Library Catalogs & Non-Roman Scripts
ALA Annual Meeting
Orlando, FL
June 26, 2004
Michael Kaplan
Director, Product Management
Ex Libris

Overview

- Unicode is all inclusive:
  - In practice: CJK, (J)A(CK)PHY+Greek+Cyrillic
  - In practice: CJK, Hebrew, Arabic, some Greek, minimal Cyrillic
  - Some sample views
- Staff and Patron aspects
- Workstation requirements
- Input Method Editor (IME) and cousins; Auto-generation
- Aggregation/Segmentation and non-Segmentation: RLIN and OCLC
- 880 fields vs. Parallel fields in viewing/editing (only) local catalogs
- Role of vernacular dictionaries
- Chinese traditional vs. Chinese simplified
- Indexing standards
  - Fields with mixed character values
- The FONT challenge
Overview 2

- Support for
  - Loading
  - Creation
  - Editing
  - Searching (staff and patron)
  - Display
  - Indexing
  - Export
  - Reports and 3rd party software

Ex Libris and non-Roman scripts

- Handled Arabic, Cyrillic, Greek, Hebrew from the beginning
- CJK added later
- Completely Unicode compliant
- Largest active CJK vernacular implementations in the world
  - National Library of China
  - California Digital Library Melvyl®
  - Harvard University HOLLIS catalog
- Greek: American/British Schools of Classical Studies (Athens)
- Hebrew: Jewish Theological Seminary, Brandeis University, Spertus Institute, Israeli catalogs
### Ex Libris and non-Roman scripts

#### Browse List: System Number

<table>
<thead>
<tr>
<th>System Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000003264</td>
<td>ドラマリーの冒険への道</td>
</tr>
<tr>
<td>0000003265</td>
<td>يوروساوي</td>
</tr>
<tr>
<td>0000003266</td>
<td>هذه الكتب صنع الفنانيين</td>
</tr>
<tr>
<td>0000003267</td>
<td>حالة الوطن</td>
</tr>
<tr>
<td>0000003268</td>
<td>Ουκρανία</td>
</tr>
<tr>
<td>0000003269</td>
<td>宝石堆山</td>
</tr>
<tr>
<td>0000003270</td>
<td>迷路的礼花</td>
</tr>
<tr>
<td>0000003271</td>
<td>真愛無處</td>
</tr>
<tr>
<td>0000003272</td>
<td>1000年復活のサルベージベア</td>
</tr>
<tr>
<td>0000003273</td>
<td>110年郊外の真誠なミシュランベーシック</td>
</tr>
</tbody>
</table>

---

### National Library of China ALEPH 500
**National Library of China ALEPH 500: Basic search**

<table>
<thead>
<tr>
<th>多种检索</th>
<th>多种检索</th>
<th>多种检索</th>
<th>多种检索</th>
<th>多种检索</th>
<th>多种检索</th>
<th>通用语言（CCL）</th>
</tr>
</thead>
<tbody>
<tr>
<td>快速检索</td>
<td>快速检索</td>
<td>快速检索</td>
<td>快速检索</td>
<td>快速检索</td>
<td>快速检索</td>
<td>通用语言（CCL）</td>
</tr>
</tbody>
</table>

**Advanced search**

<table>
<thead>
<tr>
<th>多种检索</th>
<th>多种检索</th>
<th>多种检索</th>
<th>多种检索</th>
<th>多种检索</th>
<th>多种检索</th>
<th>通用语言（CCL）</th>
</tr>
</thead>
<tbody>
<tr>
<td>快速检索</td>
<td>快速检索</td>
<td>快速检索</td>
<td>快速检索</td>
<td>快速检索</td>
<td>快速检索</td>
<td>通用语言（CCL）</td>
</tr>
</tbody>
</table>

---

**National Library of China ALEPH 500: Basic search**

**Advanced search**
National Library of China ALEPH 500: Record view

Chinese with Parallel fields (Demo)
Japanese (Brief records): Melvyl™

1. Full Record
   Title: Nihon toshi bukai seisaku shi no kiroku / hentei toshi bukai Anrei shitsu Hohushi
   Format: Journal
   Library: UCB BURLU UCI

