Deploying the Common Locale Data Repository (CLDR)

Steven R. Loomis
IBM

Mark Davis
Google

Implementation: OS/Library/Application

- Process, interchange, display all text using modern standards (such as Unicode)
- Provide the user with a localized experience that matches their own cultural and linguistic expectations

Localization: A moving target

- Often difficult to determine the "best" translation.
- Increasingly sophisticated platforms
- Emerging markets
- Constantly changing user expectations, geopolitical and linguistic landscape

The Need for Common Data

How do you spell the 12th month of the year in Catalan:

Desembre, decembre, or desembre?

- Operating systems and application software can have much variation in locale data.
- It is time consuming to keep this data up to date.
- It is difficult to get complete agreement on correctness.

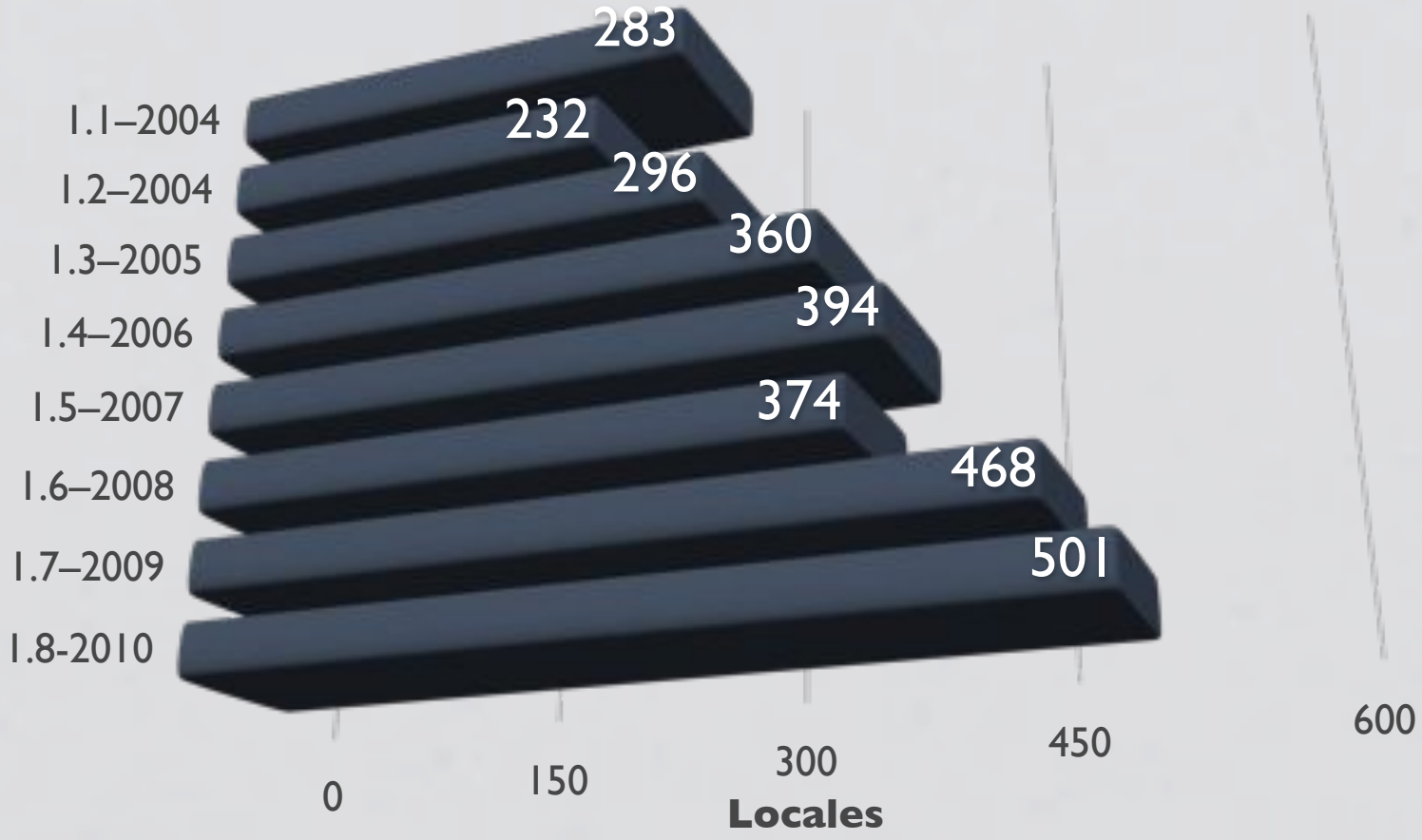
General Scope

- Dates/time formats
- Number/currency formats
- Measurement Units
- Collation Specification: Sorting, Searching, Matching
- Names for Languages, Territories, Scripts, Timezones, Currencies,...
- Characters used by a language
- ...

Common Locale Data Repository (CLDR) History

- IBM Cultural Information Repository: 1990s
