formation</u> Primaire, Secondaire, Supérieur...</p> <p><u>Institutions et politique</u> Ministères, Droit, Politique...</p> <p><u>Sciences et technologies</u> Animaux, Astronomie, Physique...</p>	<p><u>Sports et loisirs</u> Foot, Tourisme, Auto/Moto, Jeux...</p> <p><u>Art et culture</u> Littérature, Cinéma, Musique, BD...</p> <p><u>Divertissement</u> Tests/Quiz, Loteries, Humour, Sorties...</p> <p><u>Classement géographique</u> Pays, Europe, France, Paris...</p> <p><u>Références et annuaires</u> Dictionnaires, Annuaires, Cartes/Atlas...</p> <p><u>Société</u> Enfants, Gastronomie, Rencontres...</p> <p><u>Sciences humaines</u> Archéologie, Histoire, Psychologie...</p>
---	--

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 85

W3C®

I18n Overview: Cultural differences

Different approaches




The screenshot shows the Yahoo! Japan homepage with a grid of category links. An orange arrow points to the 'Arts and Culture' (芸術と人文) category, which includes links for 'Photography, Architecture, Museums, History, Literature' (写真, 建築, 美術館, 歴史, 文学).


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 ...

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 86



Overview




W3C's I18n Activity
L10n or i18n?
Content vs. presentation
I18n overview

- Characters
- Document formats
- Presentation matters
- Practical barriers
- Cultural differences

Summary


Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 87



Summary


The value of internationalization



Localization can help bridge the Digital Divide


Localization can be a lot more than just translation

Treating localization as an afterthought only makes things worse



- Higher cost during localization due to rework and support activities
- Longer localization schedules, and slower time to market/release
- Reduced quality user perceptions, despite the extra effort

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 88



Summary


The value of internationalization

Internationalization means:

- using a Quality approach to reduce the overall cost and time to market/release of multinational deliverables
- **designing** into the product an internationalized base, and a modular and easily adaptable architecture
- not always doing extra work – maybe just working in a better way

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 89



Summary

Making a difference

As content developers:

- ensure that the needs of your scripts and languages are brought to the attention of the W3C and other standards and tool developers
- push for adoption of international technology by user agents **and content creation/management tools**
- keep presentation and content separate
- always use standards and validate your code
- follow best practices for internationalization, accessibility, the mobile web, etc.

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 90

Summary

GEO resources



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

On this site

- About the Activity
- Working Groups
- Topic index
- Techniques index
- Resource types
- News filters & RSS feeds
- Translations

Shortcuts

Groups: Core, GEO, ITS

Contact

Specifications

අනුපිළිවෙල (Translations)

New translations

- 12 Aug 05 FAQ: Tipos em nome de grupo (Brazilian Portuguese)
- 12 Aug 05 FAQ: Vérifier les en-têtes HTTP (French)
- 16 Aug 05 FAQ: Quem usa Unicode? translated into Brazilian Portuguese
- 16 Aug 05 FAQ: Verificando e corrigindo HTTP translated into Brazilian Portuguese

W3C Internationalization (I18N) Activity

Making the World Wide Web truly World Wide!

The W3C Internationalization Activity has the goal of proposing and coordinating any techniques, conventions, guidelines and activities within the W3C and together with other organizations that allow and make it easy to use W3C technology worldwide, with different languages, scripts, and cultures.

[More](#)

Announcements

- 23 Aug 05 FAQ: Upgrading from language-specific legacy encoding to Unicode encoding **available for review**
- 22 Aug 05 New translations see bottom left of this page
- 17 Aug 05 FAQ: *xml:lang* in XML document schemas **available for review**
- 15 Aug 05 Tutorial: Creating *Bidi XHTML/HTML Pages* **available for review**

Web Services Internationalization (WS-I18N)

New First Working Draft 15 September 2005
 The Internationalization Core Working Group has released the First Public Working Draft of Web Services Internationalization (WS-I18N).
 The draft enhances SOAP messaging for locale and international preference negotiation and defines a locale policy. Without using Accept-Language and user identity, implementations can handle the requester's locale, locale policy and language preference.
 Editor: Addison Phillips, Mary Trumble

Changing (X)HTML page encoding to UTF-8 26 August 2005
New FAQ-based article
 How do I change the encoding of my (X)HTML pages to UTF-8?
 Author: Richard Ishida

W3C Participates in 28th Internationalization & Unicode Conference 23 August 2005
News Item
 The 28th Internationalization & Unicode Conference will be held 7-9 September in Orlando, Florida, USA. Team members Richard Ishida and Felix Sasaki will present several papers at this premier technical conference worldwide for software and Web internationalization. Read about Unicode and the W3C Internationalization Activity.

More news:
[More announcements \[RSS\]](#) • [Resources \[RSS\]](#) • [Reviews \[RSS\]](#) • [Changes \[RSS\]](#) • [Translations \[RSS\]](#) • [Tests \[RSS\]](#) • [General \[RSS\]](#)

Participate!


