# Adding Non-Latin Data to Aleph: a status report

Prepared for presentation to the
North American Aleph Users' Group
15 June 2004

Charles Husbands
Harvard University Library
Office for Information Systems
charles_husbands@harvard.edu

# Caveat Auditor

Two cautions:

- This *is* a status report
  - Development is ongoing
    - Version to version
    - Day to day
- I'm reporting from my experience which is, in effect, Harvard's experience
  - I may confuse what is Aleph with our implementation decisions.
  - You may not come to the same decisions

# Scripts in Unicode  - Overview

- The Unicode Basic Multilingual Plane
  - Holds what can be encoded in a 16-bit space
    - Capacity ca. 65,000 characters
    - Ca. 52,000 assigned
  - Houses most modern scripts
  - Additions continue to be made
    - Living scripts, obscure or poorly codified
  - “Unified Han” made initial CJK implementation possible in the BMP

# Scripts in Unicode 3.1 – 4.0

- More 64k-char. planes open for use now
  - Supplementary Multilingual Plane
    - Ca.1600 assignments
  - Supplementary Ideographic Plane
    - Ca. 43,000 assignments
  - Supplementary Special Purpose Plane
    - Ca. 100 assignments
  - The supplementary planes require more than 16-bits to encode a character
    - UTF-8 and UTF-16 still work at those altitudes.

# Limitations and Qualifications

- Not all software in our “village” can support characters above the BMP.
  - Programs frequently assume 16-bit representation.
- Practically, only CJK is affected.
  - But few of these characters, if any, are known to MARC-8

# MARC-8 CJK vs. Unicode

- MARC-8 uses 24-bit East Asian Character Code
- EACC relative to Unicode
  - Has characters that Unicode does not
    - Variant forms
      - Mapped to values in BMP private use area
    - Characters “missed” by Unicode
      - Mapped to values in BMP private use area
      - Some have since been included in Unicode
  - Mapping done by special MARBI task force
    - Reproduced in vanilla Aleph marc8_eacc_to_unicode

# CJK MARC8_TO_UTF

- Harvard's marc8_eacc_to_unicode
  - EACC characters that Unicode does not have
    - Variant forms
      - Mapped to values of the primary form
    - Characters "missed" by Unicode
      - Mapped to U+3013 "GETA"
      - As characters are added to Unicode
        - Convert to true U+ value if in BMP
        - Leave as GETA otherwise

# CJK UTF_TO_MARC8

- Harvard’s marc8_eacc_to_unicode
  - Brings primary forms together at top of table
    - They will be the ones preferred for output
  - This makes round trip mapping fail
    - But abandonment of private use area is finding favor elsewhere, at least in talk.

# CJK word indexing (Harvard implementation)

- CJK in HOLLIS since late 2002
- Version 15
  - Searching requires no special separate indexes.
  - One search retrieves all languages
  - One search retrieves traditional and simplified
  - Adjacency implied
- Version 16
  - As above, but
  - Adjacency implied – results slightly inferior
    - A configuration issue?

# CJK heading indexing (Harvard implementation)

- Version 15
  - Searching uses language-specific indexes
    - Japanese and Korean arranged by Unicode value
    - Chinese arranged by pinyin subarranged by Unicode
      - Simplified and traditional can get separated
- Version 16
  - As above, but
  - A new stroke-count filing routine is available
    - Could be interesting.  Harvard has not yet tested.

# Hebrew and Arabic

- Some features in common
  - Bidirectional writing, basically right to left
  - Grammatical particles prefixed to words
    - Definite article
    - Prepositions
    - Others

| | |
|---|---|
| bayna al-taʾlīf wa-al-tazyīf | بين التأليف و التزييف |
| ha-nimtsaʾim bi-teshuvot | הנמצאים בתשובות |

# Hebrew and Arabic

- Special word indexing requirements
  - Leading wild card to bypass prefixed particles
    - In addition to trailing or imbedded wild card
    - Not working well in HOLLIS, but okay in Israel
      - Configuration issue?
  - Combine Hebrew regular and final character forms?
    - Desirability uncertain, feasibility lacking

# Bidirectional input issues

- Pay attention to Windows locales
  - Characters on keyboard are easy
  - Others are not, especially for OPAC users
- Cursor movement can be confusing
  - Can switch field direction in Aleph client
  - OPAC users have it tough again
    - Copy and paste can solve some problems

# Cyrillic and Greek

- Not much testing done on these yet.
- Note that Greek will always be treated by Aleph as Greek
  - The so-called Greek “symbols” in MARC-8 latin contexts cannot be distinguished from real alpha, beta, gamma letters in Unicode.

# Bringing in MARC-8 non-latin

- Convert character encoding
- Squeeze out CJK inter-word spacing
  - OCLC convention preferred to RLIN
- Convert 880s to corresponding tags
  - Converted CJK fields get virtual $$9 for language
    - Used for heading indexing
    - Automatic generation from 008 or 041
      - Cataloger can override later if necessary
- Sort fields
  - Take account of $$6

# Bringing in MARC-8 non-latin

- A tab_fix excerpt

```
OCLB1  fix_doc_delete_chi_spaces
OCLB1  fix_doc_880
OCLB1  fix_doc_sort
OCLB1  fix_doc_sort_sub6
```

- We do this in all incoming record fixes
  - Does no harm to all-latin records

# Still a few bugs in the system

- Importing MARC-8 records
  - Character conversion
    - marc8_ara_to_unicode, marc8_rus_to_unicode need to be checked.
      - There should be separate tables for the extended sets with MARC-8 values reduced from the A0-FF range to the 20-7F range.
      - In the basic tables any MARC-8 values above 7F should be removed.
    - Hebrew and Arabic combining marks are not repositioned to follow their base characters

# Sending out MARC-8 non-latin

- Convert character encoding
- Retag 880s
- Construct 066
- Clean up
- Sort fields

# Sending out MARC-8 non-latin

- A tab_fix excerpt:

```
E880  fix_doc_redo_880
E880  fix_doc_create_066
E880  fix_doc_do_file_08            e880.fix
E880  fix_doc_delete_empty
E880  fix_doc_space_char
E880  fix_doc_sort
```

- e880.fix
  - Removes cataloger-inserted $$9 (non-latin specific task)
  - Insures LDR byte 09 is a space
  - Deletes technique-1 escape sequences from the 066
- Delete_empty and space_char do not refer specifically to non-latin.

# Still a few bugs in the system

- Exporting non-latin MARC-8 records
  - Character conversion
    - Some characters get mangled
      - Numerous CJK
      - One rare Greek
      - One Extended Arabic
  - 066 construction
    - Some MARC-8 escapes not provided for
      - Extended Arabic
      - Extended Cyrillic

# Tomorrow the world?

- Must have more generally supported UTF-8 exchange.
- Must deal with non-MARC-8 scripts in continuing MARC-8 exchange.
- These are not insurmountable but they need work on several levels.
  - Standards or conventions
  - Modification of local processes.