Unicode Transliteration

Mark Davis
President & Co-founder, Unicode Consortium
(Lead Int'l Architect, Google)

Unicode Consortium

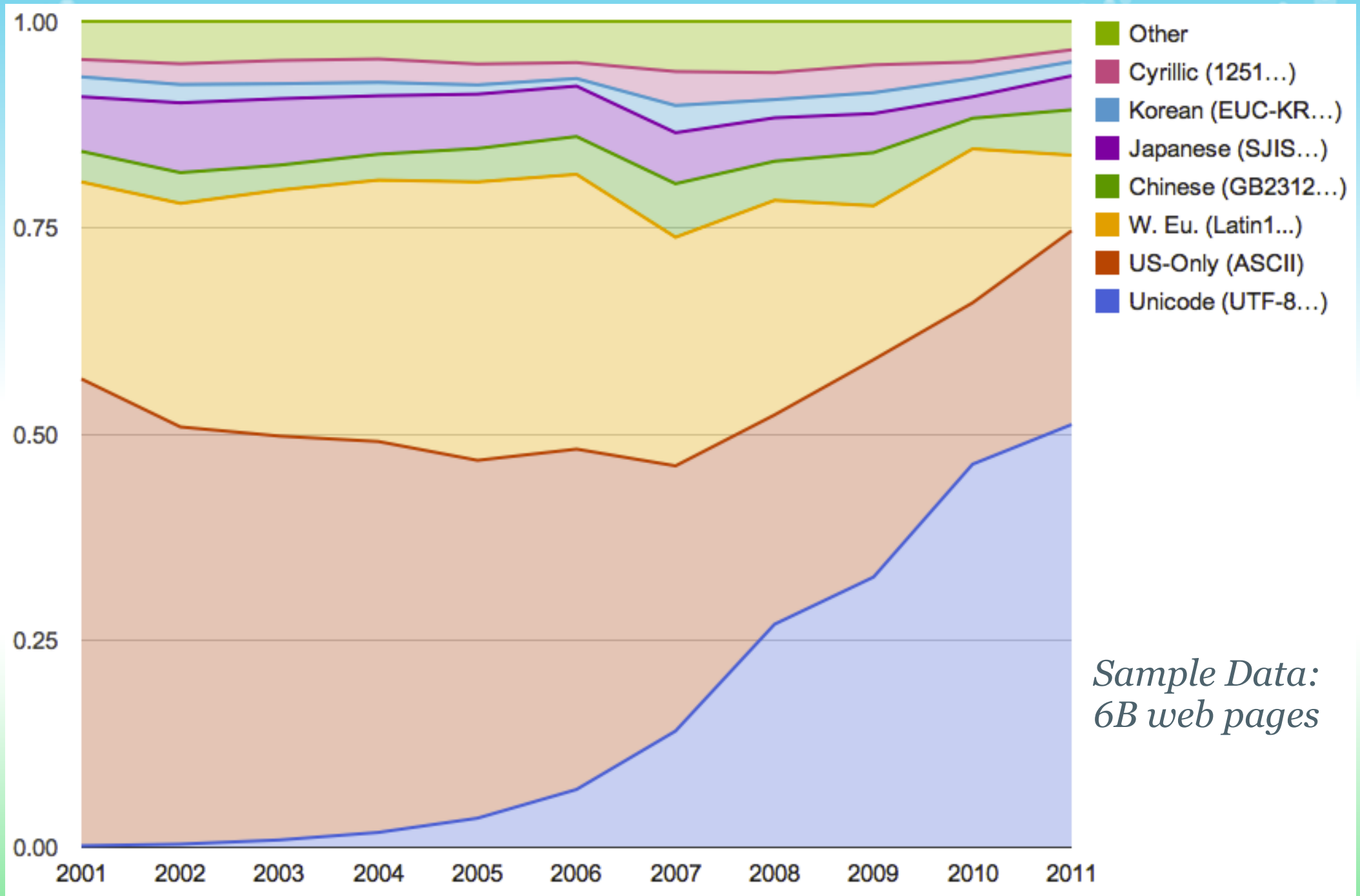
All modern software: OSs, smartphones, XML,…

