



License



This presentation and its associated materials licensed under a **Creative Commons Attribution-Noncommercial-No Derivative Works 2.5 License**. You may use these materials without obtaining permission from the author. Any materials used or redistributed must contain this notice. [Derivative works may be permitted with permission of the author.] This work is copyright © 2008-2017 by Addison P. Phillips





W3C[®] Internationalization (i18n) Activity
Making the World Wide Web worldwide!

i18n site search:

RSS Feeds 

Home Resources Techniques Topics News Groups About



The W3C Internationalization (i18n) Activity works with W3C working groups and liaises with other organizations to make it possible to use Web technologies with different languages, scripts, and cultures. From this page you can find articles and other resources about Web internationalization, and information about the groups that make up the Activity.

Learn more about the Activity...

Recent highlights

- ▶ First Public Working Draft: Ethiopic Layout Requirements September 8, 2016
- ▶ New version of Internationalization Checker released September 2, 2016
- ▶ W3C HTML5 Validator enhanced with language detection functionality July 13, 2016
- ▶ For review: Time & date, Essential concepts June 17, 2016
- ▶ by markup June 13, 2016
- ▶ Changing an HTML page to Unicode May 26, 2016
- ▶ the final OntoLex specification: lexicon model for May 20, 2016
- ▶ by Markup May 3, 2016

Get help

task-based help

site search

request a review

getting started

Quick links

- ▶ Current projects
- ▶ Articles etc.
- ▶ Tech reports
- ▶ Typography index
- ▶ Review radar
- ▶ Review comments, github
- ▶ Github issues
- ▶ Lists & archives
- ▶ i18n wiki
- ▶ Tests
- ▶ i18n checker
- ▶ MultilingualWeb
- ▶ @webi18n

The Internationalization Activity@W3C

Participate!

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

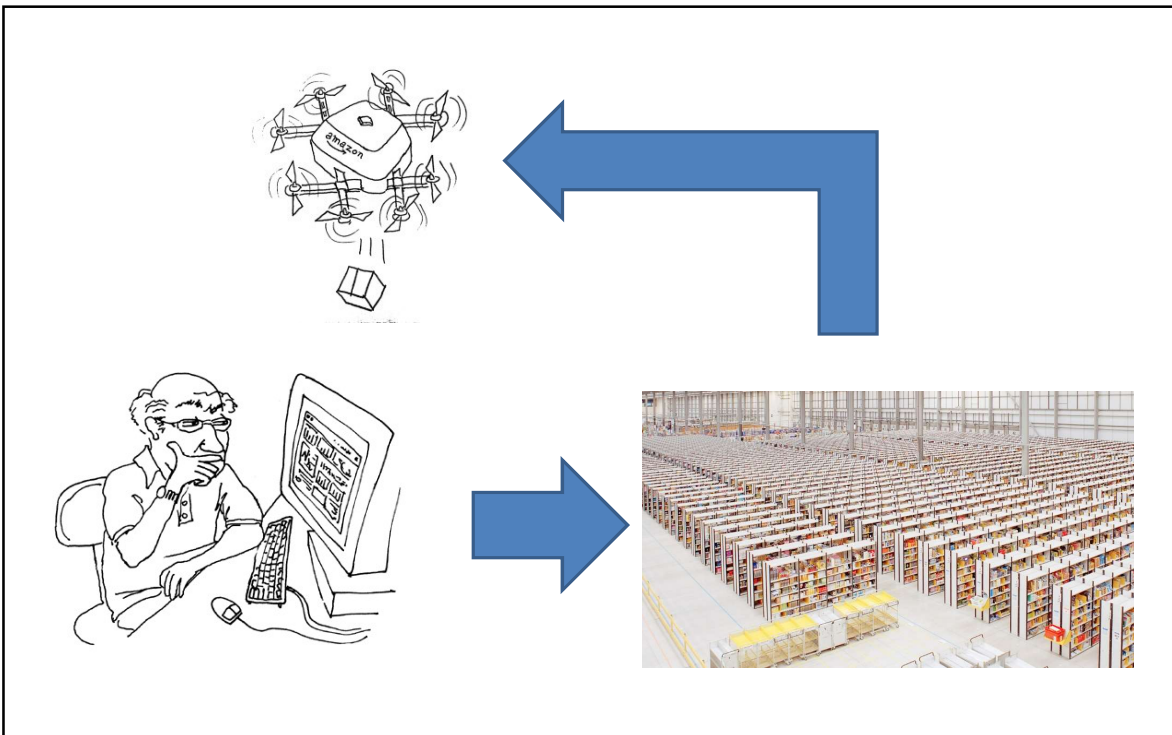
Group pages

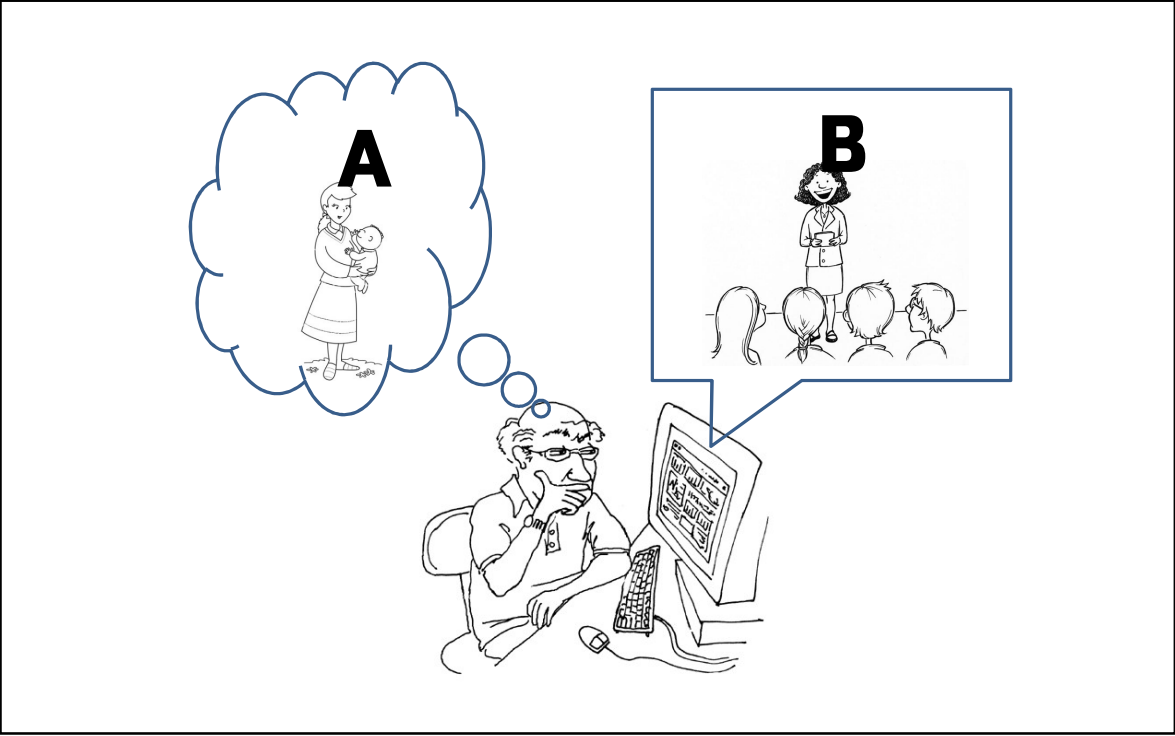
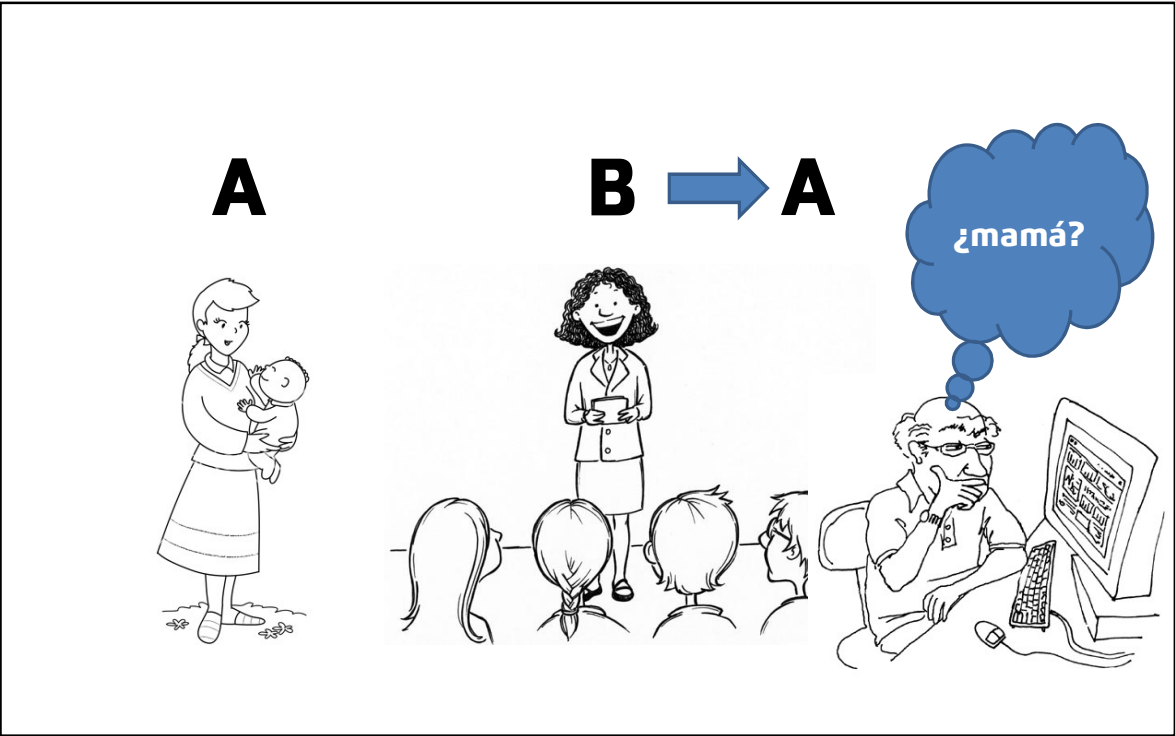
- ▶ Activity Statement
- ▶ i18n WG
- ▶ i18n Interest Group
- ▶ i18n Tag Set (ITS) IG
- ▶ Arabic Layout Task Force
- ▶ Chinese Layout Task Force
- ▶ Ethiopic Layout Task Force
- ▶ Indic Layout Task Force
- ▶ Community groups

