Multilingual Issues of Open Source ILS

Jason (Qing) Zou
Systems Librarian
Lakehead University

Grace (Guoying) Liu
Systems Librarian
University of Windsor

January 31 2009
Outline

- Writing Systems
- Romanization
- Unicode
- Open Source ILS
- Simplified Chinese Version of Evergreen
- Conclusions
Writing Systems

- Represent the sounds of a language by written or printed symbols (WordNet)

- Requirements:
  - Defined base elements/symbols (scripts)
  - Rules and conventions
  - A language (generally a spoken language)
Types of Writing Systems

- Logographic (Chinese characters: CJK)
- Alphabetic (Latin alphabet: Cyrillic, Latin)
- Abjad (Arabic alphabet, Arabic, Hebrew)
- Abugida (Indian Devanagari: India, Canadian Aboriginal Syllabics)
Scripts

10 major scripts, to write ~95% of all languages

Roman abcdeéèêøœ …
Greek αβγδε …
Cyrillic авгдеж …
Hebrew … אבגדה
Arabic … حجثبأ
Indic (11) आइईउऊ …
Thai กขฃคฅฆ …
Japanese あいうえお …
– Hiragana あいうえお …
– Katakana アイウエオ …
Korean 가각각각간 …
Chinese 甲乙丙丁 …

Arsenault, C., 2003
**Multilingual Information Systems**

- Contains records in more than one language
- The system interface is in more than one language
- The system is able to display text in more than one script
- The system allows the end user to build queries in more than one script

Arsenault, C., 2003
Multilingual Information Systems

Two models for multi-script records in MARC21:

*Model A*
- original scripts in 880 fields
- primary descriptive fields using Romanized form

*Model B*
- transcribe data directly into regularly tagged fields
Zhong guo wen hua jing dian / [Zhu Xi deng zhu; zhu bian Ren Jiyu; zhi xing zhu bian Pan Yuan].

Di 1 ban

Hangzhou : Xi leng yin she chu ban she, 2007.

14 v. ; 29 cm.

中國文化經典 / [朱熹等著；主編任繼愈；執行主編潘淵].

第1版


朱熹, 1130-1200.

任繼愈, 1916-

潘淵.

From Yale University Catalogue
<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>Chen, Xiyong</td>
</tr>
<tr>
<td>1001</td>
<td>陳錫勇</td>
</tr>
<tr>
<td>24510</td>
<td>Guodian Chu jian Laozi lun zheng / zuo zhe Chen Xiyong.</td>
</tr>
<tr>
<td>24510</td>
<td>郭店楚簡老子論證 / 作者陳錫勇.</td>
</tr>
<tr>
<td>250</td>
<td>Chu ban.</td>
</tr>
<tr>
<td>250</td>
<td>初版</td>
</tr>
<tr>
<td>260</td>
<td>Taipei shi: Li ren shu ju, 2005.</td>
</tr>
<tr>
<td>70002</td>
<td>Laozi. Dao de jing.</td>
</tr>
<tr>
<td>70002</td>
<td>老子. 道德經.</td>
</tr>
</tbody>
</table>

From McGill University Catalogue
Multilingual Information Systems

Non-Roman data in North American OPACs

Stored? yes no

Displayed? yes no

Indexed? yes no

Romanization

Vernacular

Arsenault, C., 2003
Romanization

- Representation of a written word or spoken speech with Roman alphabet
- Methods
  - Transliteration: written text, Russian
  - Transcription: spoken word, CJK
<table>
<thead>
<tr>
<th>Vernacular</th>
<th>Romanization</th>
<th>Vernacular</th>
<th>Romanization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper case letters</td>
<td>Lower case letters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>а</td>
<td>a</td>
</tr>
<tr>
<td>Б</td>
<td>B</td>
<td>б</td>
<td>b</td>
</tr>
<tr>
<td>В</td>
<td>V</td>
<td>в</td>
<td>v</td>
</tr>
<tr>
<td>Г</td>
<td>G</td>
<td>г</td>
<td>g</td>
</tr>
<tr>
<td>Д</td>
<td>D</td>
<td>д</td>
<td>d</td>
</tr>
<tr>
<td>Е</td>
<td>E</td>
<td>е</td>
<td>e</td>
</tr>
<tr>
<td>Ё</td>
<td>E</td>
<td>ё</td>
<td>ë</td>
</tr>
<tr>
<td>Ж</td>
<td>Zh</td>
<td>ж</td>
<td>zh</td>
</tr>
<tr>
<td>З</td>
<td>Z</td>
<td>з</td>
<td>z</td>
</tr>
<tr>
<td>И</td>
<td>I</td>
<td>и</td>
<td>ï</td>
</tr>
<tr>
<td>Й</td>
<td>Ì</td>
<td>й</td>
<td>Ì</td>
</tr>
<tr>
<td>К</td>
<td>K</td>
<td>к</td>
<td>k</td>
</tr>
<tr>
<td>Л</td>
<td>L</td>
<td>л</td>
<td>l</td>
</tr>
<tr>
<td>М</td>
<td>M</td>
<td>м</td>
<td>m</td>
</tr>
</tbody>
</table>

