



## 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 Groups



### 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

### Interest Group

[www-international@w3.org](mailto:www-international@w3.org)

Copyright © 2005 W3C (MIT, ERCIM, Keio)



slide 4



## W3C Internationalization Activity Objectives

- 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

Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 5

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 6

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 7

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 8



Separating content & presentation
W3C®

## Content ( XHTML )

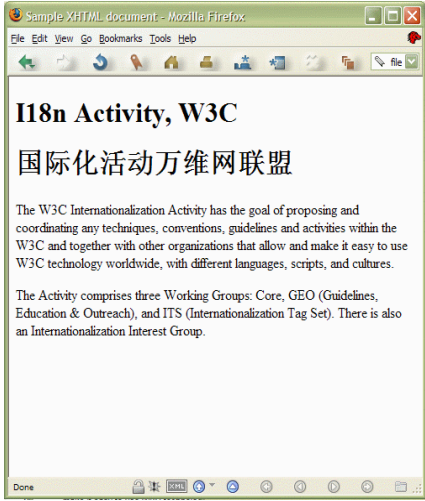
```

<!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>


```



Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 9

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


### 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</h1>
<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 10

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®

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

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.

Separating content & presentation

W3C®



The image shows a mobile phone screen displaying a web page. The page has a title 'World Wide Web...' and a subtitle 'I18n Activity, W3C'. The main content is a paragraph about the W3C Internationalization Activity. The text is enclosed in a rectangular box, demonstrating how content is separated from presentation. Below the text are 'OK' and 'Options' buttons. The phone has a standard keypad and navigation buttons.

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 12

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.



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 13

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 14




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 15


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.



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 16

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 <span class="bold">System Administrator Guide</span>`  
`for an example of re-use.</p>`


Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 17

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 <span class="doctitle">System Administrator Guide</span>  
for an example of re-use.</p>`


doctitle  
chaptertitle  
inputsequence  
etc.



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




# 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 19



# 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 20

I18n Overview: Characters

Character sets & encodings

W3C®

جعل شبكة الويب العالمية عالمية حقًا !  
缔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  
ליצור מהשהתורה כלל עולמית באמת  
वर्ल्ड वाईड वेब को सचमुच विश्वव्यापी बना रहे हों !  
ČeŕL ΔΡϙϖΡδϖ ρϙ ϙϙϙϙϙ ϙϙϙϙϙ.  
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!  
ਵਰਡ 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 21

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 22

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	ø
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	Ô	ä	ô	
5		%	5	E	U	e	u			¥	µ	Å	Ö	å	ö	
6		&	6	F	V	f	v			¦	¶	Æ	Ø	æ	ø	
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 23

Adding an 8<sup>th</sup> 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			~	±	A	P	α	ρ	
2		"	2	B	R	b	r			'	²	B	□	β	ς	
3		#	3	C	S	e	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			ı	ı	Z	Φ	ξ	φ	
7		'	7	G	W	g	w			§	'	H	X	η	χ	
8		(	8	H	X	h	x			"	'	E	Θ	Ψ	θ	ψ
9		)	9	I	Y	i	y			©	'	H	I	Ω	ι	ω
A		*	:	J	Z	j	z			□	'	I	K	İ	ı	ı
B		+	;	K	[	k	{			«	»	Λ	Ÿ	λ	ü	
C		,	<	L	\	l				¬	'	O	M	ó	μ	ó
D		-	=	M	]	m	}			-	½	N	é	v	ú	
E		.	>	N	^	n	~			®	'	Y	Ξ	ή	ξ	ó
F		/	?	O	_	o				-	Ω	O	ı	o	□	

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 24


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 25


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.

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

## 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
- Tifinagh
- Etc....



Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 26

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, 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 specifications to make them support XML 1.1.

W3C®

## I18n Overview: Characters

### Character sets & encodings

A	κ	好	丕
Code point			
41	5D0	597D	233B4

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

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				
Character sets & encodings				
	A	κ	好	丕
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 28

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 29

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>).

The screenshot shows a Flickr photo page. The title of the photo is "Château de La Napoule", where the character "t" is replaced by a large, illegible symbol. The page includes a navigation bar with links like Home, Tags, Groups, People, and Invite. Below the title, there is a large photo of a stone sculpture. To the right, there is a sidebar with "r12a's photostream" showing 1177 photos, a list of tags (0506-cote-dazur, Mandelieu-La Napoule, Alpes-Maritimes, France), and additional information about the photo's metadata.