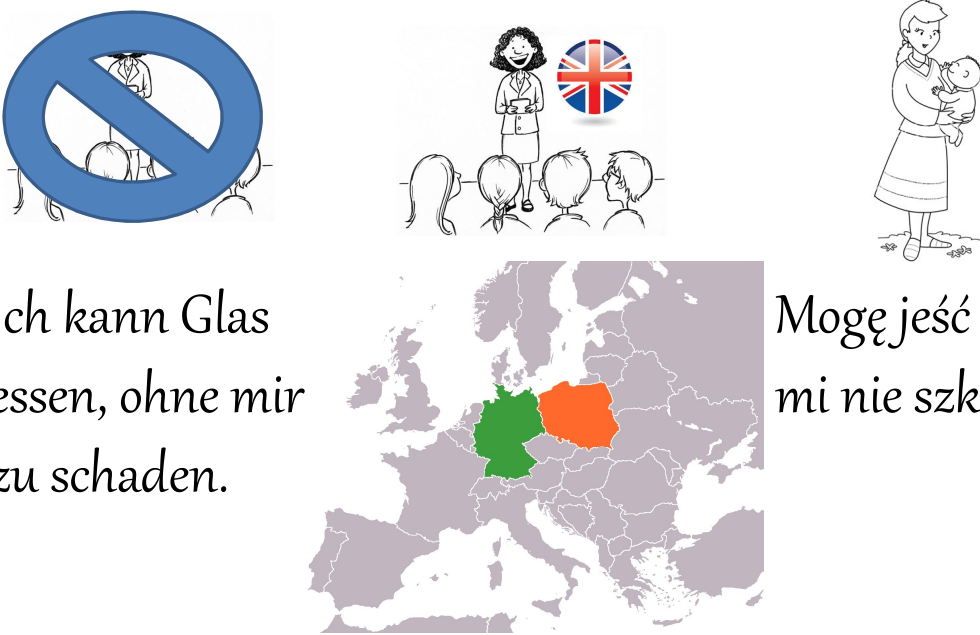
Working Draft: Ethiopic Layout Requirements

The Internationalization Working Group has published a First Public Working Draft of Ethiopic Layout Requirements.

This document describes requirements for the layout and presentation of text in languages that use the Ethiopic script when they are used by Web standards and technologies, such as HTML, CSS,

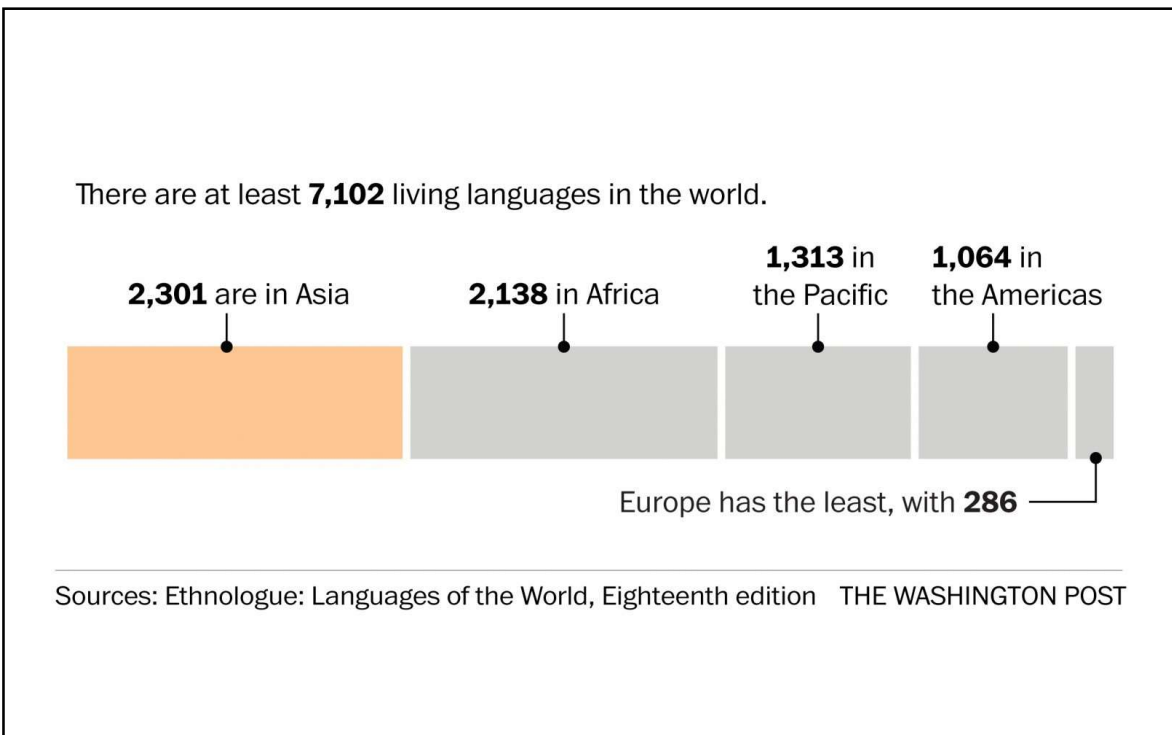






Ich kann Glas essen, ohne mir zu schaden.

Mogę jeść szkło i mi nie szkodzi.



Population range	Living languages			Number of speakers		
	Count	Percent	Cumulative	Total	Percent	Cumulative
100,000,000 to 999,999,999	8	0.1	0.1%	2,529,403,578	40.20547	40.20547%
10,000,000 to 99,999,999	82	1.2			39.42144	79.62691%
1,000,000 to 9,999,999	304	4.3			14.55462	94.18154%
100,000 to 999,999	943	13.3	18.8%	296,136,843	4.70717	98.88870%
10,000 to 99,999	1,822	25.7	44.5%	61,802,734	0.98237	99.87107%
1,000 to 9,999	1,982	27.9			0.12133	99.99241%
100 to 999	1,065	15.0			0.00738	99.99979%
10 to 99	338	4.8	92.1%	12,111	0.00020	99.99999%
1 to 9	140	2.0	94.1%	560	0.00001	100.00000%
0	206	2.9	97.0%	0	0.00000	100.00000%
Unknown	212	3.0	100.0%			
Totals	7,102	100.0		6,291,192,624	100.00000	



Choose Your Language



Culture

A combination of shared, pre-agreed, conventions and experiences.

Culture is arbitrary.

Why 360 degrees?

Base-60 Number System

"About" 360 days in a year

How is the stock market doing?

Going Up

Going Down

上海證券交易所
SHANGHAI STOCK EXCHANGE

www.sse.com.cn

2013-10-11

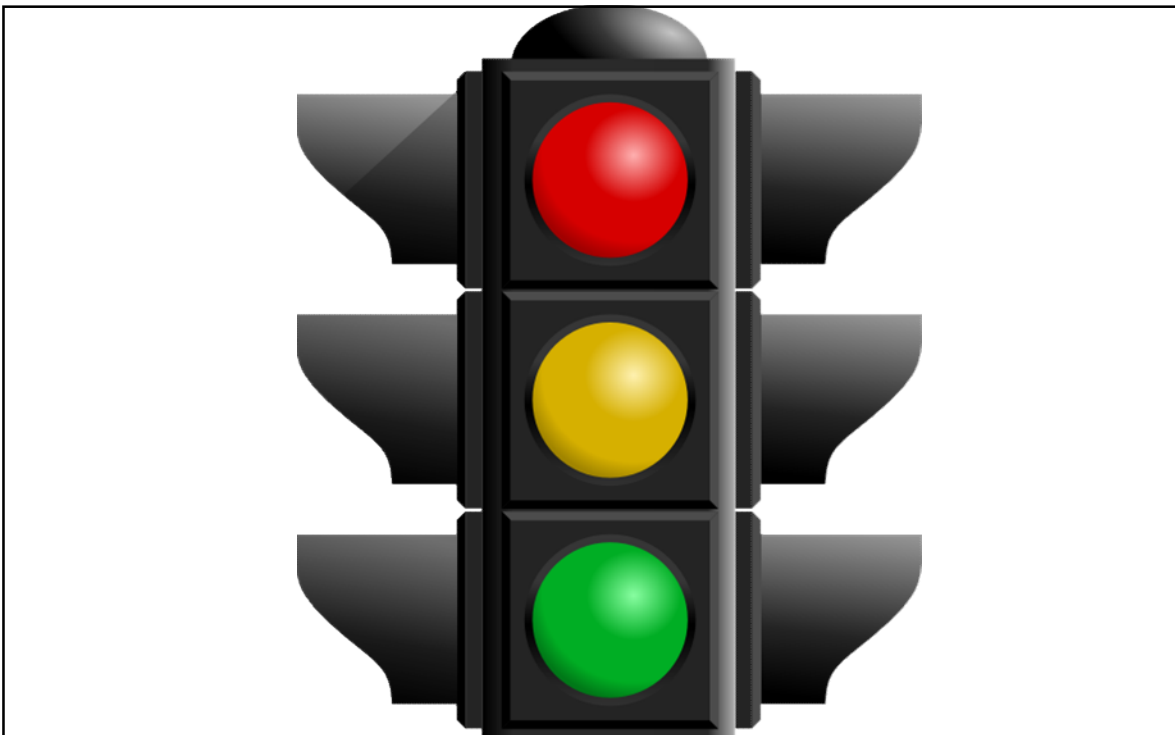
指数名称	最新	涨跌%
SSE 上证50	1669.27	+1.91
SSE 上证380	3956.69	+2.06
SSE 上证100	4074.45	+2.03
SSE 上证150	3407.13	+1.77
SSE 上证综指	2228.15	+1.70
SSE 国债指数	139.74	+0.01
SSE 180红利	2177.96	+2.04
SSE 上证转债	294.24	+0.38
SSE 上证潜力	2814.90	+1.92
SSE 上证回报	2774.22	+1.84
SSE 5年信用	148.47	+0.02
CSI 沪深300	2468.51	+1.61
CSI 中证100	2285.30	+1.59
中华120	4059.56	+1.33



Which way should I go?

