Software Globalization and Adding Languages on Computers and Mobile Devices

Craig Cummings

Unicode Technical Committee Vice-Chair

ANSI INCITS L2 Committee Chair

Unicode Emoji Subcommittee Animal/Science Category Manager

VMware Globalization Team Lead

Why Globalize Software?

IDC Worldwide Black Book: IT spend by language



Why Globalize Software? -- English vs. Localized Page Views

Ē



G11n, i18n, L10n and T9n



- Globalization (G11n): The complete process of making your application available in multiple languages consisting of 2 sub processes Localization and Internationalization.
- Internationalization (i18n): Architecting and coding for language/locale independence; allows localization for target audiences that vary in culture, region, or language.
- Localization (L10n): Adapting to specific languages/locales by translating text and adding locale-specific formatting to images and style sheets.
- Translation (T9n): Converting the meaning of text in one language into another.

https://blogs.adobe.com/globalization/2012/04/12/globalization-myth-series-myth-1-software-globalization-internationalization-localization-translation/



Localization Rocket Science – Machine Translation(MT)

Prevailing since late 2016

Neural

Dominating 10+ years until 2016

Statistic MT

Rule-based MT

java.io.InputStreamReader

(use the constructor that includes an explicit encoding)



new InputStreamReader(outputStream, explicitEncoding);



new InputStreamReader(outputStream);

Internationalization Eye Chart

- Encodings, I/O, character support
 - Regular expressions including Unicode regular expressions
 - Because [A-Za-z0-9] works only for ASCII
 - Internationalized Domain Names (IDNs)/Internationalized Resource Identifiers (IRIs)
 - Supplementary character support
- Locales (BCP-47 compliant)
 - Determination
 - Representation
 - Negotiation
 - Fallback
- Locale-sensitive APIs
 - Externalization
 - Resource files and messages

- Formatting
 - Dates, times, time zones and calendars
 - Numbers
 - Currencies
 - Honorifics
 - Addresses and phone numbers
- Linguistic (not binary) Sorting
- Wrapping/boundary analysis (words, lines, sentences, etc.)
- Keyboard support
- User Interface (including bidi), fonts, and layout support
- Search, normalization, folding, etc.
- Security
- ...and more

You Probably Know Unicode for...



Hoping you know Unicode for writing systems

Living scripts encoded in Unicode 11.0

Historical scripts encoded in Unicode 11.0 Historical scripts not yet encoded in Unicode

THE WORLD'S WRITING SYSTEMS

Atelier National de Recherche Typographique (ANRT), Nancy, France Institut Designfabor Gutenberg (IDG), Hochschule Mainz, Germany of Linuxiating, UC Berkeley, USA

				PRESSPICAL SCRIPTS FOR AN		anna an tha tar yar ar		Script Encoding Initiation	ve (SEI), Department of Ur	iguistics, UC Berkeley, US/				
PROTO-CUNEIFORM Million Million	EGYPTIAN HIEROCH		HIERATIC ANALY							ANATOLIAN MANTOLIAN MONTANA	VEDIC DECE DECE DECE DECE DECE	LINEAR S LINEAR S		
P ORACLE BODNE			DERONZE SCRIPT				LADIAN	R OLD ITALIC Scient 1		E TOCHARIAN INVESTIGATION		A LATIN TANANY	PALAEOHISPANIC	
	ĪTT				Ebidore Internet EPI-OLIECE ABERICA INTERNET					HEBREW MORAL AND	9	не селиние не селиние на селиние на на селиние на на на на на на на на на на на на на	SR MEROITIC CURSIVE	MEROITIC HIEROCITYPHS
		SATAVAHANA BOUTH AGA			K PALMYRENE MUSICAL ST	HATRAN MODULARIT MODULARIT	ANN NATIONAL MARKAN	F RUNIC Survey	SEAL MATA ANN MATA ANN MATA ANN MATA ANN		INSCRIPTIONAL PARTHIAN MODEL FAIT	CHORASMIAN Central Alia Central Alia		
			COPTIC	COTHIC NUMBER	KADAMBA Normala	SHANKHA (SHELL SCRIPT)			CEORGIAN Review I - Indep			FYU PYU Destenation fails		AVESTAN BEGORE ANT
CHALUKYA NOTWANA NOTWANA			PSALTER PAHLAVI	2 SVREAC WIGHT MART		AND THE LINE AND T				SOCOTIAN (ANNUMA) ANN	AAMIL BOOTE - MARK	VI VI MATANA		
	HIRACANA NATI ANA				J OLD TURKIC COMPARE AND	J OLD UYCHUR EXEMANA MAN	SHARADA Northe Kills	CLACOLITIC Receiped	CHOLA HOUTING AND HOUTING AND	4 HUNCARIAN EXECT	KHITAN LARGE SCHIPT MENTAN	KHITAN SMALL SCRIPT MAXAMA	SSS CHAM MINIMAN	R CYAILLIC E-DEREN No
	RANJANA BURUTH NAMA BALI - DOBAY	ST NESTERN CHAM BOUTINEAR AINS NOTINEAR AINS				ENCAL BUNGAN MINITAN			MEETEI MAVER MEETEI MAVER MEETEI MAVER			天 DIRCHEN EAT AND	OIVES ARURU BUTTER ARU	ESSU SHABA EAST SHABA LATANA LATANA
PHACE-PA PHACE-PA PHACE-PA	M THAT MATCHING				YEZIDI MICIOLA ANTI I INTE			Source A	SAL HENUTA MARKAN HENTAR		ACTIC ACTIC MARCA		BATAK BOUTHERATATIA BOUTHERATATIA	BUCINESE INTERNATION
(3	प्रत	Æ	가	m	лы	Lo	2	а	Ы	3	-	¥	£	ਕ

