East Meets Rest

Adding East Asian Scripts to Harvard’s ILS

Prepared for presentation to the
North American Aleph Users’ Group
2 June 2003

Charles Husbands, HUL Office for Information Systems
charles_husbands@harvard.edu
A short history of HOLLIS
(Harvard Online Library Information System)

- 1985: NOTIS-derived Acquisitions and Cataloging
- 1987: Circulation implementation begins
- 1988: OPAC implementation makes HOLLIS a real Integrated Library System (ILS)
- Ca. 1995: Thinking about next generation begins
- November 2000: Aleph contract signed
- July 2002: Aleph 15.2 installed as new ILS
- 2002: The name HOLLIS now encompasses Aleph ILS and other catalogs and electronic resources
Non-latin scripts at Harvard

- Pre-Aleph system could use only Latin script data
- Aleph support priorities for HOLLIS
  1. CJK
  2. Arabic and Hebrew
  3. Cyrillic and Greek
- CJK first
- Over 500,000 records
  - 60% Chinese
  - 25% Japanese
  - 15% Korean
Challenges

- Huge character repertoire
- Homonyms
- Other one-to-many issues
- Collating sequence
- Input method
- Display
- MARC management
Simplified and traditional forms and homonyms

Taiwan shi tian ye yan jiu tong xun.

臺灣史田野研究通訊．

Di 1 qi- ; 1986 nian 12 yue [Dec. 1986]-
第1期- ; 1986年12月 [Dec. 1986]-

Taipei : Taiwan shi tian ye yan jiu ji hua gong zuo shi (Zh
台北 : 台灣史田野研究計劃工作室(中央研究院).
ExLibris’s “CJK” efforts as of Mar. 2001

- Designed for Chinese sites
  - Automatic pinyinination
  - Text “segmentation”
  - Chinese Windows required
  - Collation by pinyin
  - In hospitable to Japanese or Korean
  - Not yet a mature product

- Unicode-based – a big plus
Coming to Cambridge

- Harvard scholars’ requirements
  - Truly “CJK”
  - Search traditional & simplified Chinese together
  - Search in original script or romanization
  - Cross-language character search
Coming to Cambridge

- Other development issues
  - Word division
  - Facilitating staff use
    - Retagging 880 fields
  - MARC compatibility
  - Desktop requirements
    - Input methods

- Joint specification - Jan. to Oct. 2001
- Programming Oct. 2001 to Nov. 2002 plus
Results of word search development

- For word searches on CJK characters –
  - Adjacency implied automatically
  - Multilanguage results
    - Hence, no special indexes
  - One search retrieves both simplified and traditional forms
How come implied adjacency?

- Word division issues
  - Utilities’ practices differ
    - RLIN aggregates/segments
    - OCLC does not
  - Harvard chooses not to separate words
    - Reflects the written language
    - fix_doc_delete_chi_spaces
  - Great flexibility for searcher
Results of browse development

- For browses –
- Language-specific indexes
  - Chinese
    - Pinyin order
    - subarranged by Unicode values
      - character by character
  - Japanese and Korean
    - By Unicode values
- Less than ideal
On language-specific CJK browse

- Paradox
  - Other browses not language-specific

- Chinese
  - Like Asian Aleph installations
    - Original script to pinyin dictionary
    - Indexing by automatically-generated pinyin
    - Potentially different from cataloger-input

- Japanese and Korean
  - Analogous treatment in future?
An aside

- HOLLIS language-specific browse for other non-latin scripts?
  - “Han”-based writing systems (CJK)
    - Huge repertoire
    - Many homonymys
    - Divergent sequencing principles
  - Alphabets and syllabaries
    - Small repertoire
    - Divergent sequences, but
      - More like latin-script languages, where English wins
Notes on CJK browsing

- When browsing in the HOLLIS Catalog:
  - CJK browse indexes
    - Enter search *in the original script*
  - CJK in main indexes
    - Enter search in romanized form

- In CJK browse indexes
  - Unicode values distinct for simplified & traditional
  - A mistake?
## Browse index display