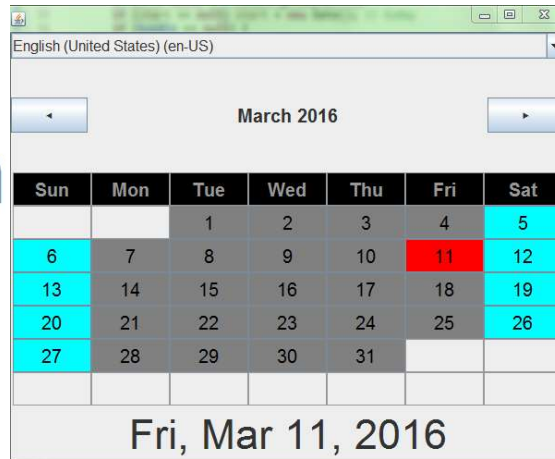


What's wrong with this magazine?



What do you know?

Th 16



English (United States) (en-US)

March 2016

Sun	Mon	Tue	Wed	Thu	Fri	Sat
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Fri, Mar 11, 2016

Your Culture

IS THE SEA YOU ARE SWIMMING IN



Internationalization

The design and development of a product that is **enabled** for target audiences that vary in culture, region, or language. [W3C]

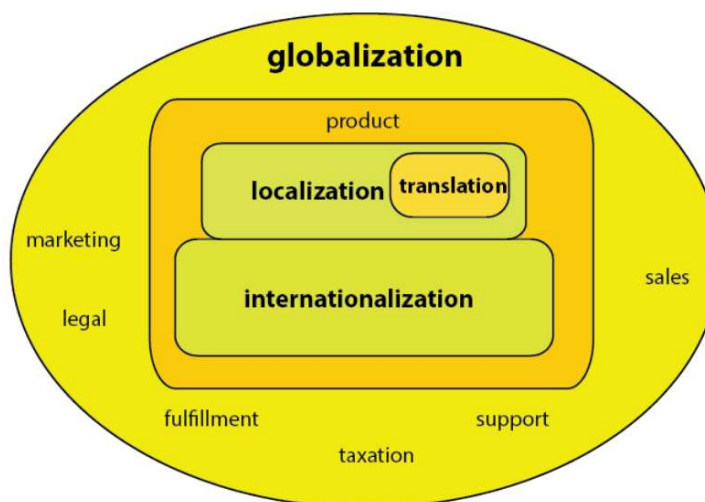
Internationalization is enabling a customer's cultural needs in software

English (United States) Locale	қазақ тілі (кириллица,Қазақстан) Locale	中文 (简体中文,中国) Locale
en_US	kk_KZ_#Cyr1	zh_CN_#Hans
MessageFormat	ICU Skeleton Dates	
Choose Date:	Choose Date:	Choose Date:
123,456 <small>(1,number,integer)</small>	12 vs. 24 hour time preference	2015年5月1日
-123,456,789,012 <small>(2,number)</small>		司五20155115593GMT-07:00
5/1/15 <small>(3,date,short)</small>		AF pattern {0,date,EyMdHmsz}
May 1, 2015 <small>(3,date,medium)</small>	1 мам. 2015 15:59 <small>>pattern: d MMM y HH:mm skeleton24:yMMMMdHm</small>	5/1周五 <small>:IMd</small>
May 1, 2015 <small>(3,date,long)</small>	1 мам. 2015 3:59 түстен кейінгі <small>>pattern: d MMM y h:mm a skeleton12:yMMMMdhma</small>	5月1日周五 <small>:MMMMd</small>
Friday, May 1, 2015 <small>(3,date,full)</small>	б.з. 2015 мамыр 1, жұма 15:59:03 Тынық уақыты <small>>pattern: GGG y MMMM d, EEEE HH:mm:ss zzzz skeleton24:GGGEE</small>	5月01日星期五 <small>:EEEEMMdd</small>
3:59 PM <small>(3,time,short)</small>		5月1日星期五 <small>:EEEEMMMd</small>
3:59:03 PM <small>(3,time,medium)</small>	б.з. 2015 мамыр 1, жұма 15:59:03 Тынық уақыты <small>>pattern: GGG y MMMM d, EEEE HH:mm:ss zzzz skeleton12:GGGEE</small>	下午3:59:03 <small>:imsa</small>
3:59:03 PM PDT <small>(3,time,long)</small>		下午3:59:03 <small>:ims</small>
3:59:03 PM Pacific Daylight Time	жұма, 1 мамыр 2015 15:59 ГОУ-7 <small>>pattern: EEEE, d MMMM y HH:mm z skeleton24:EEEEyMMMMdHm</small>	15:59:03

Related Concepts

- **Localization**: creation of a product tailored to a particular target market
- **Translation**: process of converting text from one language to another
- **Globalization**: unified approach to creating global products, especially those that support multiple geographies simultaneously

Terminology Map

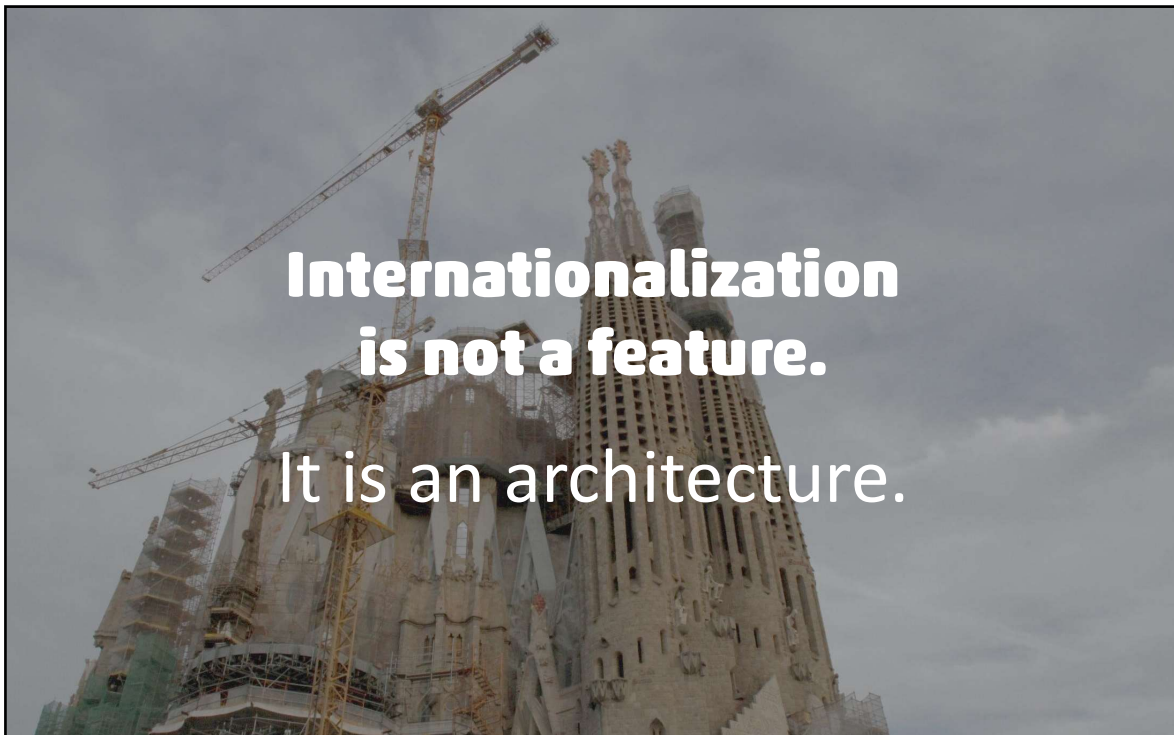


Mystic Numbering

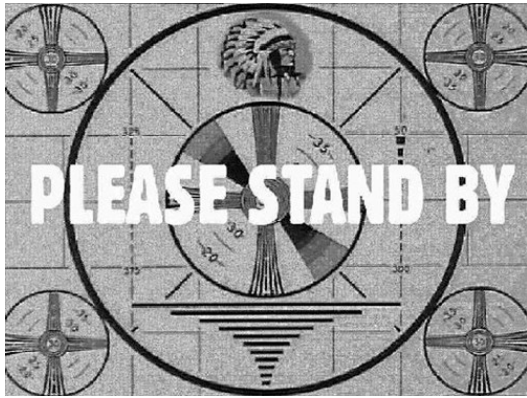
I 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 N

I18N

Localization	=	L10N
Globalization	=	G11N
Canonicalization	=	C14N
Accessibility	=	A11Y



