

What's new in CLDR 2.0

Steven R. Loomis
**IBM San José Globalization Center of
Competency**

Unicode: Global Foundation



- System of standards
- Encodes all modern languages
- Free flow and interchange of textual data

Challenge: Local Needs



- Process, interchange, display all text using modern standards including 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

- Different 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.

Unicode CLDR (Common Locale Data Repository)

- Dates
 - *2010年10月*
- Numbers
 - *€12,35*
- Units & Relative
 - *3 hours*
 - *4 hours ago*
- Characters
 - *ā ā̃ ā̄ ā̅ ā̆ ...*

...

- Names
 - for: Languages,
Regions, Scripts,
Time zones,
Currencies...
- Sorting, Searching,
Matching
- Language matching

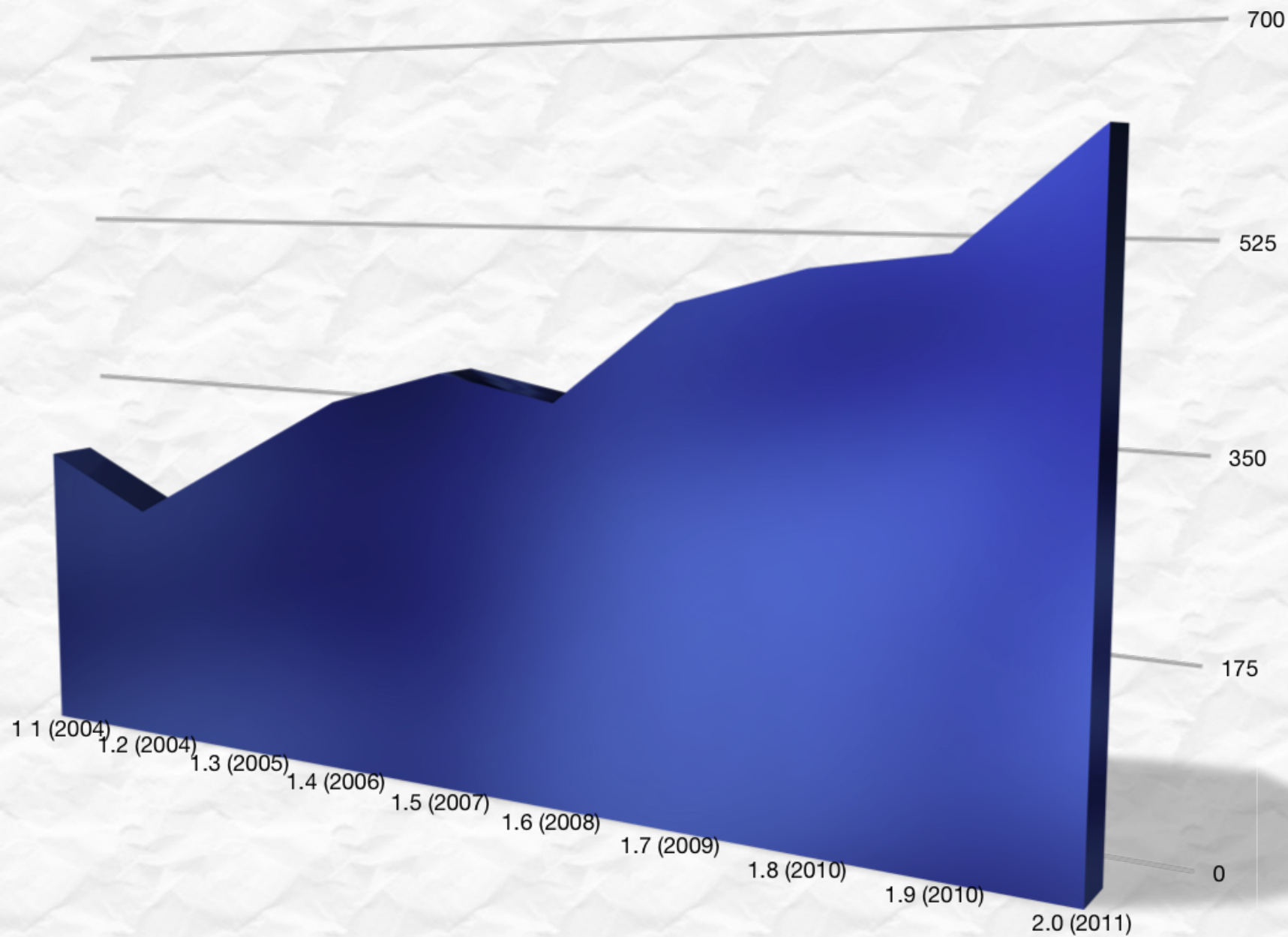
...

