Enhancing Chinese character normalization in Primo with the HKIUG TSVCC mapping table

Alan Ng
Systems Librarian
The Chinese University of Hong Kong Library
Agenda

- Primo out-of-box character normalization
- Background on CJK normalization
- HKIUG TSVCC mapping table
- Implementing TSVCC on Primo
About CUHK Library

❖ established in 1963
❖ 7 branches
❖ 200 staff
❖ 260K current patrons
❖ 130K journal subscriptions, 4.5M ebooks, 2.5M printed volumes
❖ special collections includes from oracles bones, Chinese rare books, modern Chinese literary archives …
character normalization

- different people type differently
- normal to expect “Apple” will have the exact results from “aPPLE”, “ApPlE”, “applE” …
- before indexing, Primo will first “clean up” (normalize) the data to its lower case (e.g. A —> a, B —> b …)
- Primo FE will do the same for the search term typed by users, to get a match with the index
Primo out-of-box normalizations

- Primo provides OTB normalizations for different languages at:
  - /exlibris/primo/p4_1/ng/jaguar/home/profile/analysis/specialCharacters/CharConversion/OTB/OTB

- e.g.
  - latin languages (non_cjk_unicode_normalization.txt)
  - CJK (cjk_unicode_trad_to_simp_normalization.txt)
OTB CJK normalization table

- 2700+ entries
- mainly for mapping Traditional Chinese into its Simplified form
- assume it is a 1:1 mapping, Simplified Chinese being the “lowercase” like the English language
- But in fact, Simplified Chinese is only one kind of variant form for Chinese character
- other variant forms (ideograph) of the same character need to be cover as well
extract of the OTB table

<table>
<thead>
<tr>
<th>Code 1</th>
<th>Code 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4E1F</td>
<td>4E22</td>
</tr>
<tr>
<td>4E26</td>
<td>5E76</td>
</tr>
<tr>
<td>4E7E</td>
<td>5E72</td>
</tr>
<tr>
<td>4E82</td>
<td>4E71</td>
</tr>
<tr>
<td>4E99</td>
<td>4E98</td>
</tr>
<tr>
<td>4E9E</td>
<td>4E9A</td>
</tr>
<tr>
<td>4F15</td>
<td>592B</td>
</tr>
<tr>
<td>4F47</td>
<td>4F2B</td>
</tr>
<tr>
<td>4F48</td>
<td>5E03</td>
</tr>
<tr>
<td>4F54</td>
<td>5360</td>
</tr>
<tr>
<td>4F75</td>
<td>5E76</td>
</tr>
<tr>
<td>4F86</td>
<td>6765</td>
</tr>
<tr>
<td>4F96</td>
<td>4ED1</td>
</tr>
<tr>
<td>4F9A</td>
<td>5F87</td>
</tr>
<tr>
<td>4FB6</td>
<td>4FA3</td>
</tr>
<tr>
<td>4FB7</td>
<td>5C40</td>
</tr>
<tr>
<td>4FC1</td>
<td>4FE3</td>
</tr>
<tr>
<td>4FC2</td>
<td>7CF8</td>
</tr>
<tr>
<td>4FD4</td>
<td>4F23</td>
</tr>
</tbody>
</table>
Traditional Chinese characters have been used since as early as 2nd century BC (Han Dynasty, 漢朝)

- used by people in Taiwan, Hong Kong and Macau

Simplified Chinese characters were introduced by PRC government during 1950’s

- used by people in PRC, SE Asia countries e.g. Singapore and Malaysia

“Traditional Chinese” is a retronym
in the Chinese language:

- “Traditional Chinese” is 正體字 OR 繁體字, meaning formal form, complicated format
- “Simplified Chinese” is 簡體字, meaning simplified form
Japanese adopted Chinese characters in ~5th century AD, called it Kanji (かんじ, 漢字)

Korean adopted Chinese characters in ~4th century AD, called it Hanja (한자, 漢字)

both Japanese and Korean have created variant forms for the Chinese characters (glyph), making new ideograms
CJK in Unicode character table