  - Java 1.1 (via Taligent): 1997
  - International Components for Unicode (ICU): 1999
  - Universal Locales for Linux: 2001
- CLDR 1.0: 2003 (part of the OpenI18N LADE workgroup)
- CLDR 1.1: 2004 (sponsored by the Unicode Consortium)
- Many subsequent versions…

Common Locale Data Repository History



Locale Data Markup Language (LDML)

- XML Interchange Format
- Transformed into forms optimized for use by ICU, POSIX, OpenOffice, dojo, others...
- Unicode Technical Standard #35

```
<dayWidth type="wide">
  <day type="sun">Sonntag</day>
  <day type="mon">Montag</day>
  <day type="tue">Dienstag</day>
  <day type="wed">Mittwoch</day>
  <day type="thu">Donnerstag</day>
  <day type="fri">Freitag</day>
  <day type="sat">Samstag</day>
</dayWidth>
```

Who uses CLDR?

CLDR Vetting Process

- Data Submission Phase:
  Data is entered via Survey Tool or Bug Report form
- Vetting Phase:
  Users vote for their preferred forms, make use of e-mail and forums to resolve conflicts.
- Resolution Phase:
  CLDR Technical Committee verifies data and corrects remaining conflicts.
- Final Candidate, Release:
  Final data is tested and then released. Process starts over.

CLDR Vetting

Swedish: “Bamileke”