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.


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 31	

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





Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 32

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

W3C®


axあaxあ

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

↑


Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 33



# 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 34

W3C®

## I18n Overview: Characters

Working with characters

NFC

NFD


Ízelítőül

Íőzelíőtoőuől

Ha a világ beszélni akarna, Unicode-ul szólalna meg.  
 Regisztráljon már most a Tizedik Nemzetközi Unicode  
 Konferenciára, melyet 1997. március 10-12-én rendeznek  
 Meinz-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 35

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 36

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-5wao1o.josefsson.org/m%C3%A5l/franz%C3%A9n.html

- Phishing ([www.paypal.com](http://www.paypal.com))


Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 37

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 38

W3C<sup>®</sup>

## 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 39

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/> .


Richard Ishida

39

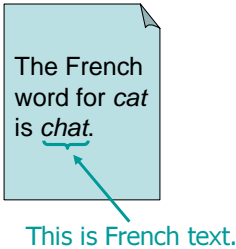

Version: 10 june 2003

## 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 40

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.


For more information see <http://www.w3.org/TR/i18n-html-tech-lang/#ri20040808.100519373>



<div> <div></div> <div>I18n Overview: Document formats</div> <div>Declaring the language of text</div> </div> <div>W3C®</div>		
<div>RFC 3066</div> <div>zh-TW ?</div> <div>zh-HK ?</div>	<div>中國語</div>	<div>RFC 3066 replacement</div> <div>zh-Hant</div> <div>zh-Hant-HK</div> <div>zh-cmn-Hant</div> <div>zh-cmn-Hant-HK</div> <div>etc.</div>
<small>Copyright © 2005 W3C (MIT, ERCIM, Keio)</small>		<small>slide 41</small>

The current way of expressing language in values for `xml:lang` and other places is to follow the rules of the IETF's RFC 3066 specification. There is a problem for Chinese, since RFC 3066 didn't allow you to label Simplified or Traditional Chinese independently of the dialect until recently. Many people used `zh-TW` for Traditional Chinese, whereas others used `zh-HK`.

A replacement for RFC 3066 has been approved by the IETF and is awaiting publication. (Members of the W3C I18n Activity have been involved in its development.) The new specification will provide a lot more power for handling language declarations. For example, in Chinese it will be possible to use the code listed above right to mean, respectively, Traditional Chinese, Traditional Chinese as used in Hong Kong, Mandarin Chinese written in Traditional Chinese, Mandarin Chinese as written in Traditional Chinese in Hong Kong, etc.



## I18n Overview: Document formats

### Locale information

**WS-i18n**

Enhancements to SOAP messaging to provide internationalized and localized operation via locale and international preference negotiation, and a general-purpose mechanism for associating a "locale policy" with messages.

**LTLI**

How document formats, specifications, and implementations should implement language and locale identifiers, as well as data structures for describing international preferences.

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

The W3C Internationalization Activity is also working on documents aimed at improving handling of language and locale information in specifications such as those relating to Web Services.

I18n Overview: Document formats
W3C®Script-specific markup

Characters as ordered in memory:

The title says "<span>פועליות הבינאום, W3C</span>" in Hebrew.



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

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 43


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.

I18n Overview: Document formats  
Script-specific markup
W3C<sup>®</sup>

Characters as ordered in memory:  
The title says "<span>פועילות הבינאום, W3C</span>" 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 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.

I18n Overview: Document formats  
Script-specific markup
W3C®

Characters as ordered in memory:  
 The title says "<span dir='rtl'>פועילות הבינאום, W3C</span>" in Hebrew.



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



Using the bidi algorithm only  
 The title says "פועילות 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.

I18n Overview: Document formats  
Markup to support localization
W3C

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 46

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



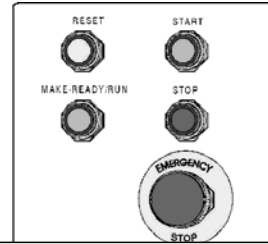
Von der VISTA-Konsole aus einen Druckauftrag übermitteln. (Siehe hierzu "Auftrag übergeben" in Kapitel 5.)

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 48

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<sup>®</sup>

## 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 49

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.



I18n Overview: Document formats  
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 50

A recent workshop in Beijing explored international requirements for markup to support speech synthesis. There are plans to organize another workshop in Crete at the end of May 2006.

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 51



# I18n Overview: Presentation matters

## Character glyph rendering



Character

vs.