Globalized Product Development



Technical Problems



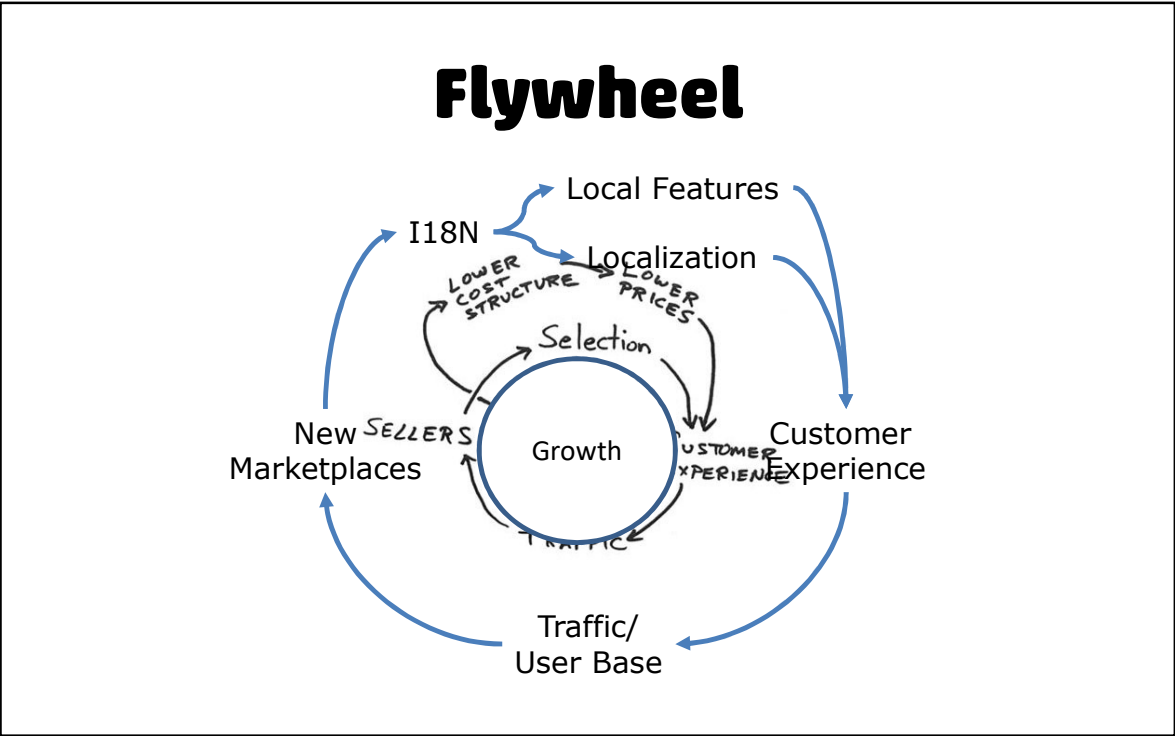
Business Decisions





THE INTERNATIONALIZATION SOFTWARE DEVELOPMENT CYCLE

“Well, it depends...”



Get Close To Your In-Country Resources
*Product managers, lawyers, sales people, etc. etc. And **especially** customers.*

Requirements

Get to know your international customers

Design

Does your design consider international needs?

Does your design consider local requirements?

UX Design

Your greatest savings in effort and cost will be realized if you design with international customers and localization in mind.

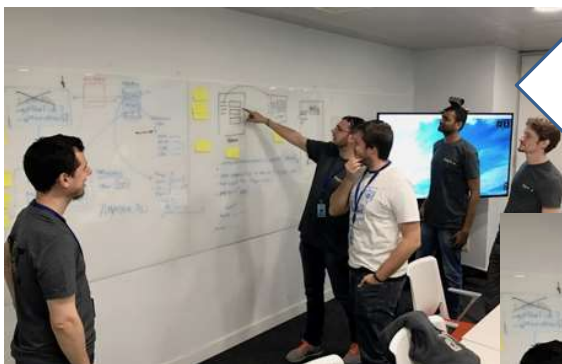
Myths

- The translator will have to abbreviate if it won't fit.
- The translation can wrap around.
- The translation can truncate with ellipses.
- Just make the font smaller.
- This feature isn't used in country X.
- We can just remove one of the features in the target language.

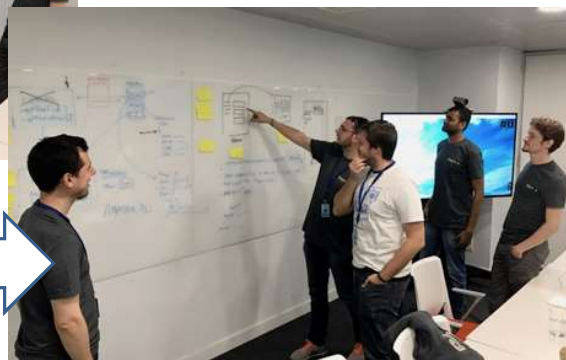
**We're making best possible English
and we can separately make the best
target language.**

Are you really going to build a separate product
with a separate code base, build, testing,
release, support, maintenance, features, etc.?

Development



The Regular Engineering Team



The I18N Engineering Team

Testing

- Does the enabled product work correctly?
 - Non-English configurations
 - Non-ASCII data and encoding support
 - Cross time zone support
 - Market specific features or customizations
- Does localization appear correctly?
 - Is the product localizable?

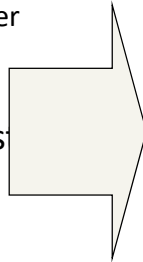
What to Test With

- Test Non-English configurations
 - Non-English locales
 - Native configurations
- Test Non-ASCII data
 - Encodings, encodings, everywhere
 - Non-ASCII character values
- Test Across Time Zones
 - Two or more time zones; consider international date line (“it’s tomorrow in Japan”) and DST issues

Planning Testing

Initially

- Get tools that are enabled!
 - Automation allows greater coverage, but only if it works.
- Plan encodings and locales as part of the test matrix.
- Acquire third-party products as necessary.
- Pseudo



Increasing Maturity

- Use test driven development practices.
- Get developers to write unit tests that are internationalized.
- Put the 'i18n' bugs into the regression suite.
- Smoke/BAT and other tests automated in all locales

Configuring Machines

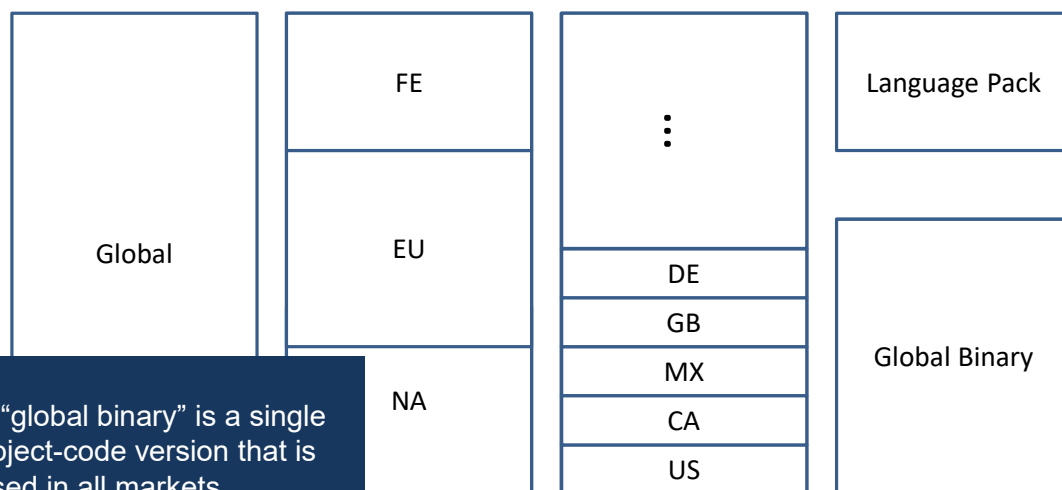
Create both native and simulated environments:

- Native operating systems may have minor but sometimes critical differences (folder names, keywords, localized registry entries)
- Most features don't run into native differences (easier to work with English-localized machines)
- Don't buy physical keyboards (use software keyboards)*

Completing the Product

- Don't forget:
 - Demos and Demo Data
 - Dictionary, Language add-ons
 - Local offers, links to Web store, etc.
 - Packaging
 - Regulatory
 - Customer service

Delivering the Product



A “global binary” is a single object-code version that is used in all markets, regardless of localization.

Simultaneous Shipment (Simship)

When you ship the target languages on the same day as the source language.

- It might not make sense for your product.
- But it might not be as difficult as you think it is.
- It promotes good internationalization—and other software development best practices.




ENABLING

Making Code Aware of Culture

What is “enabling”?

- Enabled software:
 - adapts the display, processing, validation, storage, and transmission of data according to the cultural, linguistic, and regional needs of the users
 - Text, Characters, and Encodings
 - Locale Awareness
 - Times and Time Zones

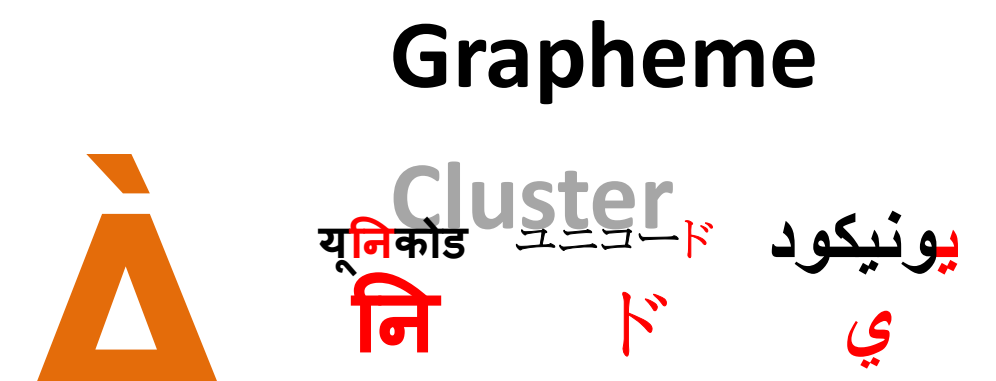




Glyph

यूनि कोड يونيكود

A single rendered shape: what the user sees on the screen, paper, or as output



Grapheme


Cluster





यूनि कोड يونيكود यूनि कोड يونيكود

न ड ي

A single visual unit of text: the smallest abstract unit of meaning in a writing system.