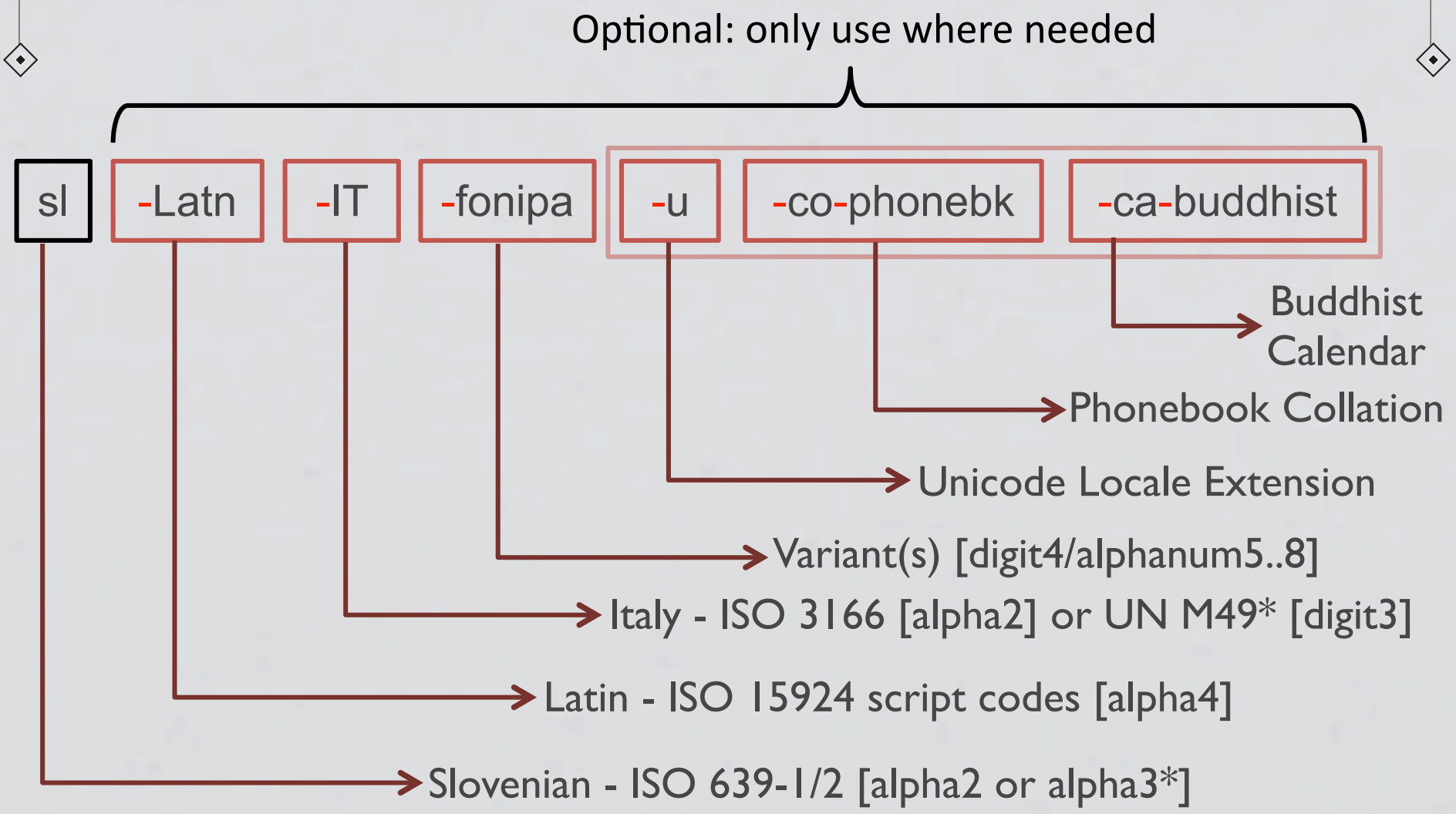
St.	Code	English	Proposed 1.8	Other
✓	bai	Bamileke Language	bamilekespråk ☆	☑ bamilekéspråk bamilekiskt språk

Bamilekespråk or bamilekéspråk or bamilekiskt språk?

CLDR Conflict Resolution

- Compare data from different platforms and experts.
- Different user levels for different types of vetters:
  - Sponsored, known experts: More weight given in conflict, higher confidence of result.
  - Guest: Anyone may apply.
- Data marked with different confidence levels according to type of conflict and type of vetter