2. Full Record
   Uniform Title: Nihon toshi shi no kiroku (San Francisco, Calif.)
   Title: Nihon toshi shi no kiroku.
   Publisher: San Francisco, Calif.: J-Town Collective.
   Format: Journal
   Library: UCB BURLU UCI UCLU UO UCD

3. Full Record
   Title: Shonan bungaku.
   Publisher: Tokyo: Shonan bungaku Bungakudai 1997-
   Format: Journal
   Library: UCB BURLU UCI UCLU UO UCD

4. Full Record
   Title: Yurika no Einka.
   Publisher: Tokyo: Sekishita.
   Format: Journal
   Library: UCB BURLU UCI UCLU UO UCD

Chinese (Brief records): Melvyl™

2. Full Record
   Title: Xi Dong xue = Xiongguo / Liang Shuhai / Zhi bi nan
   Publisher: Shanghai: Shanghai ci liu chu bian she, 2001.
   Format: Journal
   Library: UCLU

3. Full Record
   Title: Xia shu yan bao.
   Publisher: Tianjin: Sui shu yan bao chuan shu, 2003.
   Format: Journal
   Library: UCLU

4. Full Record
   Title: Hua xi cai shang jia.
   Publisher: Taipei Shi: Zhongguo gu min ding dang zhong yang wei yuan hui shi shi hao liu hao buan zuan wei yuan hui, Ningguo
   Series: Zhonghua nihon gai shang song luan, AS 1-5 2.
   Format: Journal
   Library: UCLU
Item 4 of 2868 Total
Return to Search Results List

Title: 湖北省微生物

Publisher: 华中师范大学出版社

Description: 2 v. : 22 cm.


Series: 现代生物学与可持续发展

Language: Chinese

Subject: 中国学生

Added Entry: 现代生物学与可持续发展

Succeeding Entry: 未找到

Format: Journal

Library: US Library

Call Number: 未找到

Availability: 未找到

Notes: 未找到

Specialized ALEPH CJK developments

- One search retrieves all languages
- Nonetheless, specialized searches provided for browse search of vernacular CJK authors/titles
- One search retrieves both traditional and simplified Chinese
- Adjacency always implied
- Chinese vernacular search can run off Roman input
- Tools to load and normalize RLIN and OCLC variant formats
- Import & create parallel fields from 880s
- Export with parallel fields as 880s
CJK indexes & help: Harvard University

Library Catalogs & Non-Roman Scripts

CJK indexes/Roman search of vernacular index: Harvard University

www.ala.org/alcts
CJK full record: Harvard University

CJK vernacular search against specific Chinese index: Harvard University
CJK vernacular search against specific Korean index: Harvard University

Arabic (Brief & Full): Harvard University
Greek: AMBROSIA (Athens): Brief results

Results for Θουκιδίδης.
Sorted by: Title, then Year.
Next options: «Previous Authors (A-Z)» Author/Year (A-Z) Title/Year (A-Z)

Records 1 - 10 of 14

<table>
<thead>
<tr>
<th>#</th>
<th>Author</th>
<th>Title</th>
<th>Year</th>
<th>Life/Death</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Παντοκράτωρ Καπρικιών</td>
<td>Γεωργιάδης, Θ.</td>
<td>1982</td>
<td></td>
<td>BB Book</td>
</tr>
<tr>
<td>2</td>
<td>Τσιγκριτζ, Λ.</td>
<td>Σύντομος Μνημειώδης Βιβλίο</td>
<td>1982</td>
<td></td>
<td>BB Book</td>
</tr>
<tr>
<td>3</td>
<td>Τσιγκριτζ, Λ.</td>
<td>Εικονογράφος Περιοδικών Ιστορικών Επιστημών</td>
<td>1982</td>
<td></td>
<td>BB Book</td>
</tr>
<tr>
<td>4</td>
<td>Ερμούττο, Κ.</td>
<td>Η Ιστορία της Βιβλιοθήκης στην Ακρόπολη</td>
<td>1999</td>
<td></td>
<td>BB Book</td>
</tr>
<tr>
<td>5</td>
<td>Τσιγκριτζ, Λ.</td>
<td>Αυτόνομος Περιοδικός Επιστημών Επιστημών</td>
<td>1985</td>
<td></td>
<td>BB Book</td>
</tr>
<tr>
<td>6</td>
<td>Τσιγκριτζ, Λ.</td>
<td>Αυτόνομος Περιοδικός Επιστημών Επιστημών</td>
<td>1985</td>
<td></td>
<td>BB Book</td>
</tr>
<tr>
<td>7</td>
<td>Μαρμάρο, Ε.</td>
<td>Εικονογράφος Περιοδικών Ιστορικών Επιστημών</td>
<td>1985</td>
<td></td>
<td>BB Book</td>
</tr>
<tr>
<td>8</td>
<td>Στεφανίδη, Κ.</td>
<td>Περιοδικά και Επιστημονικά Μαρτυρία</td>
<td>1983</td>
<td></td>
<td>BB Book</td>
</tr>
</tbody>
</table>