Who uses CLDR?



Many companies, organizations, and individuals contribute to CLDR data and structure

History of CLDR Locales



Locale Data Markup Language

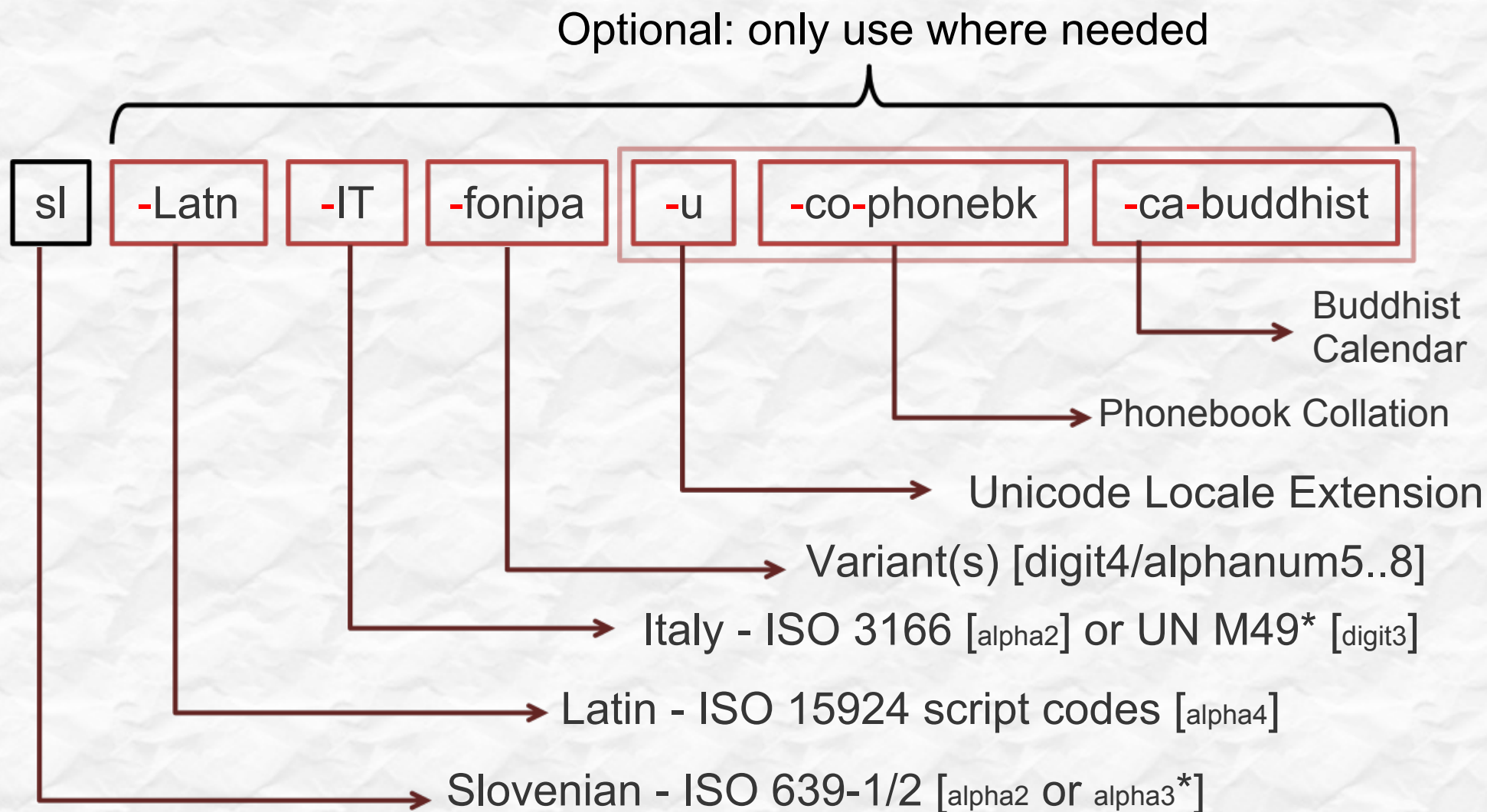
XML Interchange Format

```
<dayWidth type="wide">  
  <day type="sun">Sonntag</day>  
  <day type="mon">Montag</day>  
  <day type="tue">Dienstag</day>  
  <day type="wed">Mittwoch</day>...
```

In products, use optimized format.