Glyph

*a* *a*

雪 雪

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 52

Unicode also separates semantics from presentation. There is usually a single code point for any character. The visual representation of that character (it's glyph) however is font dependent.

**I18n Overview: Presentation matters**  
Character glyph rendering

W3C®

عندما يريد العالم أن يتكلم، فهو يتحدث بلغة يونيكود. تسجل الآن لحضور المؤتمر الدولي العاشر ليونيكود (Unicode Conference)، الذي سيعقد في 12-10 آذار 1997 بمدينة مانتس، ألمانيا. وسيجع المؤتمر بين خبراء من كافة قطاعات الصناعة على الشبكة العالمية انترنت ويونيكود، حيث سيتم، 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

In some scripts, the font glyph differences do not merely reflect style preferences. Most Arabic characters can have up to four different shapes, depending on the visual context. This is because of the joined up nature of Arabic writing. Each letter of the alphabet, however, has a single code point in Unicode, and rendering rules in the operating system and / or font are used to pick the appropriate glyph from the font at run time.

I18n Overview: Presentation matters

Character glyph rendering

W3C®

ह + ि + न + ् + द + ी

हिन्दी

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 54

These rendering rules not only affect glyph shaping, but may do more complicated things like reordering the visual placement of characters, since characters are usually stored in a 'logical' order in memory that reflects the way they are typed or spoken. The example above shows how Devanagari text (Hindi) puts all combining characters after base characters (a cardinal rule in Unicode text storage), but displays some characters to the left of the base character when printing or displaying on screen.

I18n Overview: Presentation matters

Script-specific typography


W3C®

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. Here are just a few examples.



## I18n Overview: Presentation matters

### Script-specific typography

当世界需要沟通时，请用  
Unicode。将于3月10日-12  
日在德国 Mainz 市举行的  
第十届统一码国际研讨会现  
在开始注册。本次会议将汇  
集各方面的专家。涉及的领  
域包括：国际互联网和统一  
码，国际化和本地化，统一  
码在操作系统和应用软件中  
的实现，字型，文本格式以  
及多文种计算等。

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 56


This is vertical Chinese text. Note that Latin text flows down the lines, but also that the numbers are arranged horizontally within the vertical flow.

You start reading the text at the top right, and progress towards the left of the page.



[illegible]

This is Mongolian. It is read vertically also, but you start at the top left, and progress towards the right. The question is, how do you handle a mixture of vertical Chinese and Mongolian text? The CSS Working Group is currently studying how to enable such mixtures.



## I18n Overview: Presentation matters

Script-specific typography

כאשר העולם רוצה לדבר, הוא מדבר ב־Unicode.  
הירשמו כעת לכנס Unicode הבינלאומי העשירי,  
שייערך בין התאריכים 10-12 במרץ, ב־מִיִּיִּץ  
שבגרמניה. בכנס ישתתפו מומחים מכל ענפי  
התעשייה בנושא האינטרנט העולמי וה־Unicode,  
בהתאמה לשוק הבינלאומי והמקומי, ביישום  
Unicode במערכות הפעלה וביישומים, בגופנים,  
בפריסת טקסט ובמחשוב רב־לשוני.

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 58

In addition, one has to integrate left-to-right and right-to-left text into vertical text. Again, the CSS Working Group is currently trying to finalize how to manage the combination of all these different script directions.

Note that this should just be presentational sugar. There should be no need to alter the content, just the styling, to move from a vertical to a horizontal display of text, and vice versa.

I18n Overview: Presentation matters

Script-specific typography

W3C®

Copyright © 2005 W3C (MIT, ERCIM, Keio)


slide 59

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 בישראלומי

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 ที่ปรึกษางานออกแบบหน้าจอการใช้เครื่องระบบสากล (UI)

Ρίτσαρντ Ισιντα Εύμβουλος, Εχεδιασμός διεθνών διασυνδέσεων χρήστη

Ричард Ишида Консультант по интернационализации интерфейса пользователя

रिचर्ड ईशिदा परामशरदाता, डिजिटल औफ इंटरनेशनल यूजर इंटरफेज

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


Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 60

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

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

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

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 งานออกแบบรายการใช้เครื่องระบบสากล (UI)

Ρίτσαρντ Ισιντα Εὐμβουλός, Εχεδιασμός διεθνῶν διασυνδέσεων χρήστη


Ричард Ишида Консультант по интернационализации интерфейса пользователя

रिचड ईशिदा परामशरदाता, डिजइन औफ इंटरनेशनल यूजर इंटरफेज

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

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 61



I18n Overview: Presentation matters  
Script-specific typography

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

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 62

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

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  
مسرقيات، ثم يومئ اذ تراود

**118n Overview: Presentation matters**  
Right to left layout

**W3C®**

Copyright © 2005 W3C (MIT, ERCIM, Keio)


slide 64






Directionality can also affect layout. Note, for example, how the column order is reversed in the Arabic page.