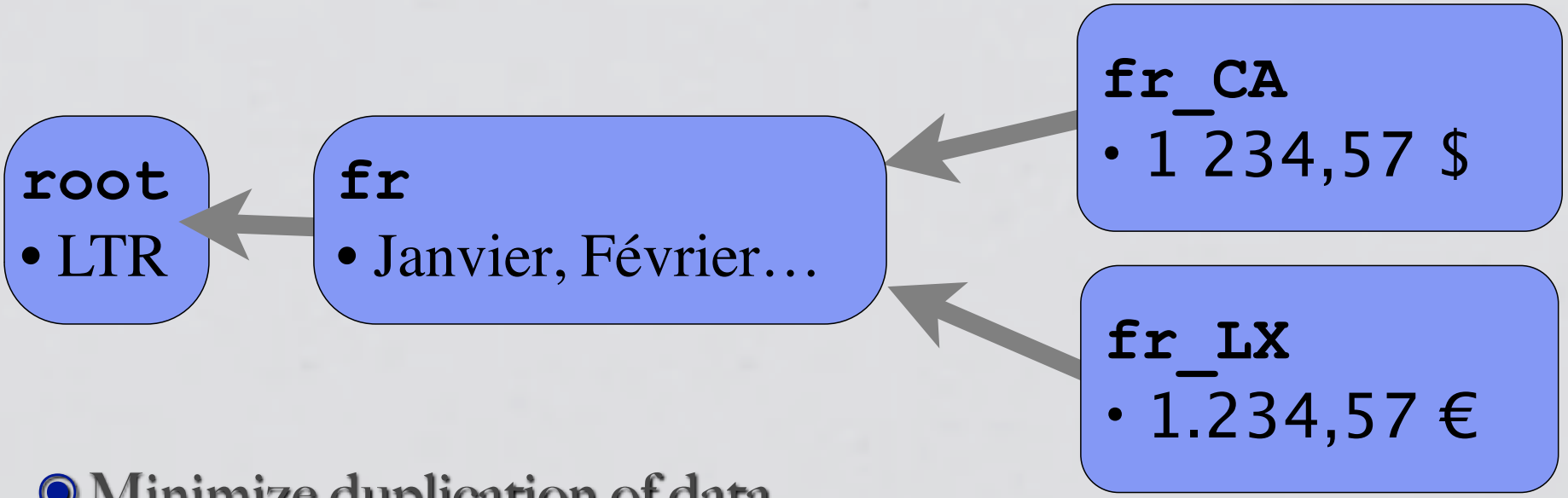
Anatomy of a Unicode Locale ID



Unicode Locale IDs

- **`de-AT-u-co-phonebk`**
  - German in Austria, Collation: Phonebook Style
- IETF BCP 47+
  - Key: 2 characters (co)
  - Type: 3-7 characters (phonebk)
- The <ldmlBCP47> element maps long and short key and type names into IETF BCP 47 format.

Locale Inheritance



- Minimize duplication of data
- Decrease maintenance cost
- Final fallback: "root" locale

Draft Status

- There are four draft values, indicating where this data field is in the vetting process.
- Implementations may choose which draft values they will accept for different types of data and locales.
  - approved: A supermajority of votes.
  - contributed: A bare majority of votes.
  - provisional: A majority of votes, but no quorum.
  - unconfirmed: Insufficient votes.

Alternate Values

- An element can have alternative forms.
  - <language type="az">Azerbaijani</language>
  - <language type="az" alt="short">Azeri</language>
- alt= is also used in the vetting process.
  - <... alt="proposed-ZZZ" draft="unconfirmed" > ...

<localeDisplayNames>

- Translated display names and formatting patterns for languages, territories, scripts, variants, keywords, keyword types, measurement systems.

code	English	German	…
de	German	Deutsch	…
fr	French	Französisch	…
nl_BE	Flemish	Flämisch	…
…	…	…	…

<characters>
<exemplarCharacters>

- Main: Letters used in the language.
  - German: a-zßäöü
- Auxiliary: Letters used in foreign and technical words.
  - German: à-âå-ïñ-ôø-ûÿāăēěīĭōŏœūŭ

<characters>
<exemplarCharacters>

- Punctuation
  - English:
    !"#%&'()*,-./:;?@[\]{}§°¶-–—‘’“”†‡…′″〈〉
- Index: Head letters that appear in an index.
  - Slovak:
    A Ä B C Č D Ď E F G H I J K L Ľ M N
    O Ô P Q R S Š T Ť U V W X Y Z Ž

<delimiters>

English	“quotation”	‘alternate’
German	„quotation“	‚alternate‘
Japanese	「quotation」	『alternate』

<dates>

- Gregorian, Buddhist, Islamic, Japanese...
- Format/Parse of dates & times
  - Eras, Years, Months ... Timezones...
- Relative day/time translations (“Yesterday”, “Tomorrow”, ...)

<dates>
Fixed and Flexible Formats

Fixed

Full	Thursday, October 14, 2010
Long	October 14, 2010
Medium	Oct 14, 2010
Short	10/14/10

Flexible

	English	Japanese
Year + Abbr. Month	Oct 2010	2010年10月
Abbr. Month + Day + Weekday	Fri, Oct 15	10月15日(金)

Date Formatting in CLDR

- Two contexts, format vs. standalone.
- For each context: wide, abbreviated, or narrow.

	Wide	Abbreviated	Narrow
Format	Μαρτίου	Μαρ	Μ
Standalone	Μάρτιος	Μαρ	Μ

Time Zone Display Names

- Based on Olson time zone database
  - Also have stable short IDs based on UN/LOCODE
    "`brfen`" = "America/Noronha" or "Brazil/DeNoronha"
- Metazones: group of equivalent zones
- Leverage Country names where possible

Time Zone Examples (French)

Generic Non-Location	HEC
	Heure de l'Europe centrale
Specific Non-Location	HAEC
	heure avancée d'Europe centrale
RFC 822	+0200
Localized GMT	UTC+02:00
Generic Location	(France)

<numbers>

- Format/Parse
  - Decimal, Scientific, Currency, Percentages, Custom
  - Example: 1234.567 (binary) → 1.234,567 (French)
- Includes localized decimal, grouping separators, currency symbols, etc.