Greek: AMBROSIA (Athens): Author browse

ASCS/BSA Catalogue
Browse an Alphabetical Index

Browse for: 
Select field to browse: Author (but same field) 

Browse List: Authors

<table>
<thead>
<tr>
<th>No. of Recs</th>
<th>Title</th>
<th>Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.</td>
<td>Μαρμάρο, Ε</td>
<td>[Author's Record]</td>
</tr>
<tr>
<td>1.</td>
<td>Στεφανίδη, Κ</td>
<td>[Author's Record]</td>
</tr>
<tr>
<td>9.</td>
<td>Στεφανίδη, Κ</td>
<td>[Author's Record]</td>
</tr>
<tr>
<td>2.</td>
<td>Στεφανίδη, Κ</td>
<td>[Author's Record]</td>
</tr>
<tr>
<td>1.</td>
<td>Στεφανίδη, Κ</td>
<td>[Author's Record]</td>
</tr>
</tbody>
</table>
### Harvard/CDL: MARC view: Parallel fields

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FMT</td>
<td>BK</td>
</tr>
<tr>
<td>LDR</td>
<td>008f0000 200202 a 4000 001 087800300 005 2002021202220 003 087800300 kar d 0300</td>
</tr>
<tr>
<td>040</td>
<td>DOD0022222</td>
</tr>
<tr>
<td>085</td>
<td>001 087800300 kar 600 0 kar d 0851 001 087800300 kar 600 0 kar d</td>
</tr>
</tbody>
</table>

### Staff and Patrons

- Non-Roman requires more workstation components, such as IME and non-Roman installed fonts
- Libraries can control only those workstations that are in their domain
- Technical discussions: Library staff are frequently either
  - Not conversant with non-Roman scripts/issues
  - Or
  - Not conversant with required technology
- Discussions on library level of issues are extraordinarily complex and difficult because of lack of overarching expertise
- Patrons often come to the library now with greater understanding of non-Roman issues and technology than Staff (e.g., use of IME)
Non-Roman “families”

- Too often non-Roman is taken to mean CJK: CJK is more difficult in many respects but not the only group
- Many more CJK examples in our online catalogs
- CJK has greater “need”: Romanization too imprecise
  - Vernacular is extremely important to disambiguate headings that are all identical in their Romanized display
- CJK: non-alphabetic
  - Keyboard input: extensive character repertoire (TransTech)
  - Unicode value input, IME input, Keyboard Roman input
- Non-CJK (Arabic, Greek, Hebrew, Yiddish, Persian, etc.): alphabetic: Bi-directional aspects
  - Limited character repertoire but some with special requirements
  - Keyboard mapping; Floating keyboards

Workstation requirements

- Changes over the last 5 years
  - Move from Chinese Windows
  - Windows 2000 and Windows XP provide more robust and built-in support
  - IME bundled in
  - Other IME options available
  - Prior requirements for higher-end workstations now largely passé
  - 2001 Configuration now eclipsed by current workstation configurations: any high-end machine should be sufficient (memory requirements for Font)
**Input Method Editor (2001)**

- IME is an operating system add-on, available from various software vendors, that serves as an aid for input of East Asian text
- IME is not a keyboard—it is a Romanization mechanism
- IME integral part of Windows 2000+ (and CJK versions of Windows 98/NT [no longer supported by Microsoft])
- Other options, at one time under consideration, included
  - Asian Suite-NJStar
- For Chinese, Ex Libris also supports use of Romanized input to search vernacular indexes by means of machine-generated lookup in Chinese dictionary