Did you know Unicode for this, too?



The Three (minimal) Requirements To Get Your Language On Computers and Mobile Devices

- 1. Fonts (already there for Latin)
- 2. Keyboards (layouts now a part of CLDR)
- 3. Unicode CLDR data (the short straw)

CLDR = Common Locale Data Repository

UN CLDR - UI	icode Common Locale Data Repository
Navigation	Unicode CLDR Project > CLDR Specifications >
Unicode CLDR Pro	
CLDR Releases/D	
CLDR Survey Tool	This document describes the minimal data needed for a new locale. There are two kinds of data:
CLDR Change Red	uests
CLDR Charts CLDR Process	1. Core XML Data - This is data that the CLDR committee needs from the proposer before a new locale is added. The proposer is expected to also get a Survey Tool account, and contribute towards the Minimal Data. 2. Minimal Data Commitment - Data that is expected to be provided for each locale. If it is not supplied in a timely fashion, the committee may remove the locale.
CLDR Specification	s (The parenthesis at the start of each line below has the approximate number of strings for each item.)
Information Hub fo	Linguists Core XML Data
Unicode Extension	for BCP
47	First, make sure you have correct language code according to Picking the Right Language Identifier. Then collect the following data. Consider using the Core Data Submission Form to submit this data.
Implementer's FAC	
ULI Subcommittee	Note to translators: If you are having difficulties or questions about the following data, please contact us. Post a follow-up to your existing bug, file a new bug, or reply to the mailing list.
Milestone Schedu	e 1. (04) Exemplar sets: main, auxiliary, index. [main/xxx.xml]
Q2/3 Tar	yets 2. (02) Orientation (bidi writing systems only) [main/xxx.xml]
	0/Data 3 (01) Plural rules [supplemental/plurals.xm]
	ration of For more information, see <u>cldr-spec/plural-rules</u> .
May 16 Start Sh	4. (or) Delaut content scipt and region (normally country with algest population using that language, and normal scipt or that). [supplemental supplemental metadata.xiii]
May 23 Start Gen	5. (14) Verify the country data (i.e. which territories in which the language is spoken chough to cleate a locate) [supplemental/sup
Submissi	
	Start 7. (N) Collation rules [non-Survey Tool]
	ubmission For details, see <u>cldr-spec/collation-guidelines</u> .
Jul 11 End Su	 The result will be a file like: <u>common/collation/ar.xml</u> or <u>common/collation/da.xml</u>. Note that the "search" collators (which tend to be large) are not needed initially.
	<i>fetting</i>
Start Re	solution Recommended Core Data
	The following are not required, but are strongly recommended:
	In (04) Exemplar Set. punctuation. [main/xxx.xim]
	changes o For more information, see <u>cldr-spec/plural-rules</u> .
	aner 3. *(N) Romanization table (non-Latin writing systems only) [spreadsheet, we'll translate into transforms/xxx-en.xml]
Sep 12 Alpha —	- Inal Data o If a spreadsheet, for each letter (or sequence) in the exemplars, what is the corresponding Latin letter (or sequence).
	idate o More sophisticated users can do a better job, supplying a file of rules like transforms/Arabic-Latin-BGN.xml.
	Ing ICU Minimal Data Commitment
	after a la l
	td, data, This data is to be entered using the Survey Tool except as noted.
	allowed 1. (44+) 4 main Date/Time formats, 12 long&abbreviated, format&stand-alone month-names, 7 long&abbreviated day-names, 2 long day periods.
	- Final 2. (01) Name of the language in the language.
Can	idate 3. (N) For any country locales, name of the country in the language, name/symbol for that country's currency. Must be at least one, for the default content locale.
	or data 4. (02) Datetime pattern, intervalFormatFallback
	allowed 5. (05) (for Latn) decimal and grouping separators; decimal, currency, percent formats after 6. (N) Names of countries (territories) with that language as official.
	arts, spec 7. (M) Names of exemplarCities in multizone countries with that language as official
change	allowed 8. (05) Timezone patterns [http://cldr.unicode.org/translation/timezones]
	release 9. (02) localePattern/Separator [http://cldr.unicode.org/translation/localepattern]
	date) 10. (03) key names 11. (14) long/short unit names (time intervals)
Oct 15 Rel	ase 11. (14) long/short unit names (time intervals)