- Join a Working Group
- Review a W3C specification
- Translate a specification or page
- Subscribe to the Interest Group list

Topics

- Characters
- Fonts
- Language
- Text formatting
- Bidirectional text
- Resource identifiers
- Locale specific data
- Forms
- Structure & metadata
- Display issues
- Miscellaneous

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 91

All the GEO materials are available from the Internationalization home page.



Summary GEO resources

- General**
 - International & multilingual Web sites
 - Monolingual vs. multilingual Web sites
 - Serving XHTML 1.0
- Characters & encoding**
 - Character encodings
 - Document character set
 - Who uses Unicode?
 - Upgrading to Unicode
 - An Introduction to Multilingual Web Addresses
 - Language tags in HTML and XML
 - The HTTP Charset parameter
 - Checking HTTP headers
 - Checking the character encoding using the validator
 - CSS character encoding declarations
 - Multilingual form encoding
 - HTML, XHTML, XML and control codes
 - Unexpected characters or blank lines
 - Setting encoding in web authoring applications
 - Setting 'charset' information in .htaccess
 - Missing characters and glyphs
 - Non-English tags
- Language**
 - Why use the language attribute?
 - Using HTTP and meta for language information
 - Language tags in HTML and XML
 - 2-letter or 3-letter language codes
 - Styling using the lang attribute
 - Setting language preferences in a browser
- When to use language negotiation**
 - Apache MultiViews language negotiation set up
 - Accept-Language used for locale setting
- Bidirectional text**
 - What you need to know about the bidi algorithm and inline markup
 - Bidi space loss
 - XHTML/HTML & bidi formatting codes vs. markup
 - CSS vs. markup for bidi support
- Resource identifiers**
 - An Introduction to Multilingual Web Addresses
- Local formats**
 - Displaying formatted dates
 - Select box sorting
 - Dates and time
- Miscellaneous**
 - Display capabilities
 - Script direction & languages
 - Introduction to Indic Scripts
 - Ruby
- Navigation**
 - Global gateway pull-downs
- Other**
 - XSLT for multilingual output
 - Using the link element

Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 92

There are many articles available on the W3C Internationalization subsite (<http://www.w3.org/International/>). This shows a selection.

Summary GEO resources

Authoring Techniques for XHTML & HTML Internationalization: Characters and Encodings 1.0

W3C Working Draft 9 May 2004

This version: <http://www.w3.org/TR/2004WD/xhtml-tech-20040509/>

Latest version: <http://www.w3.org/TR/2004WD/xhtml-tech-20040509/>

Previous version: <http://www.w3.org/TR/2003WD/xhtml-tech-20031009/>

Editor: Richard Ishida, W3C

[Copyright © 2004 W3C](#) [MIT](#) [ERCIM](#) [Keio](#) All Rights Reserved. W3C, MIT, ERCIM, Keio, and authors hereby acknowledge the W3C.

Abstract

It is important to consider character encoding matters when authoring content for the Web. This document is one of a series of documents providing internationalized HTML, using XHTML 1.0 or HTML 4.01. It focuses specifically on advice about character matters. It is produced by the Guidelines, Education & Outreach Task Force (GEO) of the W3C Internationalization Working Group (IWG).

Authoring Techniques for XHTML & HTML Internationalization: Specifying the language of content 1.0

W3C Working Draft 9 May 2004

This version: <http://www.w3.org/TR/2004WD/xhtml-tech-20040509/>

Latest version: <http://www.w3.org/TR/2004WD/xhtml-tech-20040509/>

Previous version: <http://www.w3.org/TR/2003WD/xhtml-tech-20031009/>

Editor: Richard Ishida, W3C

[Copyright © 2004 W3C](#) [MIT](#) [ERCIM](#) [Keio](#) All Rights Reserved. W3C, MIT, ERCIM, Keio, and authors hereby acknowledge the W3C.

Abstract

Specifying the language of content is useful for a wide number of applications, from linguistically sensitive searching to applying language-specific display properties. In some cases the full application is still awaiting full development, whereas in others, such as detection of language by voice browsers, it is a necessity today. Marking up language meta-information is something that can and should be done today. Without it, none of these applications can be taken advantage of.

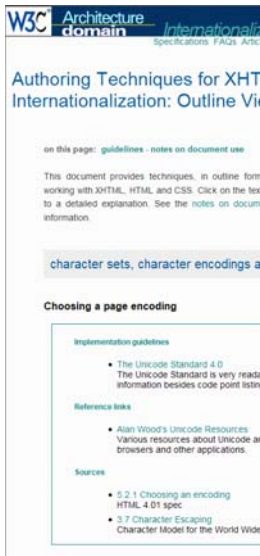
This document is one of a series of documents providing HTML authors with techniques for developing internationalized HTML, using XHTML 1.0 or HTML 4.01, supported by CSS1, CSS2 and some aspects of CSS3. It focuses specifically on advice about specifying the language of content. It is produced by the Guidelines, Education & Outreach Task Force (GEO) of the W3C Internationalization Working Group (IWG).