- Unicode table has different segments for CJK
- 2E80 - 2EFF (CJK radicals supplement)
- 3000 - 303F (CJK symbols and punctuation)
- 31C0 - 31EF (CJK strokes)
- 3200 - 32FF (enclosed CJK letters and months)
- 3400 - 4DBF (CJK unified ideographs extension A)
- 4E00 - 9FFF (CJK unified ideographs)
- F900 - FAFF (CJK compatibility ideographs)
multiple-mapped nature of CJK

- for the same character, it can be the simplified form of two Traditional characters, having 2 meanings
  
  - e.g. 干 (U+5E72) is simplified form of both 乾 (U+4E7E) and 幹 (U+5E79), meaning “dry” and “take an action”

- for the same character, it can be the simplified form of another Traditional character, but also being the traditional form for another meaning
  
  - e.g. 松 (U+677E) as Traditional form means “pine tree”, is also the simplified form for 鬆 (U+9B06), meaning “soft”
preference by different countries

- the Chinese people recognize 裴 (U+88F4) as the standard form, while Korean people use 褚 (U+88F5)
- Chinese people use 漱 (U+6F80) as Traditional form and 澸 (U+6DA9) as simplified form, while Japanese people refer 涓 (U+6E0B) for the exact same character
Chinese inputting methods

- Chinese characters can be inputted by various means
  - Pinyin, Cangjie, Zhuyin, Romanji (Japanese), HNC Romaja (Korean)
  - handwriting on smartphones and tablets
- no way to control the resulting Chinese character (ideograph) retrieved by these inputting methods
- no way to control how users key in search terms, as well as the metadata keyed in by catalogers
spot the difference

- 内 (U+5167) | 内 (U+5185), meaning “inside”
- 別 (U+5225) | 別 (U+522B), meaning “another”
- 夠 (U+5920) | 夠 (U+591F), meaning “enough”
- 歩 (U+6B65) | 歩 (U+6B69), meaning “pace”
- 瘦 (U+75E9) | 瘦 (U+7626), meaning “thin”
- 黒 (U+9ED1) | 黒 (U+9ED2), meaning “black”
these sample characters are NOT the same glyph presented in different styles (think of fonts)

they are different ideograms, having individual code-points in the Unicode table

these ideographs are variant forms of the “same” Chinese character

the OTB Primo CJK normalization table is not sufficient

need to make logical groups of ideograph based on these complicated relationship
HKIUG is a local user group for Millennium ILS, including all government funded university libraries, and other institutions.

In 2004, they attempted to compile a list which could indicate all the possible relationships among these groups Chinese ideograph.

TSCVV stands for “Traditional, Simplified, Variant Chinese Characters”
Hear the list was used as a workaround in the Millennium OPAC and its search index.

- it has been a hard-coded customization for Millennium users in Hong Kong, needs to be done by III

- Until now, Millennium and Sierra still do not have an official way to handle customer-defined normalization mapping, like Primo does
HKGU TSVCC mapping table

- The list was completed in 2006
- The list has consolidated relationships found in:
  - EACC (LC’s East Asian Character Code Tables)
  - EACC-to-Unicode mapping
  - kSimplifiedVariant and kZVariant (Unicode consortium’s Unihan database)
  - Entries defined by HKIUG (via original cataloging of Chinese material)
implement TSVCC on Primo

- analyze and compare the OTB table
  (cjk_unicode Tradable_to_simplified_normalization.txt) with the TSVCC table

- consolidate all mapping into:
  - /exlibris/primo/p4_1/ng/jaguar/home/profile/analysis/specialCharacters/CharConversion/CUSTOMER/CUSTOMER/user_defined_unicode_normalization.txt

- 3490+ ideograph groups,

- 4600+ entries, up from original 2700+ entries

- delete the index (clean_indexes.sh) and run a full re-index (Indexing_and_Didumean_and_Hotswapping)
extract of the resulting table