Character










A single logical unit of text

	A84	A85	A86	A87
0	ॐ	ॐ	ॐ	ॐ
1	ॐ	ॐ	ॐ	ॐ
2	ॐ	ॐ	ॐ	ॐ
3	ॐ	ॐ	ॐ	ॐ
4	ॐ	ॐ	ॐ	ॐ
5	ॐ	ॐ	ॐ	ॐ
6	ॐ	ॐ	ॐ	ॐ
7	ॐ	ॐ	ॐ	ॐ
8	ॐ	ॐ	ॐ	ॐ
9	ॐ	ॐ	ॐ	ॐ
A	ॐ	ॐ	ॐ	ॐ
B	ॐ	ॐ	ॐ	ॐ
C	ॐ	ॐ	ॐ	ॐ
D	ॐ	ॐ	ॐ	ॐ
E	ॐ	ॐ	ॐ	ॐ
F	ॐ	ॐ	ॐ	ॐ

Character Set

Character Repertoire



AaBbCcDd... 

A set of characters

r12a >> apps >> UniView 10.0.0

<https://r12a.github.io/uniview/>

Coded Character Set

A set of characters in which each character is assigned a numeric identifier.

r12a >> apps >> UniView 10.0.0

text area

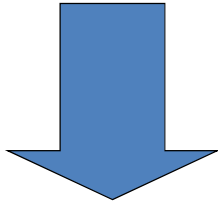
U+00C0 LATIN CAPITAL LETTER A WITH GRAVE

Code Point

Unicode Scalar Value

- Canonical combining class: 0 - Spacing, split, enclosing, reordrant, & Tibetan subjoined
- Bidirectional category: L - Left-to-right
- Character decomposition mapping: 0041 0300 À
- Unicode 1.0 name: GRAVE
- Lowercase mapping: 00E0 à
- Unicode version: 1.1

010000010101101101101000



Code Unit

010000010101101101101000 (0x41)
byte

A unit of physical storage and information interchange
Other code units exist (16-bit, 32-bit, etc.)



U+00C0



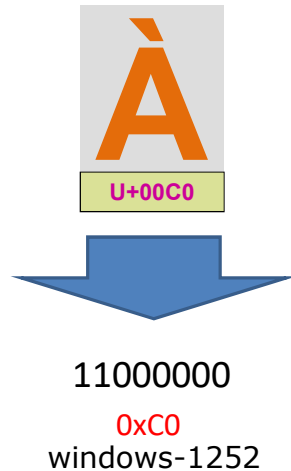
11000011 10000000

0xC3 0x80
UTF-8

Character Encoding Form

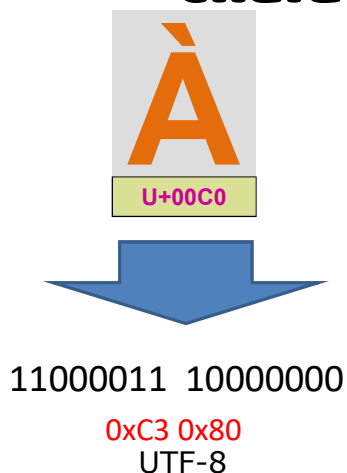
Maps code points to code units

Single Byte Character Encoding Form



- The simplest character encoding form: one character maps to one code unit (byte)
- Good for languages with small character sets and limited display requirements

Variable Width Character Encoding Form



- One character maps to differing numbers of code units (which could be bytes)

Multibyte Encoding
a variable width encoding form
whose code unit is the byte.

Shift-JIS: A Multibyte Encoding

- In order to reach more characters, Shift-JIS multi-byte characters start with a limited range of "lead bytes"
- These can be followed by a larger range of byte values ("trail byte")

Microsoft Windows Codepage : 932 (Japanese Shift-JIS)

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	MUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018							
20	SE 0020	!	!"	#	\$	%	&	'	(
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038							
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048							
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	¥ 005C]	^ 005E	~ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	007C	} 007D	DEL 007F	
80		81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A0		┌ 00A1	┐ 00A2	、 00A3	・ 00A4	ヲ 00A5	ァ 00A6	ィ 00A7	ウ 00A8	ェ 00A9	オ 00AA	ヤ 00AB	ユ 00AC	ヨ 00AD	ツ 00AE	ッ 00AF
B0	ー 00B0	ァ 00B1	ィ 00B2	ウ 00B3	ェ 00B4	オ 00B5	カ 00B6	キ 00B7	ク 00B8	ケ 00B9	コ 00BA	サ 00BB	シ 00BC	ス 00BD	セ 00BE	ソ 00BF
C0	タ 00C0	チ 00C1	ツ 00C2	テ 00C3	ト 00C4	ナ 00C5	ニ 00C6	ノ 00C7	ハ 00C8	ヒ 00C9	フ 00CA	ヘ 00CB	ホ 00CC	ベ 00CD	ペ 00CE	ポ 00CF
D0	モ 00D0	ム 00D1	メ 00D2	ヤ 00D3	ユ 00D4	ヨ 00D5	ラ 00D6	リ 00D7	ル 00D8	レ 00D9	ロ 00DA	ワ 00DB	ヰ 00DC	ヱ 00DD	ヰ 00DE	ヱ 00DF
E0	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC			

Single byte characters in Shift-JIS don't have a lead byte

Shift-JIS

- Lead bytes can be trail byte values
- Trail bytes include ASCII values
- Trail bytes include special values such as 0x5C ("\\")

@ あ 漾 烽

```
int pos = strchr(mybuf, '@');
```

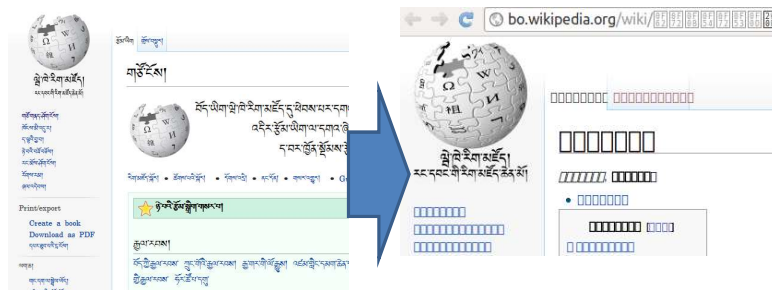
0x40 0x82 0xA0 0xE0 0x40 0xE0 0x82

Unicode / ISO-10646



Tofu

Squares (or other symbols) show instead of the characters because local fonts don't have a glyph (picture of the character).



Check with:

<http://r12a.github.io/apps/conversion/>

Mojibake: what is it?

Amazon.com ソース philips: Sci

文字化け == “garbled characters”

المغادرون				الرحلة	11:42:41
البوابة	الحالة	الوقت المقدر	الوقت	الى / عن طريق	
7	ÉÇĪÑÉ	13:05	11:35	BæŌiä 9W 533	JET AIRWAYS
2	ááØÇÆÑÉ	12:15	12:15	ÇáÉĪÑiä GF 563	طيران الخليج GULF AIR
16	ááØÇÆÑÉ	12:30	12:30	ŌáÇáÉ WY 927	
1	Çái ÇáÈæÇÈÉ	12:30	12:30	ŌíÇáBæÉ PA 951	airblue
8	Çái ÇáÈæÇÈÉ	12:45	12:45	ĪÈí FZ 038	فلاي دبي flydubai
18	Çái ÇáÈæÇÈÉ	12:55	12:55	Īáái WY 241	
5	Çái ÇáÈæÇÈÉ	13:30	13:30	āīīŌäá OX 3303	OXY
4	Çái ÇáÈæÇÈÉ	13:35	13:35	ÇÈæ ÛÈí / ĪæÇĪÑ PK 192	PK
22	Ýí äæÚĪáÇ	13:40	13:40	ĪBÇ WY 315	
8	Ýí äæÚĪáÇ	13:50	13:50	ÈÑĪÇáĪÑä WY 215	

Question Marks



Ransom Note

私たちは、正しい文字を見ることはできますか？
 我们能看到正确的字符吗？
 我們能看到正確的字符嗎？

In memory, on disk, on the network, etc.

**All text has a
character encoding**

When things go wrong, start by asking what the encoding is, what encoding you expected it to be, and whether the bytes match the encoding.

Unicode Encodings

- Characters (*graphemes*) can require multiple Unicode codepoints.

न ि

 यूनिकोड



UTF-16

- Each character (*codepoint*) can require either 1 or 2 code units

語

U+8A9E

0x8A9E



U+1F404

0xD83D.DC04



UTF-8

- 7-bit ASCII is UTF-8: all other characters (*codepoints*) require 2, 3, or 4 bytes.