(see Note 1)
新年好  xin nian hao
温莎大学  wen sha da xue
刘国英  liu guo ying
明清小说比较研究  Ming Qing xiao shuo bi jiao yan jiu
李白和他的诗歌  Li Bai he ta de shi ge
Chinese Romanization

- Only Transcription is possible
- Two Romanization systems for bibliographic control in North America:
  - Wade-Giles (through October 2000)
  - Pinyin (After October 2000)

唐宋全诗

Wade-Giles: Tʻang2 Sung4 chʻüan2 shih1
Pinyin: Táng Sòng quán shī
Pinyin 拼音

- Literally “spell the sound”
- Based on Hanyu Pinyin (Chinese Phonetics) which was adopted in 1958 by mainland China
- Used for many years in libraries in Europe and Australia
- Oct. 1, 2000, LC and other libraries in US adopted Pinyin
Issues of Pinyin

Chinese characters

- 47,043 in 1716 (康熙字典)
- ~60,000 in 1990 (漢語大字典)
- sharing only around 1,300 syllables in spoken Chinese (Arsenault, 2001)
- High level of Homophonous ambiguity

liu: 刘 六 流 留 柳 … (over 30 possibilities)
Issues of Pinyin

Word division (syllable integration)

下午我去图书馆了. (I went to the library this afternoon)
Pinyin: Xia wu wo qu tu shu guan le.
or: Xiawu wo qu tushuguan le.
Issues of Pinyin

No consistent rules on syllable integration

中国话
zhong guo hua
zhong-guo hua
zhongguo hua
zhongguohua

More difficult to form queries
Issues of Pinyin

- A standard based on Mandarin
- In cataloguing, impossible to maintain consistency
- Infrequently used characters may be impossible accessed by phonetic scripts
# ASCII Code Chart

<table>
<thead>
<tr>
<th>Bit Pattern</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 0 0 0</td>
<td>NUL</td>
<td>DLE</td>
<td>SP</td>
<td>0</td>
<td>@</td>
<td>P</td>
<td>¬</td>
<td>p</td>
</tr>
<tr>
<td>0 0 0 1 1 1</td>
<td>SOH</td>
<td>DC1</td>
<td>!</td>
<td>1</td>
<td>A</td>
<td>Q</td>
<td>o</td>
<td>q</td>
</tr>
<tr>
<td>0 0 1 0 0 2</td>
<td>STX</td>
<td>DC2</td>
<td>&quot;</td>
<td>2</td>
<td>B</td>
<td>R</td>
<td>b</td>
<td>r</td>
</tr>
<tr>
<td>0 0 1 1 3 3</td>
<td>ETX</td>
<td>DC3</td>
<td>#</td>
<td>3</td>
<td>C</td>
<td>S</td>
<td>c</td>
<td>s</td>
</tr>
<tr>
<td>0 1 0 0 4 4</td>
<td>EOT</td>
<td>DC4</td>
<td>$</td>
<td>4</td>
<td>D</td>
<td>T</td>
<td>d</td>
<td>t</td>
</tr>
<tr>
<td>0 1 0 1 5 5</td>
<td>ENQ</td>
<td>NAK</td>
<td>%</td>
<td>5</td>
<td>E</td>
<td>U</td>
<td>e</td>
<td>u</td>
</tr>
<tr>
<td>0 0 1 0 6 6</td>
<td>ACK</td>
<td>SYN</td>
<td>&amp;</td>
<td>6</td>
<td>F</td>
<td>V</td>
<td>f</td>
<td>v</td>
</tr>
<tr>
<td>0 0 1 1 7 7</td>
<td>BEL</td>
<td>ETB</td>
<td>'</td>
<td>7</td>
<td>G</td>
<td>W</td>
<td>g</td>
<td>w</td>
</tr>
<tr>
<td>1 0 0 0 8 8</td>
<td>BS</td>
<td>CAN</td>
<td>(</td>
<td>8</td>
<td>H</td>
<td>X</td>
<td>h</td>
<td>x</td>
</tr>
<tr>
<td>1 0 0 1 9 9</td>
<td>HT</td>
<td>EM</td>
<td>)</td>
<td>9</td>
<td>I</td>
<td>Y</td>
<td>i</td>
<td>y</td>
</tr>
<tr>
<td>1 0 1 0 10</td>
<td>LF</td>
<td>SUB</td>
<td>*</td>
<td>:</td>
<td>J</td>
<td>Z</td>
<td>j</td>
<td>z</td>
</tr>
<tr>
<td>1 0 1 1 11</td>
<td>VT</td>
<td>ESC</td>
<td>+</td>
<td>;</td>
<td>K</td>
<td>[</td>
<td>k</td>
<td>{</td>
</tr>
<tr>
<td>1 1 0 0 12</td>
<td>FF</td>
<td>FS</td>
<td>,</td>
<td>&lt;</td>
<td>L</td>
<td>\</td>
<td>l</td>
<td>/</td>
</tr>
<tr>
<td>1 1 0 1 13</td>
<td>CR</td>
<td>GS</td>
<td>-</td>
<td>=</td>
<td>M</td>
<td>]</td>
<td>m</td>
<td>}</td>
</tr>
<tr>
<td>1 1 1 0 14</td>
<td>SO</td>
<td>RS</td>
<td>.</td>
<td>&gt;</td>
<td>N</td>
<td>^</td>
<td>n</td>
<td>~</td>
</tr>
<tr>
<td>1 1 1 1 15</td>
<td>SI</td>
<td>US</td>
<td>/</td>
<td>?</td>
<td>O</td>
<td>_</td>
<td>o</td>
<td>DEL</td>
</tr>
</tbody>
</table>
Language and Unicode