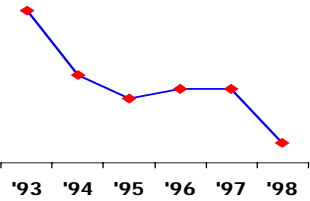
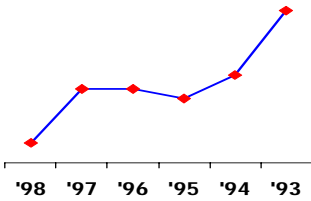
## I18n Overview: Presentation matters

Right to left layout



-  Decrease Indent
-  Increase Indent
-  Numbering
-  Bullets
-  Align Left

-  Undo
-  Redo
-  Repeat
-  Find Next
-  Replace...


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

Text direction also affects icons and graphics. The icons shown on this slide may need to be mirror imaged or, in some cases, redrawn for use with Arabic or Hebrew content.

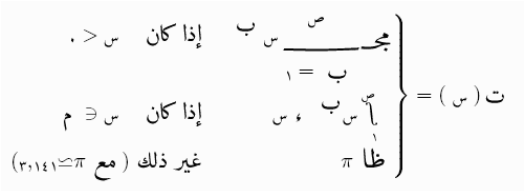
Also tables, collated pictures, graphs, spreadsheets, etc. commonly flow from right to left.

**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 66

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

:first-letter feedback request



One ought to know whether first letter styling has special implications for languages in non-Latin scripts.




Copyright © 2005 W3C (MIT, ERCIM, Keio)


slide 67

The W3C I18n Activity has begun an experiment to seek input regarding international requirements by posting a summary of a particular area on our web site.

Here is our first such page. It relates to the use of :first-letter in non-Latin scripts or Latin scripts with accents (see [http://www.w3.org/blog/International/2006/01/20/request\\_for\\_feedback\\_usefulness\\_of\\_first](http://www.w3.org/blog/International/2006/01/20/request_for_feedback_usefulness_of_first) )



# 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 68



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 69

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.

■
■
W3C®

## 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 70

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

W3C®

The < > has been disabled.

printer                      stacker

                         stapler options

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 71

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.

W3C®I18n Overview: Practical barriers
Screen usage

Interface Language

Sprache der Benutzeroberfläche

Interface Language


Sprache der Benutzeroberfläche

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 72


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 73

■ I18n Overview: Cultural differences  
■ Data formats
W3C®

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

Country:

First name:

Last name:

Address:

City:

State:

Zip code:

Telephone: (  )

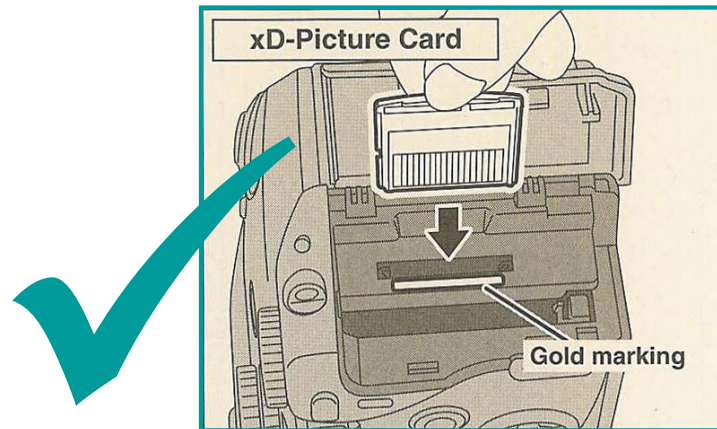
Application date:

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

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...



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.