A

U+0041

0x41

Á

U+00C0

0xC3.80

語

U+8A9E

0xE8.AA.93



U+1F404

0xF0.9F.90.84



Counting Things

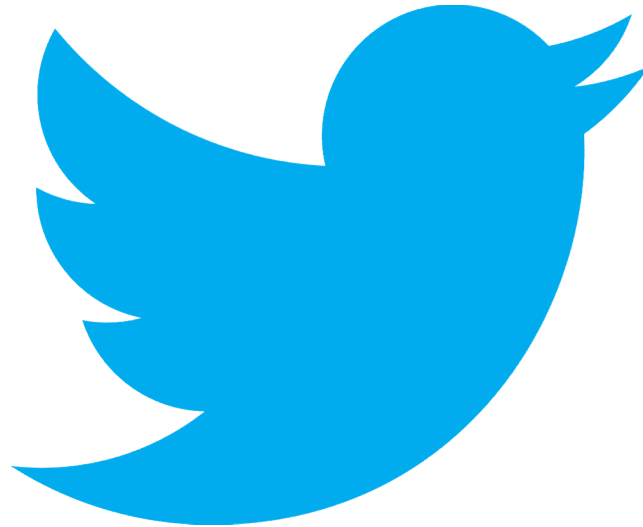
5
9
10
28

यूनि कोड 

092F 0942 0928 093F 0915 094B 0921 1F404 FE0F

092F 0942 0928 093F 0915 094B 0921 D83D DC04 FE0F

E0 A4 AF E0 A5 82 E0 A4 A8 E0 A4 BF E0 A4 95 E0 A5 8B E0 A4 A1
F0 9F 90 84 EF BC 8F



Combining Emoji

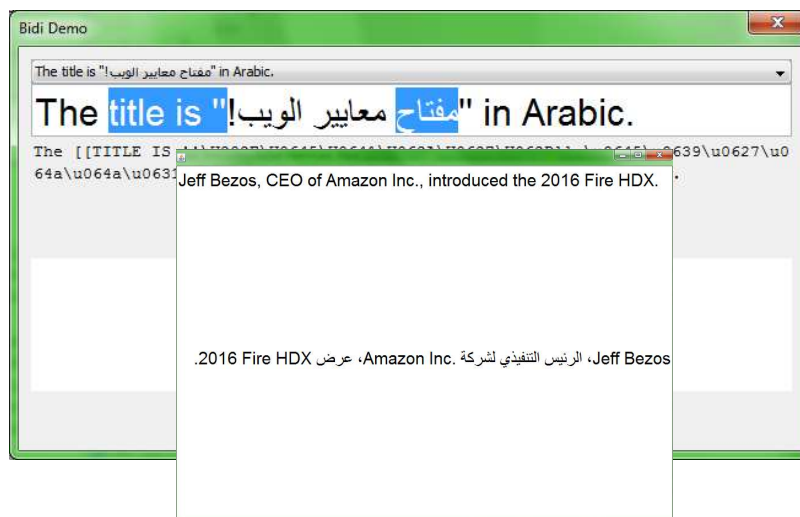


U+1F45A



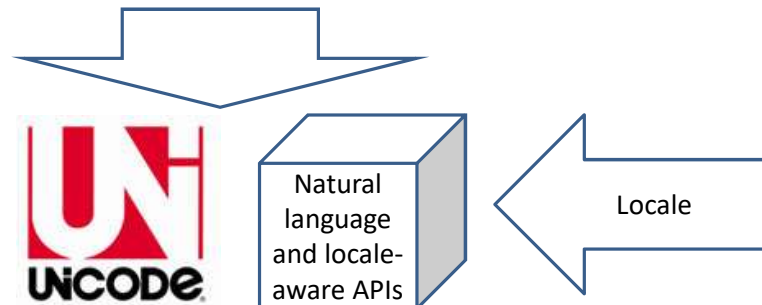
U+1F468 U+200D U+1F469 U+200D
U+1F467 U+200D U+1F467

Bidirectionality



Text Processing

- Line breaking. Sentence breaking. Finding words. Hyphenation. Text selection. UPPER or lower or Titlecasing. Sorting. Searching. Regular expression matching. And more...



Locale

- an identifier or data structure that allows programmers to access culturally and linguistically affected functionality in a system.

Many systems based on IETF BCP 47.

Unicode's **C**ommon **L**ocale **D**ata **R**epository project (CLDR) uses BCP 47.

Months, Days, Abbreviations

English (United States) (en-US)	▼	Jan	Sat
français (France) (fr-FR)	▼	Janv.	sam.
Deutsch (Deutschland) (de-DE)	▼	Jan	Sa
中文 (中国) (zh-CN)	▼	1月	周六
العربية (مصر) (ar-EG)	▼	يناير	السبت
हिंदी (भारत) (hi-IN)	▼	जन.	शनि
русский (Россия) (ru-RU)	▼	янв.	сб
español (España) (es-ES)	▼	Ene.	Sáb.
ไทย (ประเทศไทย) (th-TH-u-nu-thai-x-lvariant-TH)	▼	ม.ค.	ส.

Numbers

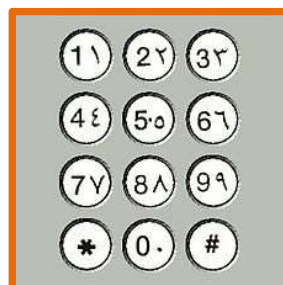
Tutorial Number Demo

-123456789.9876

Language	Formatted Number
English (United States) (en-US)	-123,456,789.988
français (France) (fr-FR)	-123 456 789,988
Deutsch (Deutschland) (de-DE)	-123.456.789,988
中文 (中国) (zh-CN)	-123,456,789.988
français (Suisse) (fr-CH)	-123'456'789.988
हिंदी (भारत) (hi-IN)	-12,34,56,789.988
(ar-EG) العربية (مصر)	١٢٣,٤٥٦,٧٨٩,٩٨٨-

Digit Shaping

default	-123,456.789
arab	-١٢٣,٤٥٦,٧٨٩



Lists and Sorting



Sorting and Organizing Information

“Alphabet” differences

ASCII vs. the world

Mixed language information sets

123	123	123	123
			a
			b
c	サ	г	с
d	タ	г	d
e	ナ	д	e
f	ハ	её	f
g	マ	ж	g
h	ヤ	з	h
i	ラ	ий	i
j	ワ	к	j
k		л	k
l		л	l
m		н	m
n		о	n
o		п	o
p		р	p
q		с	q
r		т	r
s		у	s
t		ф	t
u		х	u
v		ц	v
w		ч	w
x		шц	x
y		ъ	y
z		ы	z
		ь	г
		э	а
		ю	б
		я	

Unicode Collation

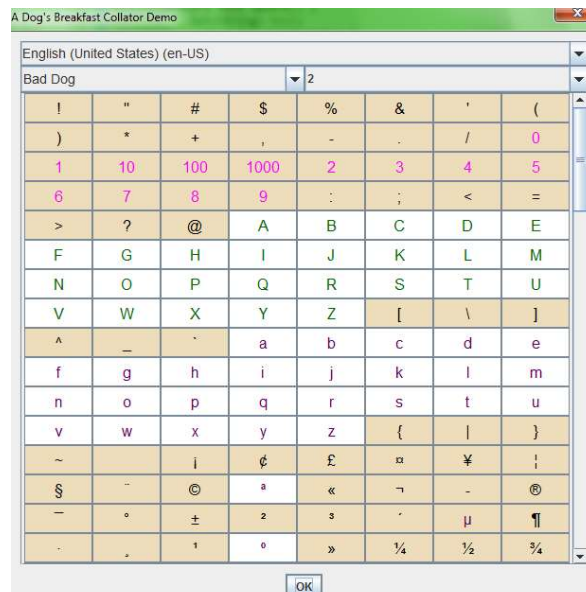
- Unicode Technical Standard #10: Unicode Collation Algorithm (aka “UCA”) is the basis for many implementations of sorting Unicode



Unicode Technical Standard #10 UNICODE COLLATION ALGORITHM

Version	6.2.0
Editors	Mark Davis (markdavis@google.com), Ken Whistler (ken@unicode.org), Markus Scherer
Date	2012-08-30
This Version	http://www.unicode.org/reports/tr10/tr10-26.html
Previous Version	http://www.unicode.org/reports/tr10/tr10-24.html

Collation Demo



What's the problem?

Your Kindle Library

View: OR

Showing 466 - 480 of 764 items

Title ▼	Author
<input type="checkbox"/> The Dog Said Bow-Wow	Swanwick, Michael
<input type="checkbox"/> The Door Through Space	Bradley, Marion Zimmer
<input type="checkbox"/> The Door into Summer	Heinlein, Robert A.
<input type="checkbox"/> The Draco Tavern	Niven, Larry
<input type="checkbox"/> The Dragon Book: Magical Tales from the Masters of Modern Fantasy	Dann, Jack
<input type="checkbox"/> The Dragon and the Raven	Henty, George Alfred
<input type="checkbox"/> The Dragon-Child (Hyborean Dragons)	Lawson, M. P.
<input type="checkbox"/> The Drawing of the Dark (Del Rey Impact)	Powers, Ian
<input type="checkbox"/> The Drowning City (The Necromancer Chronicles 1)	Downum, Amanda
<input type="checkbox"/> The Duellists	Keith Carradine, Harvey Keitel

I have 12 pages of "T"?


What did you say?

- Some language content is sorted by pronunciation. Chinese and Japanese, for example.
 - text doesn't contain this information directly!



“Should I be writing all of this down...?”

- Wide range of variation
- Obscure formats
- Difficult to obtain reliable information on formats
- Lots of work to implement and maintain



Enabling means not having to know (m)any of the details

Locale Aware APIs

- Each programming language or operating environment provides its own. Sometimes these are complimentary (e.g. Java + Android)
- Add-on libraries (notably ICU) or platform specific add-ons (for example, PHP *intl* mod) are sometimes useful.
- (Go find and learn yours)

CLDR

Unicode CLDR Project

News

- 2013-05-15: [CLDR v23.1 Released](#)
- 2013-04-30: [CLDR v24 Open for Data Submission](#)
- 2013-03-15: [CLDR v23 Released](#)

What is CLDR?