---

**Aggregation and Segmentation**

- RLIN needed aggregation because of the problem of indexing the Romanized forms of CJK characters
- Aggregator for Romanized fields used in RLIN is proprietary
- On export, RLIN inserts a space to indicate word breaking by inserting a blank space
- OCLC does not keep spaces in any CJK vernacular field. If incoming records have vernacular aggregation, it is removed from the records
  - (“They [OCLC] think it looks better”)
Aggregation v. Segmentation implications

CJK records from RLIN and OCLC have different treatment in the vernacular fields: RLIN records have aggregation in the vernacular, OCLC records have the characters strung together with no spaces

This poses significant issues for building a merged catalog based on records from the two fundamental sources

Role of Chinese dictionaries in ALEPH

- Chinese dictionaries in use:
  - One consists of 300,000 words for use in segmentation (Chinese-Pinyin): determine words
  - One consists of Chinese characters in Unicode with Pinyin and stroke string: automatic “translation”
  - And Traditional to Simplified mapping to enable searching of Simplified to also retrieve Traditional
    - Not a mainland problem (Simplified only in use)
    - Research library problem
- Nothing comparable for Japanese or Korean
Indexing standards

- Indexing for non-Roman alphabets (!) is based on character mapping and well-understood Romanization tables
- Not so for non-alphabetic languages (CJK)
- Indexing and ordering must be predictable and intuitive to users
- Display must likewise be clear
- Single index
- At Harvard, separate browse index for each of the 3 CJK languages

Indexing standards: Chinese

- Ex Libris indexes based on Pinyin value of the characters, and in case of duplicate Pinyin values they subsort on Unicode values
- In support of ability to search Chinese vernacular from Romanized input search of Chinese indexes, Chinese index is ordered by machine-generated Pinyin values
  - Potential conflict with cataloger-created Pinyin
  - Nonetheless, always predictable
Indexing standards: Japanese and Korean

- Because of character set issues that cut across Chinese, Japanese, and Korean—and because of obvious sensitivities—it is not desirable to use Pinyin Romanization for Japanese and Korean characters.
- Unicode for now (accepted by communities and seems to be well-understood).
- Plan for overall indexing based on Stroke Count.

Indexing standards

- Choice of language for determining which language indexes should receive vernacular headings from CJK records will be driven by the language code in the 008/041; that is, at the document level.
  - While this has some potential problems, there is no other ready solution.
- By this means browse searches will be directed explicitly to Chinese, Japanese and Korean indexes. The Web OPAC screen offers choices for vernacular browse searching (for example):
  - Names (including subjects) in Chinese
  - Titles in Chinese
  - Names (including subjects) in Japanese
  - Titles in Japanese
  - Names (including subjects) in Korean
  - Titles in Korean
**Indexing standards 2**

- Database merged of OCLC and RLIN records posed special challenges
  - At Harvard, aggregator symbols were stripped as a means of handling a database of merged OCLC and RLIN records
  - Any spaces in vernacular fields were stripped out
  - Use of these choices for browse searching forces the search automatically to be considered a phrase with forced adjacency, regardless of any spaces that may be in the search string

**Indexing standards 3**

- An ideal solution for all languages would be to base the indexing on ‘stroke’ count
  - Placeholder now present in ALEPH 16
- An open and intractable question is what to do about fields that internally contain characters from more than one character set
The Font problem

- Serious lack of suitable, comprehensive, and freely available Unicode font, for both Staff and Public users

Reports and 3rd party software

- Reporting against databases that contain non-Roman is complicated where 3rd party reporting packages (Crystal Reports, Brio, etc.) are concerned
  - Non-Roman data ‘infiltrates’ administrative data
  - Reporting on bibliographic data of course involves reporting on non-Roman data
  - Not all 3rd party tools are Unicode compliant
  - Of those that are, it is often only the (expensive) Enterprise edition that supports Unicode
- This is an evolving situation that will no doubt improve as Unicode support becomes mandatory for these vendors