What is Unicode?

Unicode provides a unique number for every character,
no matter what the platform,
no matter what the program,
no matter what the language.

Unicode

- Support by large software companies
- Support by organizations
- Support by countries
**UTF-8**

- Unicode Transformation Format
- UTF-8

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>U-00000000 – U-0000007F</td>
<td>0xxxxxxx</td>
</tr>
<tr>
<td>U-00000080 – U-000007FF</td>
<td>110xxxx 10xxxxxx</td>
</tr>
<tr>
<td>U-00000800 – U-0000FFFF</td>
<td>1110xxx 10xxxxxx 10xxxxxx</td>
</tr>
<tr>
<td>U-00010000 – U-001FFFFF</td>
<td>11110xxx 10xxxxxx 10xxxxxx 10xxxxxx</td>
</tr>
<tr>
<td></td>
<td>家 (home)</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>UTF-8</td>
<td>E5 AE B6</td>
</tr>
<tr>
<td>UTF-16</td>
<td>5B B6</td>
</tr>
<tr>
<td>UTF-32</td>
<td>00 00 5B B6</td>
</tr>
</tbody>
</table>
Commercial ILS

- Major commercial ILS products
  - Voyager
  - Aleph
  - Symphony
  - Millennium
Open Source ILS

- Major Open Source ILS Products

  - Koha
  - Evergreen
Open Source ILS

- What is open source?
## Open Source Software

<table>
<thead>
<tr>
<th>Operating System</th>
<th>Windows</th>
<th>Linux variants (Red hat, Debian, Ubuntu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>Oracle, DB2, SQL Server</td>
<td>MySQL, PostgreSQL</td>
</tr>
<tr>
<td>Web Server</td>
<td>Internet Information server</td>
<td>Apache</td>
</tr>
<tr>
<td>Browser</td>
<td>Internet explorer</td>
<td>Firefox</td>
</tr>
<tr>
<td>Programming Language</td>
<td>Microsoft C++, Visual Basic, C#</td>
<td>Perl, Php, Python</td>
</tr>
<tr>
<td>Office</td>
<td>Microsoft Office</td>
<td>Open Office</td>
</tr>
</tbody>
</table>
Integrated Library Systems

- Regular modules
  - Circulation
  - Acquisition
  - Cataloguing
  - System Administration
  - OPAC
Language Issues in ILS

- Store
- Display
- Index/Search
- Sort
## Language Issues in ILS

### Display

<table>
<thead>
<tr>
<th>Formats: Standard</th>
<th>Citation</th>
<th>MARC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>▶️ Zhang, Boxun.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶️ 张宝顺.</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Wei Yuan &quot;Laozi ben yi&quot; yan jiu / Zhang Boxun zhu.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>白元《老子本義》研究 / 張寶順.</td>
<td></td>
</tr>
<tr>
<td>Edition</td>
<td>Chu ban.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>中編.</td>
<td></td>
</tr>
<tr>
<td>Published</td>
<td>Taipei Xian Yonghe Shi : Hua mu lan wen hua chu ban she, 2007.</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>176 p.; 27 cm.</td>
<td></td>
</tr>
<tr>
<td>Series</td>
<td>▶️ Gu dian wen xian yan jiu ji kan. Si bian ; di 18 ce.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>大千文化研究期刊館. 第版 ; 第 18 卷.</td>
<td></td>
</tr>
</tbody>
</table>
Language Issues in ILS

