Managing Transliteration of Bibliographic Data

Deborah W. Anderson, University of California, Berkeley
Steven R. Loomis, IBM Global Foundations Technology Team

Moderator: Margaret Hughes, Stanford University Libraries

June 27, 2015

Sponsor: Committee on Cataloging Asian and African Materials
Co-Sponsors:
Africana Librarians Council
Asian, African, Middle Eastern Section - ACRL
Committee on Research Materials on Southeast Asia
Committee on South Asian Libraries & Documentation
Middle East Librarians Association

#alctsAC15
Character encoding in Unicode, transliteration, and the future of multilingual search

Deborah Anderson
Researcher, Dept. of Linguistics, UC Berkeley
ALA June 2015
- The Unicode Standard (and related specs)
- Unicode website: http://unicode.org
- Other projects, including CLDR (locale data)
  - Includes some text transliteration data
Unicode Standard

- International standard, synchronized with ISO/IEC 10646
- Supported on modern browsers, mobile devices, and computers
- Backbone of multilingual text representation on the Internet, in email, text messages, word-processing docs, etc.
- Basis of Unicode-enabled fonts, keyboards, and OCR
The unlabelled scripts of India are:
(west) Gurmukhi, Gujarati, Kannada, Malayalam,
and (east) Tamil, Telugu, Oriya, Bengali, Barmese.
Unicode basics -1

- Unicode Standard assigns to letters and symbols of the world’s writing systems a unique number (code point)

  Latin letter b is “0062”

  Devanagari ब is “092D”

- Numbers (code points) stay the same on any modern device, whether an iPhone, on Android device, tablets, computers, etc.
Unicode basics -2

• New script/characters must be approved by two standards committees

• Proposals provide information on
  • characters, glyphs and names
  • sort order (i.e., a, A, b, B, c, C, etc.)
  • directionality of the script
  • other information needed to implement the script on computers
Languages and Scripts

- Number of languages: over 6,000 (*Ethnologue*)
- Number of scripts: ca. 223 (modern and historical)
  - Number in Unicode: 123
  - Not yet in Unicode: over 100 (approximately 35 modern)
UC Berkeley Script Encoding Initiative

- Works with users to get eligible characters and scripts into Unicode
- Remaining modern unencoded scripts are primarily in Africa, S/SE Asia

Medefaidrin
(Nigeria)

Mandombe
(Congo)
UC Berkeley Script Encoding Initiative

Masaram Gondi
Components

- Language
- Script
- Orthography (for non-Latin script=transliteration scheme)
- Text representation (fonts, keyboard/IME, rendering, software)
  - Example of rendering: क + ृ + ष → क्ष
  - Unicode code points: (&lt;0915, 094D, 0937&gt;)

Example of rendering: "क + ृ + ष → क्ष"
Example 1:
Language: English

- Script: Latin
  Orthography 1: Standard English Spelling
  Text representation: `cat`
  Unicode: `<0063, 0061, 0074>`

- Script: Latin
  Orthography 2: IPA (phonetic)
  Text representation: `khæt` (with Unicode-compliant font, etc.)
  Unicode: `<006B, 02B0, 00E6, 0074, 031A>`
Example 2:
Language: Modern Greek

- Script: Greek
  Orthography 1: Standard Modern Greek Spelling
  Text representation: γάτα (with Unicode-compliant font, etc.)
  Unicode: <03B3, 03AC, 03C4, 03B1>

- Script: Latin
  Orthography 2: ALA-LC Greek Romanization table
  Text representation: gata
  Unicode: <0061, 0041, 0074, 0061>
Example 3: Language: Japanese

Script: Han
  Orthography 1: Standard Japanese (as kanji)
  Text representation: 猫 (with Unicode-compliant font, etc.)
  Unicode: <732B>

Script: Hiragana
  Orthography 2: Standard Japanese (spelled out in hiragana)
  Text representation: ねこ (with Unicode-compliant font, etc.)
  Unicode: <306D, 3053>

Script: Latin
  Orthography 3: Standard Romanization of Japanese
  Text representation: neko
  Unicode: <006E, 0065, 006B, 006F>
Transliteration Tables for non-Latin scripts
(Romanization tables)

- ALA-LC: ca. 129 tables for languages 40 different scripts
- BGN/PCGN: 45 tables 17 scripts
- UNGEGN: 45 tables 26 scripts
- ISO standards:
  [Total number of scripts 220+ scripts]
Background on Romanization tables -1

ALA-LC Romanization tables* page:

- Tamil (2011)
- Romanian (in Cyrillic) (2014)
- Mande languages (in N'ko script) (2015)

*http://www.loc.gov/catdir/cpso/roman.html
Background on Romanization tables -2

LC Guidelines*:

• “should enable machine-transliteration as much as possible and preferably reversible transliteration”
• take equivalent Latin letter used from MARC Basic Latin, avoid rarer letters
• diacritics can be used to accommodate pronunciation; when using diacritics, avoid those not widely supported or whose position may interfere with printing/display of Latin letter (i.e., those diacritics occurring below).