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 76

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

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

W3C®




Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 77

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

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

W3C®



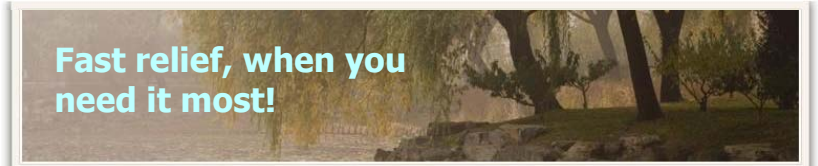
Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 78

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®



Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 79

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 80

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



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

W3C®



Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 81

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

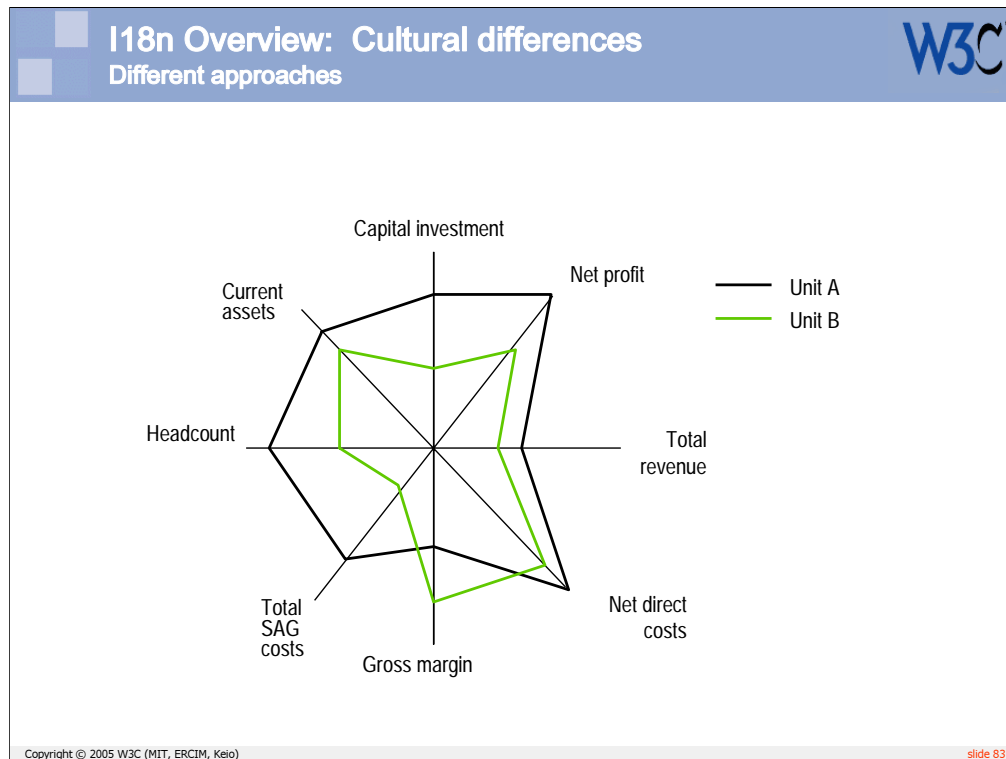
I18n Overview: Cultural differences  
Symbolism, color, graphics...

W3C®




Copyright © 2005 W3C (MIT, ERCIM, Keio)slide 82

It is unusual for women to wear black at a wedding in the West.



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 84

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

**Web Site Directory** Sites organised by subject.

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

Sign to  
FREE!

ES

ISSUES

ad in

archive

EL

1,1000

other

0

city

over to  
glossary  
p you

18  
last

Read  
about

TV

Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 85

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

The subcategories chosen for Arts & Humanities for the UK & Northern Ireland home page are Literature, History and Photography.

W3C®

## I18n Overview: Cultural differences

### Different approaches

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

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

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 86

Subcategories for this same subsection in French list Literature, Cinema, Music and Graphic Novels.

Yahoo is not only translating, but also adapting content for the different market places.

**118n Overview: Cultural differences**  
Different approaches



W3C®

Yahoo! カテゴリ サイトの推薦

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

Copyright © 2005 W3C (MIT, ERCIM, Keio) slide 87

The same subsection in Japanese carries the following subcategories:  
Photography, Architecture, Museums, History, Literature.

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 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



## In summary

### Different approaches

### How do I ...

- Ensure that XHTML forms return data in the right encoding?
- Make my Urdu, Arabic or Hebrew text display correctly?
- Declare language and encoding for XML documents?
- Order XSL output according to French rules?
- Approach the creation of multilingual documents in HTML?
- Help users navigate to the right localized page?
- Ensure the table I'm about to write has all the right i18n features?
- etc


Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 90


The GEO Working Group provides information to developers and content authors about how to use international aspects of W3C technologies.