<table>
<thead>
<tr>
<th>78</th>
<th>79</th>
<th>80</th>
<th>81</th>
<th>82</th>
<th>83</th>
<th>84</th>
<th>85</th>
<th>86</th>
<th>87</th>
</tr>
</thead>
<tbody>
<tr>
<td>4E79</td>
<td>4E7E</td>
<td>4E81</td>
<td>5E79</td>
<td>69A6</td>
<td>6E0B</td>
<td>6F80</td>
<td>6F81</td>
<td>6FC7</td>
<td>88F4</td>
</tr>
<tr>
<td>5E72</td>
<td>5E72</td>
<td>5E72</td>
<td>5E72</td>
<td>5E72</td>
<td>6DA9</td>
<td>6DA9</td>
<td>6DA9</td>
<td>6DA9</td>
<td>88F5</td>
</tr>
<tr>
<td># 車 --&gt; 干 (1, tsvcc)</td>
<td># 乾 --&gt; 干 (1, otb, tsvcc)</td>
<td># 乾 --&gt; 干 (1, tsvcc)</td>
<td># 幹 --&gt; 干 (1, otb, tsvcc)</td>
<td># 幹 --&gt; 干 (1, otb, tsvcc)</td>
<td># 漬 --&gt; 澀 (1, tsvcc)</td>
<td># 漬 --&gt; 澀 (1, otb, tsvcc)</td>
<td># 漬 --&gt; 澀 (1, otb, tsvcc)</td>
<td># 漬 --&gt; 澀 (1, otb, tsvcc)</td>
<td># 漬 --&gt; 澀 (pure tsvcc)</td>
</tr>
</tbody>
</table>
three sample ideograph groups

- **RED** is the group representative in OTB
- **BLACK** is from OTB “cjk_unicode_trad_to_simp_normalization.txt”
- **BLUE** is from TSVCC table (new entries)
  - 涩 (U+6DA9) | 渋 (U+6E0B) | 澀 (U+6F80) | 澲 (U+6F81)
  - 澠 (U+6FC7) | 澥 (U+7012)
  - 裘 (U+88F4) | 襅 (U+88F5)
  - 干 (U+5E72) | 軆 (U+4E79) | 乾 (U+4E7E) | 乾 (U+4E81)
  - 幹 (U+5E79) | 干 (U+69A6)
case 1: BEFORE adding TSVCC

- search term: 涩谷 (12 results, OTB normalized)
case 1: BEFORE adding TSVCC

- search term: 渋谷 (12 results, OTB normalized)
case 1: BEFORE adding TSVCC

- search term: 渋谷 (33 results, NOT normalized)
case 1: BEFORE adding TSVCC

- search term: 滋谷 (8 results, NOT normalized)
case 1: AFTER adding TSVCC

- search term: 浪 (71 results, normalized)
case 1: AFTER adding TSVCC

- search term: 澀 (71 results, normalized)
case 1: AFTER adding TSVCC

- search term: 湯 (71 results, normalized)
case 1: AFTER adding TSVCC

- search term: 洵 (71 results, normalized)
case 2: BEFORE adding TSVCC

- search term: 裨 (1757 results, OTB normalized)
case 2: BEFORE adding TSVCC

- search term: 亵 (0 result, NOT normalized)
case 2: AFTER adding TSVCC

- search term: 裳 (9 results, normalized)
case 2: AFTER adding TSVCC

- search term: 我 (9 results, normalized)
points of interest

- make sure not group the characters in loop
  - CORRECT:
    - B → A
    - C → A
    - D → A
    - E → A
  - NOT SO SURE:
    - B → A
    - C → A
    - A → D
    - D → E

- make sure all OTB entries appear in the user-defined table, not sure how Primo take precedence among the files
Reference (1)

- The Unicode consortium
  - http://www.unicode.org/reports/tr38/index.html
  - http://www.unicode.org/charts/unihangridindex.html
  - http://www.unicode.org/charts/unihan.htm

- Unicode character table

- Wikipedia
Reference (2)

- Hong Kong Innovative User Group TSVCC table

- Library of Congress EACC
  - http://www.loc.gov/marc/specifications/specchareacc.html
Q & A
Thank you!

Alan Ng (alanng@cuhk.edu.hk)