The First Seven Steps to 'Core XML Data'

- 1. Exemplar characters (dictionaries, grammar, references)
- 2. Left-to-right, Right-to-left, other directions
- 3. Plural rules (one duck, two ducks, etc.)
- 4. Script/region data (ISO 15924, 3166, 639)
- 5. More region data
- 6. Casing (titlecase for Latin-based languages)
- 7. Collation (fancy word for sorting; order with exemplars)

Getting To A Better CLDR Interface

← → C ▲ Not Secure | st.unicode.org/cldr-apps/v#/lkt/Alphabetic_Information/

Survey Tool 35 Read-Only 🔅 - Coverage: Basic Instructions -Survey Tool is now closed. Lakota / Core Data / Alphabetic Information > ← Previous Next → Toggle Sidebar 🚍 This locale, Lakota, supplies the default content for Lakota (United States). If there are that all the changes that you make here are appropriate for Lakota (United States). If there are please try to pick the one that would work for the most other sublocales. Code English A Winning **Characters In Use** Main Letters [abcdefghijklmnopgrstuvwxyz] a á {aŋ} b č {čh} {č'} e é g ğ h ň i í {iŋ} k {kh} {kň} {k'} l m n ŋ o ó p {ph} {pň} {p'} s š t {th} {tň} {t'} u ú {uŋ} w y z ž ' 🙀 Others: [áàăâââãāæçéèĕêëēíìĭîïīñóòŏôöøōœú V c d f {ĥ'} j q r {s'} {š'} v x 🖄 auxiliary ùŭûüūÿl Others: index [A B C D E F G H I J K L M N O P Q R S T U V W X Y Z] V A B Č E G Ğ H H I K L M N Ŋ O P S Š T U W Y Z Ž 🛠 1 Others: numbers [-, . % % + 0123456789][\-,.%‰+0123456789] 🛠 Others: [\----,;\:!?....'''""()\[\]§@*/\&#†‡'"] ~ \----,;\:!?." " " () \[\] @ * / \& # 🛠 punctuation **Quotation Marks** End 11 X " 🏠 Start " × " 🛠 1 🛠 embedded-End 1 × embedded-Start 1 1 🛠 Yes And No

Getting to Exemplars: Step 1A - Main Letters



a	a:	b	ch	ch'	chw	
father	p <u>a</u> lm	bear	church	(ch with catch)	inchworm	
whila'	whina:'	bo:se	mindich	which'ich'	chwich	
(my hand)	(my eye)	(cat)	(bobcat)	(my elbow)	(firewood)	
d	dz	e	e:	g	gy	
deer	adze	met	men	geese	figure	
<u>dinday</u>	didzit	whixe'	n <u>e:</u> s	niwhgit	digyun	
(bullet)	(short)	(my foot)	(long)	(I'm afraid)	(here)	
h	i	j	k	ky	k'	
hen	h <u>i</u> t	jar	<u>k</u> eep	than <u>k y</u> ou	(k with catch)	
xontah	mis	je:nis	king	kya'	k'ina'	
(house)	(riverbank	-	(stick) (dress)		(Yurok)	
ky'	I	ł	m	n	ng	
(ky with cate	ch) let	(breathy I)	mill	now	ri <u>ng</u>	
ky'oh	lah	<u>ł</u> a'	<u>m</u> ili <u>m</u> il	<u>n</u> u <u>n</u> dil	whing	
(porcupine)	(seaweed)	(one)	(flute)	(snow)	(song)	
0	0:	q	q'	S	sh	
tote	c <u>o</u> ne	(guttural k)	(q with catch)		ru <u>sh</u>	
dingq'och	to:-nehwa:n	go	whig'os	sa:ts'	no <u>sh</u> t'ah	
(sour) it)	(obsidian)	(worm)	(my throat)	(bear)	(I don't believe	
t	ť	tł'	ts	ts'	u	
tea	(t with catch)	(tł with catch)	ca <u>ts</u>	(ts with catch)	r <u>u</u> n	
to	<u>t'</u> e'	<u>tł'</u> oh	<u>ts</u> e	ts'ilting'	łix <u>u</u> n	
(water)	(blanket)	(grass)	(stone)	(weapon, rifle)	(sweet)	
w	wh	x	XW	у	6	
word	whirred	(guttural h)	(guttural wh)	yes	(catch)	