The Unicode CLDR provides key building blocks for software to support the world's languages, with the largest and most extensive standard repository of locale data available. This data is used by a wide spectrum of companies for their software internationalization and localization, and software to the conventions of different languages for such common software tasks. It includes:

- Locale-specific patterns for formatting and parsing:** dates, times, timezones, numbers and currency values
- Translations of names:** languages, scripts, countries and regions, currencies, eras, months, weekdays, day periods, timezones, cities, and time units
- Language & script information:** characters used; plural cases; gender of lists; capitalization; rules for sorting; searching; writing direction; transliteration rules; rules for spelling out numbers; rules for segmenting text into words, and sentences
- Country information:** language usage, currency information, calendar preference and week conventions, postal telephone codes
- Other:** ISO & BCP 47 code support (cross mappings, etc.), keyboard layouts

A Unicode Consortium project to gather high-quality locale data. Used by many (not all) operating environments.

CLDR Charts

Locale Data Summary for de_CH [Swiss High German]

CLDR Version 23 [Other charts and help](#) 2013-02-22 1


[root \[Root\]](#) > [de \[German\]](#) > [de_CH \[Swiss High German\]](#)

No	Path	English	Parent	Native	D?
1	names currency	CHF:symbol	=	CHF	
2	names language	be	Weißrussisch	Weissrussisch	
3	names territory	BD	Bangladesch	Bangladesh	
4	names territory	BN	Brunei Darussalam	Brunei	
5	names territory	BW	Botsuana	Botswana	
6	names territory	BY	Belarus	Weissrussland	
7	names territory	CV	Cape Verde	Kapverden	
8	names territory	DJ	=	Dschibuti	Djibouti
9	names territory	GB	United Kingdom	Vereinigtes Königreich	Grossbritannien
10	names territory	MH	Marshall Islands	Marshallinseln	Marshall-Inseln
11	names territory	QO	Outlying Oceania	Außeres Ozeanien	Ausseres Ozeanien
12	names territory	RW	=	Ruanda	Rwanda
13	names territory	SB	Solomon Islands	Salomonen	Salomon-Inseln
14	names territory	ST	São Tomé and Príncipe	São Tomé und Príncipe	Sao Tomé und Principe
15	names territory	ZW	=	Simbabwe	Zimbabwe
16	misc delimiters	alternateQuotationEnd	'	'	
17	misc delimiters	alternateQuotationStart	'	'	
18	misc delimiters	quotationEnd	"	"	
19	misc delimiters	quotationStart	"	"	
20	number pattern	currency	¤#,###0.00;(¤#,###0.00)	¤ #,###0.00 ¤	¤ #,###0.00;¤ #,###0.00
21	number symbol	decimal	.	.	
22	number symbol	group	,	,	

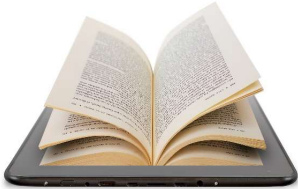
Data Structures

- Identify your locale bias
 - Field names matter!
 - “Postal Code”, not “ZIP code”.
 - Family Name/Given Name, not First Name/Last Name
 - Avoid problems if possible.
 - Postal address parsing? Area code? Etc.
 - Identify your assumptions.






Customer Locale



Content Language



Jurisdiction

Time Zones

October 2016						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

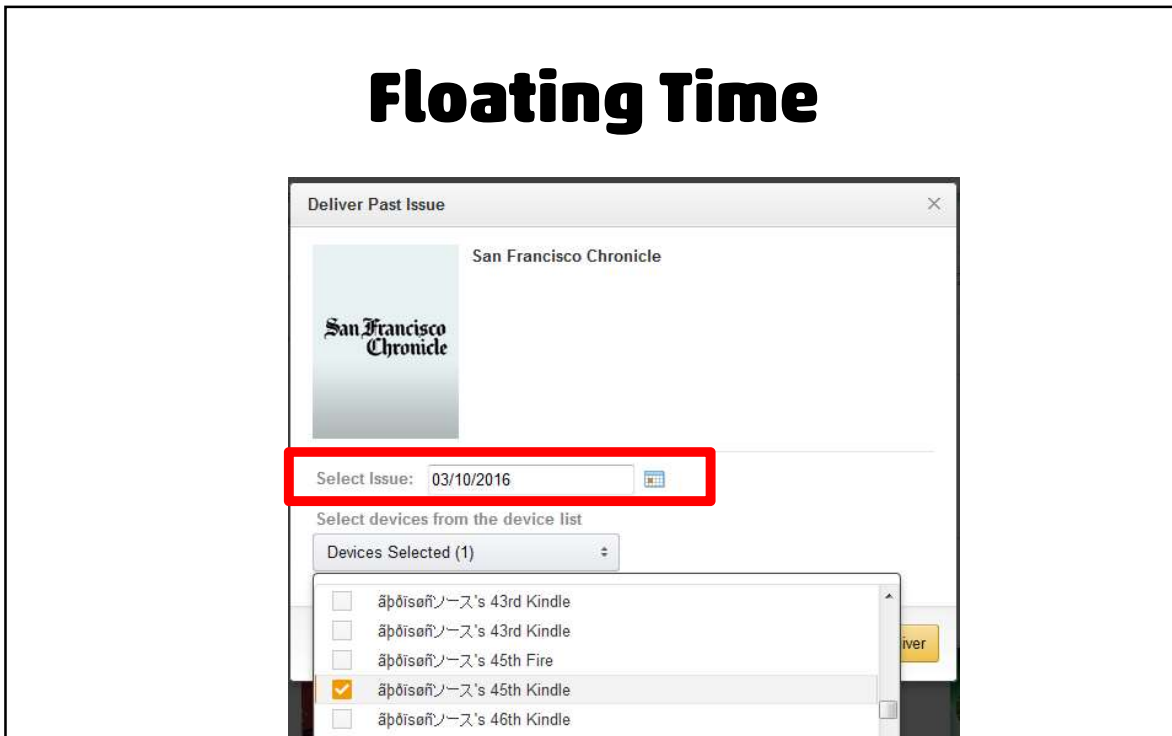
North America

Pacific Standard Time (Los Angeles) GMT-0700

5:30 PM Set

Coordinated Universal Time Oct-29 00:30	Japan Standard Time Oct-29 09:30	India Standard Time Oct-29 06:00
Central European Summer Time Oct-29 02:30	Eastern Daylight Time Oct-28 20:30	Pacific Daylight Time Oct-28 17:30
Hawaii Standard Time Oct-28 14:30	GMT+14:00 Oct-29 14:30	GMT-12:00 Oct-28 12:30

Floating Time



ENABLING SUMMARY

Understand Encodings and Unicode

All text has an encoding!

Be Locale-Aware

Create locale-neutral data structures

Separate display from storage

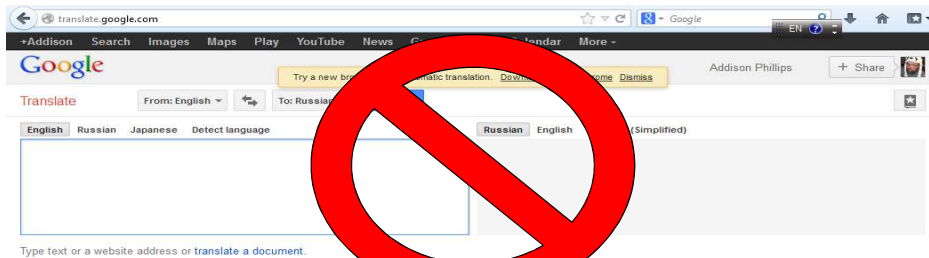
Understand time and time zones



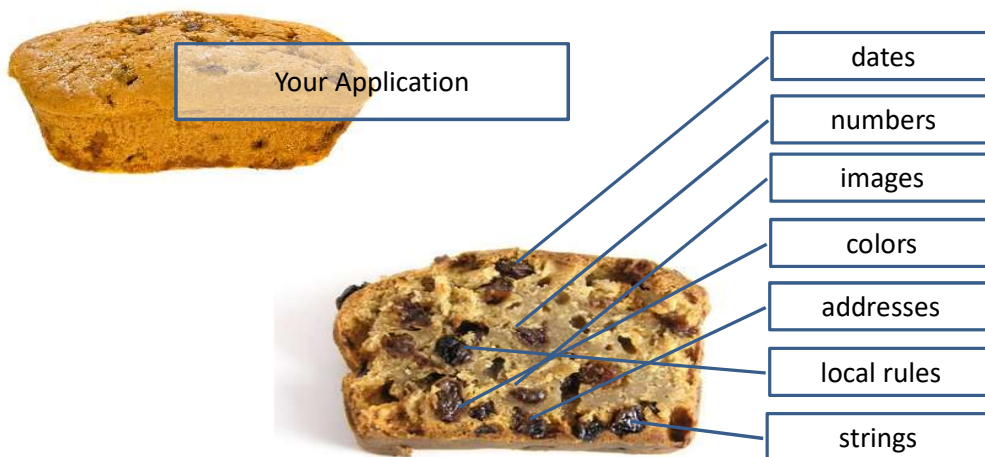
Externalization

Making software localizable

Localization

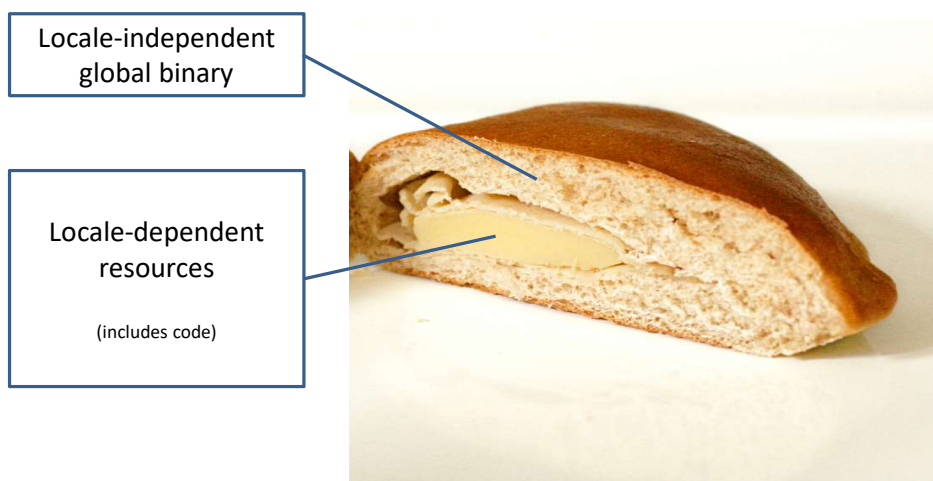


The Problem



local rules, regulatory requirements, postal addresses, default bookmark lists, your company's customer service phone numbers

The Solution



Why Resources?