* http://www.loc.gov/catdir/cpso/romguid_2010.html
ALA-LC Romanization Tables: Adding New Tables

- 6 months - 1 year (typically)
- If controversial, can take 2-4 years (or longer)
Transliteration: Advantages

• Consistent set of rules to follow
  • Can find book title if script is not in Unicode or if no Unicode-enabled font is available


• Can find book if there is an error in a record in the original script (in Unicode), example for Arabic
Transliteration: Problems

- Different transliteration schemes (and legacy data) not conformant with ALA-LC Romanization may make it hard to find a title

<table>
<thead>
<tr>
<th>яйца Фаберже</th>
<th>Fabergé eggs</th>
<th>yaytsa Faberzhe</th>
<th>BGN/PCGN</th>
</tr>
</thead>
<tbody>
<tr>
<td>jajca Faberže</td>
<td>Scholarly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>âjca Faberže</td>
<td>ISO</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Many scripts missing from ALA-LC Romanization tables
- Takes time to propose transliteration table and get approved
ALA-LC Romanized Tables: Exs. of Missing Scripts with Printed Materials

- **Africa (4)**: Bamum, Bassa Vah, Mende Kikakui, Osmanya
- **South Asian (15)**: Chakma, Grantha, Kaithi, Khojki, Khudawadi Mahajani, Meetei Mayek, Modi, Mro, Saurashtra, Siddham, Syloti Nagri, Takri, Tirhuta, Warang Citi
- **SE Asian (7)**: Kayah Li, New Tai Lue, Pahawh Hmong, Pau Cin Hau, Tai Le, Tai Tham, Tai Viet
- **Indonesia and Oceania (3)**: Buginese, Rejang, Sundanese
- **E Asia (3)**: Lisu, Miao, Yi
Components

- Language
- Script
- Orthography (for non-Latin script = transliteration scheme)
- Text representation (fonts, keyboard/IME, rendering, software)
- Unicode code points (<XXXX, XXXX>)
Issues with fonts, keyboards, and software

• Font issue
Issues with fonts, keyboards, and software

N’Ko: Using older rendering engine software/OS:

On Windows 8:
Issues with fonts, keyboards, and software (or messy data?)

- Vietnamese

Correct:

Dại Việt sử kỳ toàn thư.

on OCLC FirstSearch:

Dai Viet Su Ky Toan Thu.

Đại Việt sử kỳ toàn thư.
Components

- Language
- Script
- Orthography (for non-Latin script=transliteration scheme)
- Text representation (fonts, keyboard/IME, rendering, software)
- Unicode code points (<XXXXX, XXXXX>)
Issues with Unicode

- Missing scripts or characters
  - About 100 scripts are known to be missing
Transliteration tools -1

- CLDR has 16 script-script transliteration tables*, possible to have more added
- Process of adding more tables requires submitting rules in a special syntax which needs to catch the edge cases, like casing (UTR #35)

*See http://www.unicode.org/cldr/charts/latest/transforms/index.html
Transliteration tools -2

- Google transliteration input tool* has 25 languages, but is not rule-based
- Type the word in phonetically in Latin, pick from list:

*http://www.google.com/inputtools/services/features/transliteration.html
The Future....

• Will fonts/software support the world’s scripts?
• Be able to search in more of the original scripts?
• Add ALA-LC transliteration schemes to CLDR?
Thank you

• Thanks to Bruce Johnson (LC); UC Berkeley librarians Shayee Khanaka, Virginia Shih, Adnan Malik, Haiqing Lin, Noriko Nishizawa, and Jaeyong Change; Google Input Tools members Xiangye Xiao, Yuanbo Zhang, Yingbing; Unicode Technical Director Ken Whistler

• Support for SEI project comes from a Google Research Award and NEH grant PR-50205-15
Questions?

Debbie Anderson
dwanders@berkeley.edu

Script Encoding Initiative project:
http://linguistics.berkeley.edu/sei