<numbers> Currencies

	English	Serbian
USD	$35.72 US dollar / US dollars 1 US dollar 2 US dollars 5 US dollars	35.72 US$ амерички долар / долара 1 амерички долар 2 америчка долара 5 америчких долара
EUR	€35.72 euro / euros 1 euro 2 euros 5 euros	35.72 € евро / евра 1 евро 2 евра 5 евра

<units>

- Currently: Year, Month, Week, Day, Hour, Minute, Second

English	Czech
1 hour	1 hodina
1 hr	1 hod.
2 hours	2 hodiny
2 hrs	2 hod.
5 hours	5 hodin
5 hrs	5 hod.

<listPatterns>

English	Japanese
John and Mary	鈴木、田中
John, Mary, and Ted	鈴木、田中、渡辺

<posix>

- **Yes** and **No** strings and expressions used for compatibility with POSIX
- Used by POSIX locale generation tools to generate the **`LC_MESSAGES`** section correctly.

Rule Based Number Format (RBNF)

#	12,345
English	twelve thousand three hundred forty-five
German	zwölftausenddreihundertfünfundvierzig
Italian	dodicimilatrecentoquarantacinque

Text Segments

- UAX #29 and locale-specific tailorings

User Character Breaks	|I| |l|i|k|e| |a|p|p|l|e|s|.| |(|D|o| |y|o|u|?|)|
Word Breaks	|I| |like| |apples|.| |(|Do| |you?|)|
Line Breaks	I |like |apples. |(Do |you?)
Sentence Breaks	|I like apples. |(Do you?)|

Transforms

キャンパス	**kyanpasu**
Αλφαβητικός Κατάλογος	**Alphabētikós Katálogos**
биологическом	**biologichyeskom**

Supplemental Data I

- Likely Subtags: **hi⇔hi-Deva-IN**
- Territory↔Language↔Script:
  - Côte d'Ivoire: 49% French, 11% Baolé, ...
  - French: 54,449,130 in France, 10,102,379 in Côte d'Ivoire, ...
  - Serbian ⇔ Cyrillic Script, Latin Script, ...
- Territory → Currency
  Botswana: South African Rand [**ZAR**] from 1961-1976,
  Botswanan Pula [**BWP**] from 1976-present, ...
- Territory Containment (UN M.49):
  Central America [**013**] = Belize + Costa Rica + ...

Supplemental Data II

- Zone → Tzid: Windows Timezone IDs to Olson
- Language Plural Rules:
  Arabic: “zero”, “one”, “two”, “few” (3-10), “many” (11-99), ...
- Character Fallback Substitutions:
  <U+20B9>  (Indian Rupee Sign) → “Rs.”
- Aliases: **cmn** (Mandarin) → **zh** (Chinese)

<collations>

- Unicode Collation Algorithm (UTS #10)
- Tailoring of DUCET for languages
- New in CLDR 1.9 – Root tailoring
  - Spaces, Punctuation, Symbols, Currencies, Numbers in groups.
  - U+FFFE lowest weight, U+FFFF highest.
  - Only spaces and punctuation ignorable.

Collation example

German	Swedish
01: Åkersberga	02: Alingsås
02: Alingsås	04: Oskarshamn
03: Äppelbo	07: Utting
04: Oskarshamn	06: Üttfeld
05: Östersund	08: Zwickau
06: Üttfeld	**01: Åkersberga**
07: Utting	**03: Äppelbo**
08: Zwickau	**05: Östersund**

Questions?

- Unicode CLDR web site:
  http://unicode.org/cldr
- LDML specification:
  http://unicode.org/reports/tr35
- These Slides:
  http://macchiato.com
- srloomis@us.ibm.com
- mark@macchiato.com