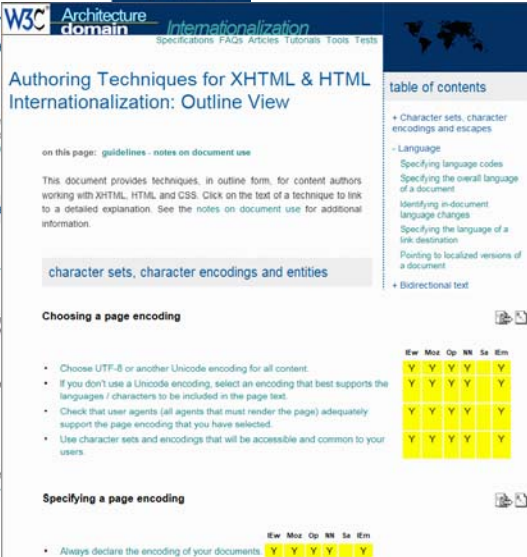
Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 93

In addition, the GEO Working Group is developing techniques documents, to provide advice on accomplishing particular tasks.


Summary
GEO resources
W3C®






Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 94

We are trying to provide multiple points of entry and levels of detail to the techniques information.



Summary


Making a difference



As designers and developers:


- be aware that well localized deliverables and content begins with you, not the localization folks
- develop an awareness of good internationalization practice
- put in place a mechanism for obtaining user feedback on an international level
- don't be locked into a short-term view with development costs – realize the value of enabling expansion beyond the local market/users
- always use standards (especially Unicode)

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 95



Summary

Making a difference

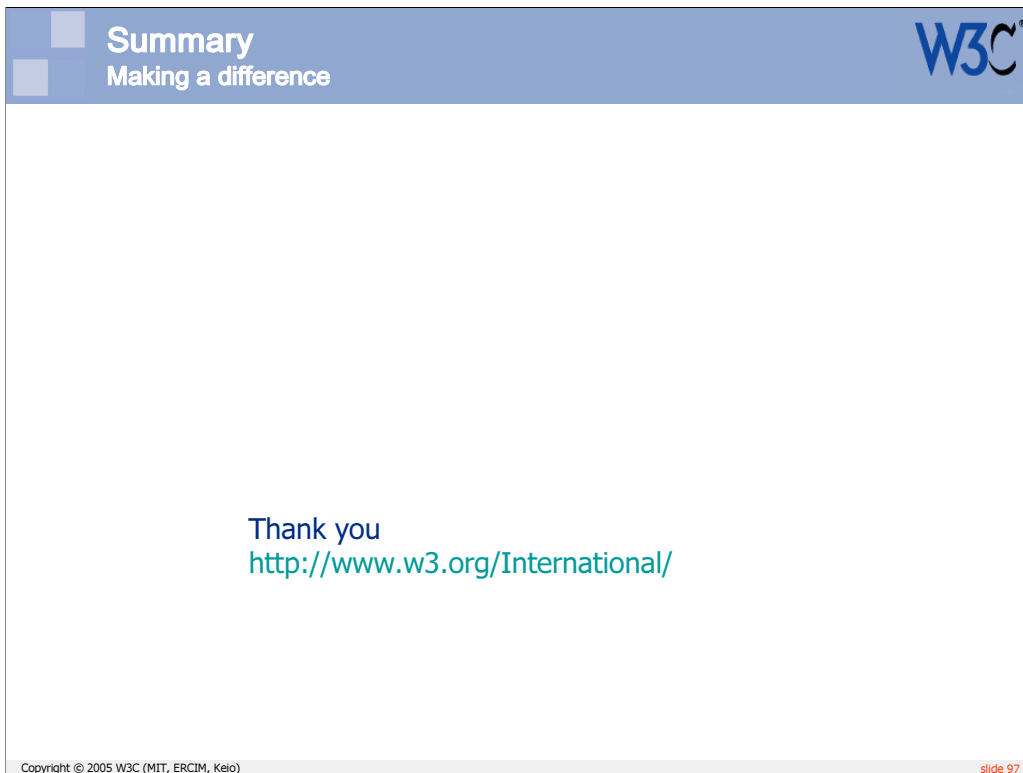


Get involved:

- visit the I18n Activity Home Page
- join a W3C Internationalization Working Group, or the Interest Group (www-international@w3.org)
- offer to help with reviews, or provide local knowledge for other Working Groups
- provide translations of W3C specifications or articles
- take advantage of the i18n-readiness of W3C technology

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 96



Summary
Making a difference

W3C®

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

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 97

This slide is a summary slide with a blue header bar containing the text 'Summary' and 'Making a difference', and the W3C logo. The main content area is white and contains the text 'Thank you' and a link to 'http://www.w3.org/International/'. The footer is a light gray bar with copyright information and the slide number 'slide 97'.

There a number of well-accepted goals in development these days, expressed by words ending in -bility – interoperability, scalability, portability, accessibility. In my mind, in today's ever increasingly global community, another key -bility is localizability.