- Index/Search
- Sort
Koha

- Features
  - Debian/Linux, Windows
  - Perl
  - MySQL/Zebra
  - LibLime Company
Important links here.

http://koha.wikispaces.com/
“ti,wrld: perl” 總共找到 2 筆。

1. Perl 5.X 指令速查手冊 作者: Micheal O Foghlu著
   峯岫資訊 2001- [16], 378面, 21公分.
   ...Perl 5.X指令速查手冊 ... Perl 5 quick reference ... Perl(電腦程式語言) ...
   Copies available: 輔仁大學 (1),
   □ 312.932P39

2. Perl 程式設計 作者: Larry Wall, Tom Christiansen, Randal L. Schwartz原著
   歐萊德 1999- 10,750面, 23公分.
   ...Perl程式設計 ... Programming Perl 2/e ... Perl(電腦程式語言) ...
   Copies available: 輔仁大學 (1),
   □ 312.932P39
Koha

- Debian/Linux
- Perl
- MySQL/Zebra
- Large community
- Fully developed ILS
- Fully support standards
- Language supports
Evergreen

- Debian/Linux, Windows
- OpenSRF
- PostgreSQL
- C, Perl
- Equinox software Company
Evergreen in Canada

- BC public consortium (Sitka)
- UPEI
- Project Conifer
Chinese Localization

- Introduction
- Goals
Chinese Localization

- Features are crucial to localization
  - Supports Unicode
  - Indexing/Searching: PostgreSQL Tsearch2
Simplified Chinese Version

- Interface localization
- Indexing
- Searching
- Sorting
- Others
Simplified Chinese Version

- Interface
  - Utilizes Pootle
  - Uses gettext tools to convert Portable Objects to Document Type Definition files
  - Updates DTD files hourly
<table>
<thead>
<tr>
<th>Original</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>主题词</td>
</tr>
<tr>
<td>Series</td>
<td>丛书名</td>
</tr>
<tr>
<td>Keyword</td>
<td>关键词</td>
</tr>
<tr>
<td></td>
<td>检索类型</td>
</tr>
<tr>
<td>Format</td>
<td>文献格式</td>
</tr>
<tr>
<td>Loading</td>
<td>正在加载中...</td>
</tr>
</tbody>
</table>
Indexing

- Tsearch2 (PostgreSQL full text search engine)
  - Configure Tsearch2 to be able to handle Chinese records
  - Utilize Chinese words segment algorithms
Indexing

- Tsearch2 (to_tsvector, to_tsquery)

  # SELECT to_tsvector('Social history of China') ;
  to_tsvector

  ------------------------------------
  'china': 4 'social': 1 'histori': 2 (1 row)
Indexing

- Tsearch2 (to_tsvecto, to_tsqury)

```sql
# SELECT to_tsvecto('中国社会历史')

  to_tsvecto

-------------------------

'中国社会历史': 1 (1 row)
```
Indexing

- Tsearch2 (to_tsvector, to_tsquery)

# SELECT olis_cn_index('中国社会历史') as title

------------------------------------
“中”：1  ‘史’：6  ‘国’：2  ‘会’：4  ‘历’：5  ‘社’：3  ‘中国’：7  ‘国社’：8  ‘会历’：10  ‘历史’：11  ‘社会’：9
Searching

- Searching is the opposite process of indexing

- Utilize Chinese words segment algorithms to divide and form queries which can be understood by the system
The search phrase “中国社会历史” will be formed as “中 & 国 & 社 & 会 & 历 & 史” and then fed into the system.
Sorting

- Romanize Chinese characters with tone information
- Sort the corresponding Pinyin of Chinese records to obtain A-Z order
Sorting

# Select utf8_pinyin('中国社会历史') as pinyin;
Pinyin
-------------------------------------
zhong4guo2she4hui4li4shi3
魏源《老子本義》研究
Zhang, Boxun.
Chu ban. | 2007 Hua mu lan wen hua chu ban she | print 176 p.; 27 cm.

您的检索结果几乎为零。
或许您指的是: A

提示：如果您未找到您想要的，请试用查询条右边的范围选项来扩大您的检索范围
Import Issues

- CMARC
- CNMARC
- 880 fields
Conclusions

- Unicode is not the answer for everything
- More practical to tweak a system language by language
Thank you for your interest and attention!

Any Questions?
Contact Information

Jason Zou
qzhou@lakeheadu.ca

Guoying Liu
gliu@uwindsor.ca