Getting to Exemplars: Step 1B - Index Letters



Pamela Munro and Catherine Willmond

CHICKASAW-ENGLISH VOCABULARY

aachompa' store, town (1, 10) aahámbi (irregular HNgr. of aabi) (8) aahashtahli' window (10RS, 18) aahilha' dance hall (18) aahobachi to copy (something) from (a thing) (A,Noun,Noun) (16) aaholloppi' graveyard, cemetery (18) aaholhponi' pot (9RS), kitchen (10) aahonkopa to steal (something) from (an institution) (A.Noun.Noun) (16) aahopooni to cook (something) in, cook (something) on (A.Noun,Noun) (9RS) Aaiklanna' part of Nittak Hollo' Aaiklanna' aaiksaa to be made from (A.Noun) (13RS) aaimpa' restaurant (10) aaimpa' abooha dining room (10)

áa to go along, to be going along (A) (1, 20) alhínchi to be enacted, made true (Noun) (17RS) ashaka', ashka' behind, rear, back (SP); in back of (loc. n.) (SP) (10)

bala' bean (2)

sinti' bala' / bala' sinti' snake bean balaafokha' / balaafka' pants (14) banna to want (S,Noun) (9) basha to be cut, to be operated on (S) (6) itti' basha' abooha frame house bashafa to be cut (of a long subj.) (S) (16) bashaffi to cut (a long obj., usually with a knife) (A,S) (5) bashli to operate on, to cut (A,S) (5) bashpo knife (1, 7) baafa to stab (A.S) (9) bila to melt, to be melted (S) (1, 4) biliili to melt (something) (A.S) (16) billi'vacha billi'va forever and ever (15RS) bini'cha loshka sit down and tell lies! (an invitation to gossip) (3) biniili to sit down, get into a sitting position (sg. subj.) (A) (12RS, 19) bínni'li to sit, be seated, be in a sitting position (A); to sit in, to be located (sitting) in (a place) (A,Noun) (sg. subj.) (8, 10)

aaissachi to leave (A,S) (14RS) aaittanaa' church (7RS, 10) aaittapiha' group (20RS) aai'pa' table (10) aai'pa' holiito'pa' holy table (12RS) aalhakoffichi to save from (A,S) aalhpi'sa to be right, be correct; to do the right thing (A) (6) aalhponi' kitchen (10) aaminti to come from (A.Noun)

aaipita' in issoba aaipita' piini'

aanosi' bedroom (10) aaomhiniili' chair (12) aapihlichika' kingdom (DP) (15RS) aaponta to borrow (something) from (an institution) (A,Noun,Noun) (16) aayimmi to believe in (something) (S:Noun) (17RS)

áshwa to stay, be there (A); to be located in (a place) (A,Noun) (dl, subi.) (12) imáshwa to have (relatives) (dl. human obj.)

binohli to sit down, get into a sitting position (pl. subj.) (A) (19) binohmáa to sit, be seated, be in a sitting position (A); to sit in, be located (sitting) in (a place) (A,Noun) (tpl. subi.) (10) bíyyi'ka all the time, always (aux.) (1, 8); (aux. used with ±a'hi "can" verbs) (5); it's all, it was all (7RS); truly (17RS); several, many (used as a stative verb following a noun, with focus endings) (18) bo'li to hit (more than once); to beat up; to pound; to hit (pl. subi.) (A.S) (1, 13) bo'wa to be hammered (Noun) (18) bohli to lay down, put down (sg. long obj.) (A,S) (1, 13) ishtombohli to accuse, blame; to accuse about, blame for [isht ombohli] ombohli to lay (something) down on, put (something) down on

chaffa one (3RS); to be one in number (A) (12); part of nittak chaffaka one day

The Still Easy 'Minimal' Remainder

- Just Over 50 additional easy-to-find translations
 - (38) The brunt are some 12 month names and 7 day names
 - Translations often in dictionaries, apps, or other good references
 - (1) Name of the language in that language
 - (7) "Year", "Month", "Week", "Day", "Hour", "Minute", "Second"
 - (6) Next slide
- With caveats
 - Some translations may require a bit more 'wordsmithing'
 - Abbreviations may not be a concept in a given language

The Last Six Strings To Minimal

- "Standard Time"
- "Daylight Time"
- "Gregorian Calendar"
- "Standard Sort Order"
- "Western Digits"
- <date> "at" <time>

Months of the Year

← → C ▲ Not Secure | st.unicode.org/cldr-apps/v#/lkt/Gregorian/

Survey Tool 35 Read-Only 🔅 🗸 Coverage: Basic Instructions 🗸

Survey Tool is now closed.