ICU, POSIX, OpenOffice, dojo, others...

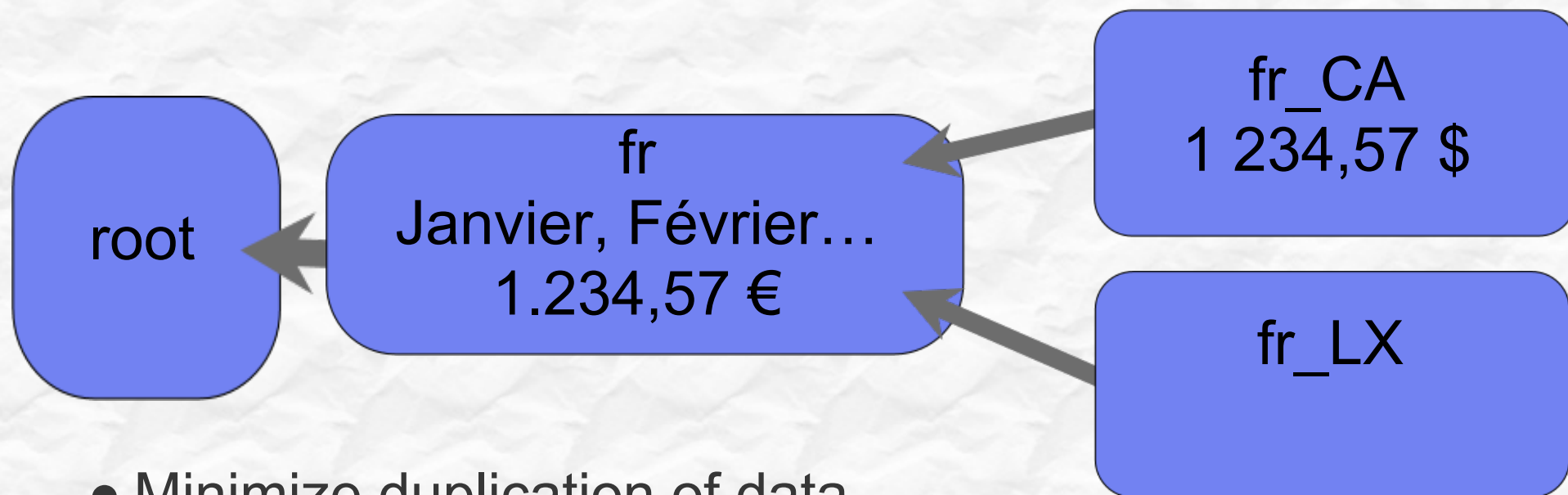
Anatomy of a Unicode Locale ID



Unicode Locale/Language ID

- UTS #35 *Unicode Locale Data Markup Language* (LDML)
- Based on [BCP 47](#) + [RFC 6067](#) + [language-subtag-registry](#)
.
- Some restrictions & extensions
 - Both '_' and '-' as separators
 - No extlang, no irregular (grandfathered) tags
 - Uses “zh” for compatibility, not “cmn”, etc.
 - Private use codes defined
 - “ZZ” for Unknown Region

Locale Inheritance



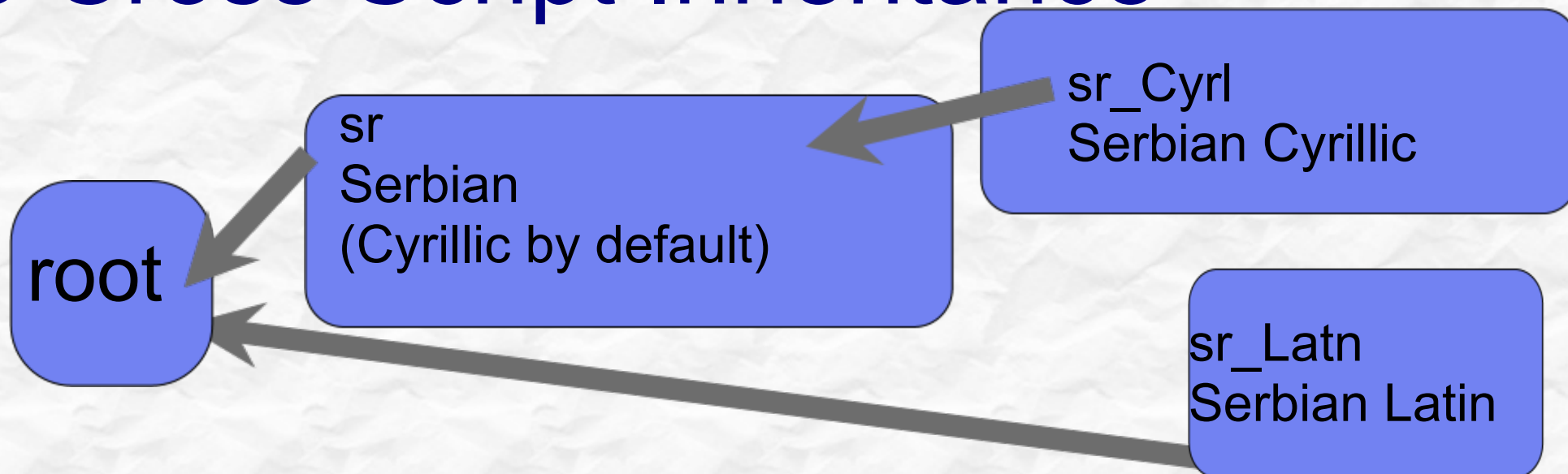
- Minimize duplication of data
- Decrease maintenance cost
- Final fallback: “root” locale