Core Globalization Standards *and* Data

- Encoding (the Unicode Standard)
- Locales (CLDR/LDML)
  - **Transliteration**
- IDNA Compatibility
- Collation (Sorting/Matching)
- Regular Expressions
- Security
- …

http://www.unicode.org

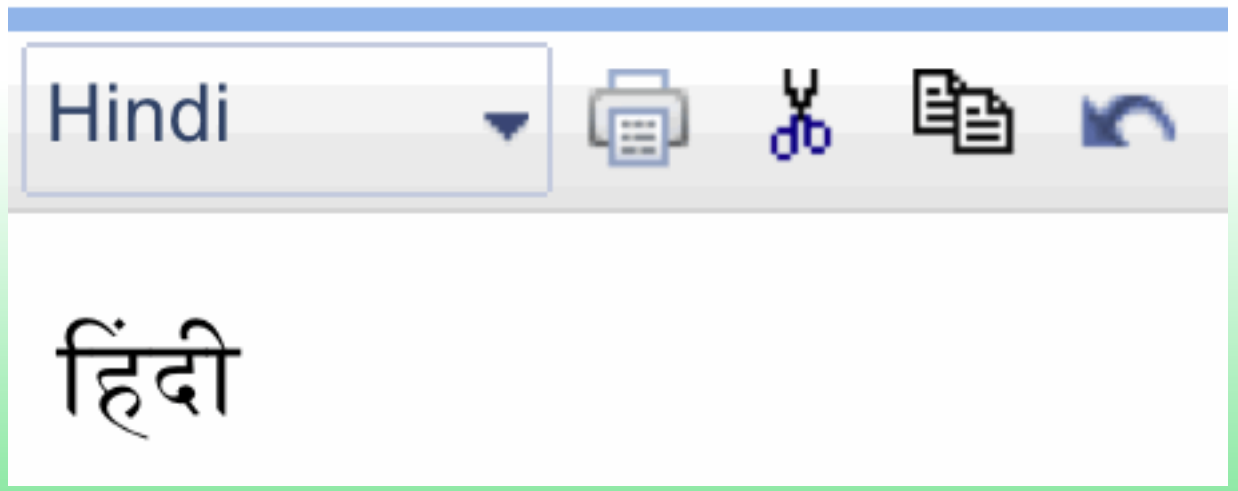
Unicode > 50% of the web



Sample Data: 6B web pages

Input Transliteration

- Keyboard input
- Typically multiple choices
  - User can pick if the 1st isn't suitable
  - Corpus driven: correspondences in bilingual usage
  - Adaptive: add choices users actually make
- Typically Latin → Y



Display Transliteration

- No choices, just display
  - Algorithmically driven
  - Sometimes augmented with dictionary lookup
- X → Y
  - Amharic → Latin
  - Cyrillic → Katakana
  - ...
- Different possible goals
  - Customary modern usage: *What people are used to*
  - Reversible: *Allows recovery of original*
- Proper Names - often augmented by dictionary lookup

Unicode Transliteration Identifiers

Syntax: *source-target/variant*

- *Russian to Japanese:*
  - ru-ja
- *Latin to Cyrillic:*
  - Latn-Cyrl (= und_Latn-und_Cyrl)
- *Latin to Cyrillic (UNGEGN):*
  - Latn-Cyrl/UNGEGN
- *Latin to Cyrillic (BGN):*
  - Latn-Cyrl/BGN

Unicode Language Identifiers (BCP47+) + variant codes

Unicode Transliteration Rules

Purely algorithmic

$\alpha \rightarrow a$;
$\xi \rightarrow x$;

Account for case

$A \rightarrow A$;
$\Xi \rightarrow X$;

Context required

$\theta \rightarrow th$;
Θ **} [:Ll:]** $\rightarrow Th$;
$\Theta \rightarrow TH$;

http://unicode.org/reports/tr35/#Transform_Rules

Chaining and Pivots

Malayalam-Latin =

- Malayalam-InterIndic ; InterIndic-Latin

Malayalam-Devanagari =

- Malayalam-InterIndic ; InterIndic-Devanagari

Spanish-Japanese =

- Spanish-Spanish(IPA) ; Spanish(IPA)-Japanese

Usage Example: Google Maps

For presentation, see http://goo.gl/wegil

Submissions (es_419-ja)

Valid rules file(s)

b → β ;
ch → ʧ ;
c } [eéií] → θ ;
c → k ;
————————
β → | b;
ð → | d;
a → ア;
ba → バ;
bb → ッ | b;

Test cases

ababuj → アバブフ

abades → アバデス

abadia → アバディア

abadiano → アバディアノ

…

Coöperation

Open-source Unicode transliterators

Used in a wide variety of products

UNGEGN rules welcome

This Presentation: http://goo.gl/jSDyP
(Action>Show speaker notes)

Questions and Discussion