Lakota / Date & Time / Gregorian

← Previous Next → Toggle Sidebar =

Months - Wide - F	Formatting			
Jan	January	V	Wiótheȟika Wí 🛠	M01
Feb	February	V	Thiyóȟeyuŋka Wí 🛠	M02
Mar	March	V	Ištáwičhayazaŋ Wí 😒	M03
Apr	April	V	Pňežítňo Wí 🖄	M04
May	May	V	Čhaŋwápetȟo Wí 🛠	M05
Jun	June	V	Wípazukȟa-wašté Wí 😒	M06
Jul	July	V	Čhaŋpȟásapa Wí 📩	M07
Aug	August	V	Wasútȟuŋ Wí 🛠	M08
Sep	September	V	Čhaŋwápeği Wí 🛠	M09
Oct	October	V	Čhaŋwápe-kasná Wí 🛠	M10
Nov	November	V	Waníyetu Wí 🛠	M11
Dec	December	~	Tȟahékapšuŋ Wí 🛠	M12

Days of the Week

← → C ▲ Not Secure st.unicode.org/cldr-apps/v#/lkt/Gregorian/

Survey Tool 35 Read-Only 🔅 - Coverage: Basic Instructions -

Survey Tool is now closed.

Lakota / Date & Time / Gregorian

>

← Previous Next → Toggle Sidebar 重 Days - Wide - Formatting Sun Aŋpétuwakȟaŋ 😭 sun ~ Sunday Mon Aŋpétuwaŋži 😭 mon V Monday Aŋpétunuŋpa 📩 Tue tue V Tuesday Wed Aŋpétuyamni 🖄 wed V Wednesday Thu Aŋpétutopa 🖄 thu Thursday ~ Fri fri Aŋpétuzaptaŋ 😭 ~ Friday Sat sat Owáŋgyužažapi 🙀 V Saturday

What Success Looks Like On A Device

..... $\bullet \bullet \circ \circ < >$ Language & Region Q Search Language & Region preferences control the language you see in menus and dialogs, and the formats of dates, times, and currencies. Preferred languages: \Diamond Region: **United States** GWУ Cherokee — Primary $\hat{\mathbf{C}}$ First day of week: Sunday Lakhól'iyapi Lakota \bigcirc Calendar: Gregorian $\Delta \sigma_{P} \cup \nabla c$ Inuktitut 24-Hour Time Time format: Diné bizaad \Diamond List sort order: Universal Navajo Guarani OOVLOSO, OZAWO 5, 2019 7:08:09 HOG PST Guarani 1/5/19 7:08 UOJ 1,234.56 \$4,567.89 + | -

(?)

A Success Story: Cherokee (in their own words)



https://youtu.be/EEEu8ufwW08?t=1498

How To Get Started

- File an new ticket with the 'Core XML Data' here:
 - <u>http://unicode.org/cldr/trac</u>
 - For reference, Osage, Muscogee/Creek, Chickasaw examples here:
 - <u>https://unicode.org/cldr/trac/ticket/10721</u>
 - <u>https://unicode.org/cldr/trac/ticket/11424</u>
 - <u>https://unicode.org/cldr/trac/ticket/11983</u>
- Ping me with ticket #
 - craig@unicode.org



Saving the World with Computing

Kathy Yelick EECS Professor, U.C. Berkeley

Associate Laboratory Director for Computing Sciences and NERSC Director, Lawrence Berkeley National Laboratory



1:15/40:18

Q&A

craig@unicode.org

ccummings@vmware.com

Postscript

- CLDR Release Timing
 - Spring and fall of each year; fall is a bigger release
 - Calendar on CLDR home page
- Unencoded scripts require more work (e.g., Mi'kmaq)
 - Happy to guide towards a Unicode character encoding proposal
- Patience is required
 - Data may only take an hour to prepare, but keyboards, fonts, and system integrations will take longer.
- Full localizations of user interfaces, online help, and other documentation are a different discussion
 - Happy to have that discussion, too.