Before



Text
Error messages
Icons
Pictures
Fonts
Colors
Graphics
Sizes
Positions
Magic Numbers
Mnemonics
Dictionaries
Glossaries
Grammar Rules
Culturally specific code

After



Localization

- The process of tailoring a product to a specific target market.
 - **Translation** of messages
 - Adaptation to local preferences
 - Addition (or subtraction) of content or features

Localization is obvious

... but it isn't "internationalization"

- Localizability is internationalization.
 - Externalize text
 - Externalize presentation
 - Enable dynamic composition
 - Plan distribution of language content
 - "Plug-in" features

It's Part of the Process

Localization is part of the release process too.

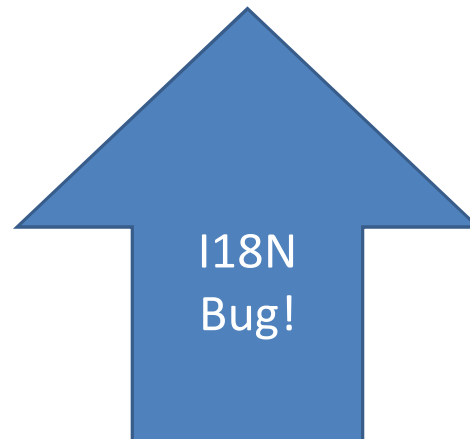
- Changes to the user interface cost the localization team time and money.
- (Changes to the product cost the documentation and QA folks too)
- May need to institute change control or a UI freeze

Two Basic Rules

- Don't Hard Code Stuff
 - Put strings in resources.
 - *Never hardcode your user interface.*
- Learn and use message formatting APIs.
 - *Do not concatenate strings.*
 - *Do not write your own replacement syntax.*

Your first program...

```
public static main(Object[] args) {  
    System.out.println("Hello, world!");  
    System.exit(0);  
}
```



Fixing “Hello World”

```
private String hello() {  
    ResourceBundle rb =  
        ResourceBundle.getBundle(“helloResources”);  
    return rb.getString(“hello”);  
}
```

“Hello, World!”
« Bonjour Monde! »
...

Getting personal...

```
private String hello(String name) {  
    return “Hello, “ + name + “!”;  
}
```

Hello, George!

You can find used flash drives in the warehouse deals.

```
String value =  
    "You can find used "  
    + product +  
    " in the warehouse deals.");
```

String Building

English Resource File:

```
msg.part.1: "You can find used "  
msg.part.2: " in the warehouse deals."
```



French Resource File:

```
msg.part.1 : "Vous pouvez trouver d'occasion "  
msg.part.2 : " dans les offres de destockage."
```

116

String Building

en-US You can find used flash drives in the warehouse deals.

fr-FR Vous pouvez trouver des clés usb d'occasions dans les offres de destockage.

Correct version:

fr-FR Vous pouvez trouver des clés usb d'occasions dans les offres de destockage.

117

Message Formatting

Internationalized APIs insert values into ***complete thought sentences***, formatting appropriately:

[] files out of [] were deleted.
 An error occurred at [] on [].
 Page [] of []
 Processing: []% complete.

Use Message Formatting APIs

Number and type the replacement variables!

“There were “ + numTables + “ tables on “ + date + “.”

There were {0} tables on {1}.

There were {0,number,integer} tables on {1,date,short}.
{1,date,short}に{0,number,integer}のテーブルがあった。

What's wrong here?

“You have {0,number,integer} {1} in your cart.”

More Issues With Text Composition

- There were **one errors** found.
- You have earned your **22th set** of bonus points.



More than one pattern string
is needed when the message
varies according to the value
being formatted?

What's wrong here?

```
public String makePlural(String noun) {  
    return noun + "s";  
}
```



Some Sheeps

Does this work?

There were no errors.
There was 1 error.
There were 2 errors.

0:There were no errors.
1:There was {0} error.
2:There were {0} errors.

Counting Things

0: не было ошибок
 1: была {0} ошибка
 2: были {0} ошибки
 5: были {0} ошибок
 10-20: были {0} ошибок

x1: была {0} ошибка
 x2-x4: были {0} ошибки
 x5-x0: были {0} ошибок

* Not just a Russian thing. Other languages have similar needs.

PluralFormat

```
String pattern = "{0,plural,
  =0    {Zero items}
  one   {An item}
  other {{0,number,integer} items}}";
```

```
MessageFormat fmt = new
  MessageFormat(pattern, myLocale);
```

What's wrong here?

1 000,00, 1 001,00, 1 002,00, 1 003,00,
 1 004,00, 1 005,00, 1 006,01, 1 007,01,
 1 008,01, 1 009,01, 1 010,01

Culture and Language Variance



amazon Prime
 Get Early Access



Images

- Beware your biases—even “good” ones.
- Avoid metaphors
Avoid cultural sensitivities
Avoid body parts
Replace as necessary

Meet your friends on our new social website for India



Text Swell

- 30% in length (alphabetic)
- May require larger font sizes (ideographics)
- But... **a rule of thumb**, not a "fact"

Measure your results with care.

925 IDEAS TO HELP YOU SAVE MONEY, GET OUT OF DEBT AND RETIRE.
925 IDEAS TO HELP YOU SAVE MONEY, GET OUT OF DEBT AND R...

will give you and your family a greater sense of who y

6. Get it at t

bookstore. Y

People, even parents, all need a little time to

themselves and to learn new things. Be afraid t

afraid not to have one. If your b

free, like genealogy, walking a

expensive to be enjoyed; a nob

Kindle Edition

Copyright 2014 by L. Danielle

you

be f

ebook may be reproduced, scanned or distributed in

any manner whatsoever without written permis

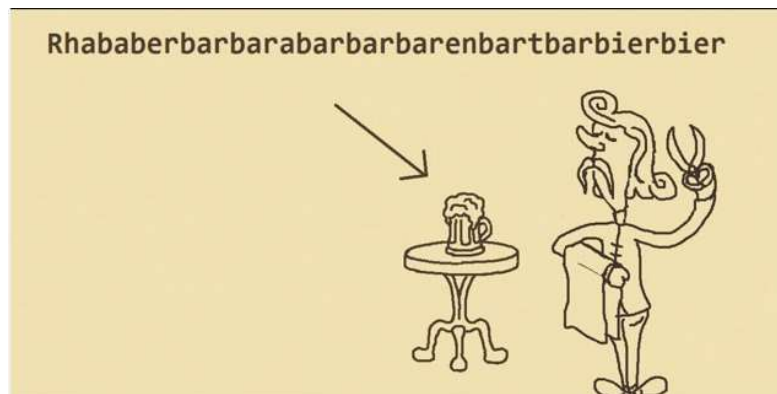
from the autor

That says "Goodreads"

読書の速さを測定中
0%

“But the German fit...”

- German does have (some) longer words.
- So if the German fits, we’re done right?



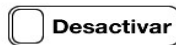
What is the longest language in the world?

クラウド | 端末

アイテム一覧 (20) ▼ 最新 ▼

Japanese

A Cautionary Tale



Positioning and Truncation Rules

Device name, "notification bullet", and
time all overlapping



サクライタカフミさんの 2番目の Kindle2015



Ask Yourself...

- What are the rules for expansion?
- Would you design an English dialog that follows those rules?
- Do you seek to avoid the expansion in English?
- What assumptions am I making about culture or language?
- Does the feature pertain to the language of the UX or of the content?

Recap: things to do.

- Use Resources.
- Use MessageFormat to build strings.
- Use MessageFormat to format insert values.
- Use complete thoughts.
- Handle plurals correctly.

Locale isn't everything...

price { 1134.0, "JPY" }



"We charged your card US\$1,134.00."

Non-Translatable Resources

- Some content should be externalized but not translated
 - Sometimes referred to as "**DNT**" for "do not translate"
- Externalize? Yes...
 - Segregate DNT material from translated material if possible (by using separate resource files or separate resource blocks within a file).
 - Developers can't always tell when something should or should not be DNT... and neither can translators (context is missing)



Style Guides

- Instructions to the translators on how to be consistent, your house style, the tone you want.

Naming and Branding



Consider doing Product Name Analysis.



Modifying your software to address a specific linguistic, cultural, or jurisdictional set of requirements.

CUSTOMIZATION

Externalization Review

- Externalize all resources—strings, numbers, fonts, images, settings
- Use international APIs for string assembly, plurals, genders, and formatting
- Prefer built in formats or skeletons
- Test drive with pseudo
- Customize features where it makes sense—using code as a resource

Internationalization

... is a fundamental architectural approach: it is how software is built.

- Design
- Enabling
- Externalization
- Customization
- Testing and Support
- Lifecycle



Q&A

Would you write the code for I18N on the whiteboard before you go?

```
#define UNICODE  
#import I18N.h
```