Specialized Inheritance



- Further reduce cost/effort
- Coordinate translations in related sublocales
- Identify places to save translation costs for applications

No Cross Script Inheritance



- sr_Latn does NOT inherit from sr !
- Avoid “ransom note effect” if some are/aren't translated.
- Similarly, zh_Hant (Traditional Chinese) doesn't inherit from zh (Simplified Chinese)

Locale Display Names

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

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

Exemplar Characters

Main: Letters used in the language

ä ö ß ü v-z

Auxiliary: Foreign and technical letters

à â ã ä å æ ç è é ê ë ē ... œ ú û ü ū Ÿ

Index: "Head" letters

À Ä Å Ć Č Ď Ě Ğ ... X Y Z Ž

Punctuation

- _ , ; : ! ? ‘ ’ ‚ / “ ” ’ () [] / @ &
...

Delimiters

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

Fixed/Flexible Date 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)

Time Zone Formatting

Generic NL - Short	HEC
Generic NL - Long	Heure de l'Europe centrale
Specific NL - Short	HAEC
Specific NL - Long	Heure avancée d'Europe centrale
RFC 822	+0200
Localized GMT	UTC+02:00
Generic Location	France

Unit Formatting

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.

- Year, Month, Week, Day, Hour, Minute, Second
- With plural support

Relative times

English
Yesterday
Tomorrow
3 days ago*
In 3 days*
...

*** New in CLDR 2.0, with plural support**

Rule Based Number Formatting

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

- Many improvements to the data

Currencies

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

List Patterns

English	Japanese
A and B	A、B
A, B, and C	A、B、C

Text Segments

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

Transforms

キャンパス	kyanpasu
Αλφαβητικός Κατάλογος	Alphabētikós Katálogos
биологическом	biologichyeskom

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

Collation (Sorting/Matching)

- Unicode Collation Algorithm (UTS #10)
- Tailoring (Customizing) for languages
- Root tailoring
 - Rearrange groups:
 - Spaces, Punctuation, Symbols, Currencies, Numbers, Latin, Cyrillic, Greek, ... CJK
 - U+FFFE lowest weight, U+FFFF highest.

Collation New Features

- Search Collator
 - Korean, Arabic, Hebrew (but located in Root)
 - Assigns primary weights to make searching easier (i.e. consider several different ALEF as equivalent)
- “Import”
 - Simplify maintenance
 - Example: Many European Languages will import “European Ordering Rules”

CLDR Process

Data Submission

English	Swedish
bai Bamileke Language	Bamil



Vetting

St.	Code	English	Proposed 1.8	Other
✓	bai	Bamileke Language	<input type="radio"/> bamilekespråk ☆	<input type="radio"/> <input checked="" type="checkbox"/> bamilekéspråk <input type="radio"/> bamilekiskt språk



Resolution / Verification (Technical Committee)



Final Testing / Release



What's Ahead?

- CLDR v21* – December, 2011
- CLDR v22* – June, 2012
- Structural changes to support new types of data (while keeping compatibility for existing users)
- Continual improvements to the voting process and policies
- Speed and reliability improvements to the Survey Tool

(*Note: CLDR 21 follows CLDR 2.0 - change in version numbering scheme.)

Questions?

CLDR	<u>http://unicode.org/cldr</u>
LDML	<u>http://unicode.org/reports/tr35</u>
Author	<u>srloomis@us.ibm.com</u>
Thanks	<u>Mark Davis and others on the CLDR-TC for comments and content.</u>