## Summary

### GEO resources



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



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

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

Supporting authors and implementers
W3C®

The screenshot shows two browser windows. The left window displays a table of contents for the 'W3C International Technology Language (i18n)' website. The right window shows the 'Choosing language values' page, which provides detailed guidelines for selecting language codes.

**Table of Contents (Left Window):**

- Language
- Content-Language
- browser settings
- declaring
- flags
- Hans and Hant language codes
- hreflang
- IANA language tags
- ISO language codes
  - ISO 639 Codes for the Representation of Names of Languages
  - 2-letter and 3-letter codes
  - Notes: Two-letter or three-letter language codes
- ISO country codes
- ISO script codes
- language tag values
- language negotiation
- link target language
- primary language
- RFC 3066
- text processing language
- xml:lang


**Choosing language values (Right Window):**

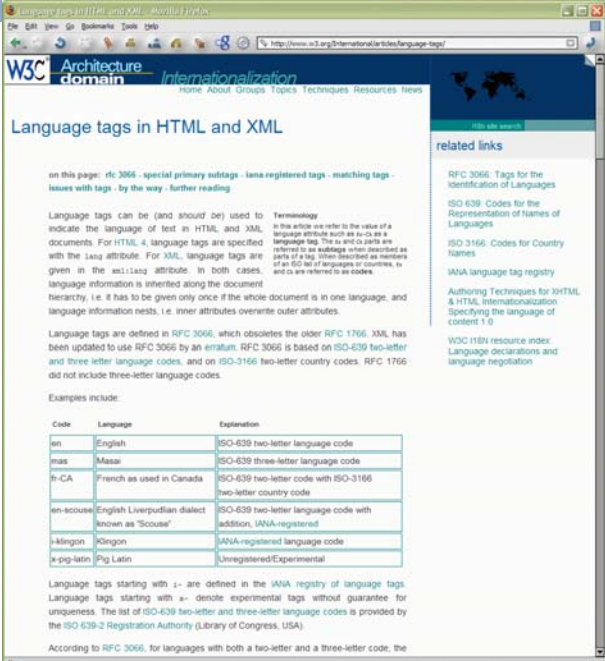
- How to choose language values**
  - W3C techniques document (Choosing Techniques for XHTML, & HTML, Internationalization: Specifying the language of content)
  - Language tags in XHTML and XML
  - How to choose the right attribute values
  - Specifying language tag values (W3C tutorial: Declaring Language in XHTML and HTML)
  - Two-letter or three-letter language codes
  - Should I use two-letter or three-letter language codes? (W3C article)
- Best practices**
  - Follow the guidelines in RFC3066 or its successors for language attribute values
  - Use the two-letter ISO 639 codes for the language code where there are both 2- and 3-letter codes
  - Use the two-letter ISO 639 codes for the language code where there are both 2- and 3-letter codes
  - Where possible, use the codes in their original form to refer to Simplified and Traditional Chinese, respectively
  - Use the codes in their original form to refer to Simplified and Traditional Chinese, respectively
- Particularly useful links**
  - IANA Assigned Language Tags (IANA language tag registry)
  - RFC 3066 Tags for the Identification of Languages (The IETF document that defines how to use language tags to identify languages)
  - ISO 3166: Codes for Country Names (ISO country codes)
  - ISO 639: Codes for the Representation of Names of Languages (ISO language codes)
- Other references**
  - Specifying the language of content: the lang attribute (in the HTML 4.01 spec (section 5.1))
  - Language identification and lang in the XML spec (section 2.12)
- Test data**
  - Automatic font assignment for CJK text (W3C test page)
  - Automatic font assignment for CJK text (W3C test results)

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

There is also a topic index and a techniques index to help you find the information you need. (Note that we have just started developing these, and there is still some way to go, although there is already plenty of useful information there.)


## Supporting authors and implementers





Copyright © 2005 W3C (MIT, ERCIM, Keio)
slide 93

Much of the GEO material is made available as short articles, often answering a specific frequently asked question. There are also tutorials and tests, as well as some summaries of best practices which are still in development.



## 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](mailto:www-international@w3.org))
- offer to help with reviews, or provide local knowledge for other WGs
- provide translations of W3C specifications or articles
- take advantage of the i18n-readiness of W3C technology

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 94



## Summary

Making a difference

- ◆ this is your Web – not the W3C's – if something isn't right, get involved to fix it

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

Copyright © 2005 W3C (MIT, ERCIM, Keio)

slide 95