### FULL CATALOG - Browse an Alphabetical List

#### Browse List: Titles, All Chinese

<table>
<thead>
<tr>
<th>No. of Recs</th>
<th>Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>台湾乙未战纪</td>
</tr>
<tr>
<td>1</td>
<td>台湾医学五十年. Chinese</td>
</tr>
<tr>
<td>1</td>
<td>台湾艺术散文选</td>
</tr>
<tr>
<td>1</td>
<td>台湾轶事 聶华苓短篇小说集.</td>
</tr>
<tr>
<td>1</td>
<td>台湾赢家秘籍</td>
</tr>
<tr>
<td>2</td>
<td>台湾幽默精选</td>
</tr>
<tr>
<td>1</td>
<td>台湾游记选</td>
</tr>
<tr>
<td>1</td>
<td>台湾邮政光复二十年纪要: 中华邮政七十周年纪念</td>
</tr>
<tr>
<td>1</td>
<td>台湾与海外华人作家小传</td>
</tr>
<tr>
<td>1</td>
<td>台湾与前苏联交往秘录</td>
</tr>
</tbody>
</table>
Record 1 out of 1

Author: Oda, Toshio, 1892-1989.

Title: Taiwan igaku gojūnen. Chinese

Title: Taiwan yi xue 50 nian / Xiaotian Junlang zhu; Hong Youxi yi.

Edition: Xiu ding ban.

Published: Taipei Shi: Qian wei chu ban she, 2000.
MARC21 compatibility issues: “alternative graphic representation”

Paired fields from 880 and mate

- Simpler index construction
- Better display for catalogers
- Maintained as a pair
- Subfield 9 in ex-880
  - Automatically generated
  - Contains a language code from 008 or 041
  - Can be overridden by cataloger
  - Only one subfield 9 allowed per pair
Paired fields in cataloger’s view

01
Oda, Toshio,
1892-1989.

01
小田俊郎，
1892-1989.

02
Taiwan igaku gojūnen.
Chinese

02
台湾医学五十年。
Chinese

03
Taiwan yì xue 50 nian /
Xiaotian Junlang zhu ; Hong Youxi yì.

03
台灣醫學50年 /
小田俊郎著；洪有錫譯。

04
Taiwan yì xue wu shi nian

04
Xiu ding ban.

04
修訂版。

05
Taipei Shi :
Qian wei chu ban she,
2000.

05
台北市：
前衛出版社,
2000.

14-152 p.
MARC21 compatibility issues: “alternative graphic representation”

- Typical p_manage_25 tab_fix group for importing CJK MARC21 records to Aleph

  - fix_doc_delete_chi_spaces modify RLIN-style data
  - fix_doc_880 retag fields
  - fix_doc_sort rearrange fields by tag
  - fix_doc_sort_sub6 subarrange to unite pairs
  - fix_doc_marc21_spaces “standard” blank replacement
  - fix_doc_do_file_08 x.fix other fussing as needed locally, e.g. delete unwanted fields
MARC21 compatibility issues: “alternative graphic representation”

- Exporting CJK MARC21 records from Aleph
  - Variant procedures required depending on the character encoding desired – UTF8 or MARC8.
  - Two new Ex Libris routines required for non-latin export are in hand but not yet tested.

- A tab_fix group for p_print_03 will include
  - `fix_doc_redo_880`  
    - *restore 880 fields*
  - `fix_doc_create_066`  
    - *only for MARC8 output*  
    - *066 not defined in UTF8 records*
MARC21 compatibility issues: Character encoding

- MARC8 EACC and Unicode CJK
  - More variants encoded separately in EACC
  - Harvard’s decision:
    - Go with Unicode
    - Modify Ex Libris CJK conversion table
    - Two EACC values can become one Unicode value
    - Imperfect reversibility
Harvard desktop requirements for CJK

- Staff client for CJK character input
  - Windows 2000 Professional
  - “Language setting for the system” Japanese, Korean, Chinese traditional, Chinese simplified
  - “Input locales” as needed
  - MS Arial Unicode font

- Staff client for view-only CJK
  - Windows 2000 professional or NT 4.0
  - A CJK enabler such as Unionway’s Asian Suite
  - MS Arial Unicode font
Harvard desktop requirements for CJK

- Web Browser /OPAC for all users
  - Windows 2000 or NT 4.0
  - Internet Explorer 5.01 or higher
  - MS Arial Unicode font
- For NT
  - IE Language packs Chinese simplified, Chinese traditional, Japanese, Korean
- For 2000
  - “Language setting for the system” Japanese, Korean, Chinese traditional, Chinese simplified
  - “Input locales” as needed
Things as they are today

- CJK added to .5 million existing records
- In production
  - Cataloging
    - OCLC XPO
    - RLIN PUT
- In testing
  - OCLC batch record import
  - Export of MARC records