True Scripts in Library Catalogs – The Way Forward

Joan M. Aliprand
Senior Analyst, RLG

© 2004 RLG

Why the current limitation?

• Coordination!
• As complex as Format Integration!
Script Capability and Data Exchange

Before Unicode

- **JACKPHY** = East Asian scripts (CJK), Arabic, and Hebrew
- **CJK** = scripts for Chinese, Japanese, Korean
- **HAPY** = scripts for Hebrew, Arabic, Persian, Yiddish

Script Capability and Data Exchange

With unrestrained Unicode

- **JACKPHY** = East Asian scripts (CJK), Arabic, and Hebrew
- **CJK** = scripts for Chinese, Japanese, Korean
- **HAPY** = scripts for Hebrew, Arabic, Persian, Yiddish
Script Capability and Data Exchange

With unrestrained Unicode

System U+

Unicode capable

Latin data

> CJK data

> HAPY data

Other non-Roman data

System A

JACKPHY capable

>…data – Outside the scope of the individual MARC 21 sets

☒ – Can only use the romanized data in the record

CHARACTERS OF THE MARC 21 RECORD

"MARC-8"

• Latin character sets (default)
  • Basic Latin “ASCII” (US standard ANSI X3.4)
  • Extended Latin “ANSEL” (ANSI/NISO Z39.47)
  • Euro sign € and Eszett ß for UKMARC alignment

• Supplementary character sets
  • Superscripts
  • Subscripts
  • Greek symbols α β γ

• Non-Roman character sets
  • Arabic, Cyrillic, EACC (CJK), Greek, Hebrew
CHARACTERS OF THE MARC 21 RECORD

Use of Unicode

- UTF-8 encoding form
- Designation of Unicode as encoding
  - Leader position 09, Character coding scheme
    - Single character set, so no 066 field, Character Sets Present
- Initial limitation to MARC-8 character repertoire
  - MARBI Proposal 98-18
- Exception for Canadian Aboriginal Syllabics
  - MARBI Proposal 2002-11
    - Encoded only as Unicode characters
      - Addition to MARC-8 sets not requested

Character Sets in the MARC 21 Record

<table>
<thead>
<tr>
<th>Record Component</th>
<th>Character Sets Allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leader</td>
<td>ASCII graphic characters</td>
</tr>
<tr>
<td>Directory</td>
<td>ASCII numeric characters</td>
</tr>
<tr>
<td>Control Fields (00X)</td>
<td>ASCII</td>
</tr>
<tr>
<td>Data Fields (except 880)</td>
<td>Basic Latin, Extended Latin plus € and ß, Superscripts, Subscripts, Greek symbols</td>
</tr>
<tr>
<td>Alternate Graphic Representation (880)</td>
<td>At least one non-Roman character set required; ASCII</td>
</tr>
</tbody>
</table>
**Unicode in the MARC 21 Record**

<table>
<thead>
<tr>
<th>Record Component</th>
<th>MARC-8 Character Sets</th>
<th>Unicode (UTF-8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leader</td>
<td>ASCII graphic characters</td>
<td>ASCII graphic characters</td>
</tr>
<tr>
<td>Directory</td>
<td>ASCII numeric characters</td>
<td>ASCII numeric characters</td>
</tr>
<tr>
<td>Control Fields (00X)</td>
<td>ASCII</td>
<td>ASCII graphic characters</td>
</tr>
<tr>
<td>Data Fields (except 880)</td>
<td>Basic Latin, Extended Latin plus € and ß, Superscripts, Subscripts, Greek symbols</td>
<td>Latin</td>
</tr>
<tr>
<td>Alternate Graphic Representation (880)</td>
<td>All</td>
<td>At least one Script other than Latin; ASCII characters</td>
</tr>
</tbody>
</table>

**Accented Letters?**

- Two alternatives:
  - Strictly decomposed only?
    - Base letter followed by combining characters
  - Both composite characters and combined character sequences permitted
- Use Unicode-defined canonical equivalence and canonical ordering
The Order of Data

• Sorting; Filing order; Collation
• Not only for presentation of results
  – Essential part of query matching
• Sorting by character code?
  – A primitive approach
  – **DOES NOT WORK FOR UNICODE**

Unicode Collation Algorithm

• Default order for all characters of each script
• Four levels of differences
  • Diacritics significant at Secondary level
  • Case significant at Tertiary level
• “Tailoring” methodology provided
  – For alphabetical order of a particular language
  – For library practices
• Does not do everything
  – Separate character folding needed in some cases
    • Medial, final forms in Hebrew; CJK ideographic variants
True Scripts in Authority Records

One of the factors that continues to delay the implementation of original scripts in NACO authority data is that not all places that support copies of NAF for input/update of NACO records can yet cope with all of the original scripts that might be included in these records. In order to maintain virtually identical copies of NAF, users are not allowed to create headings with original scripts that might be invisible or deleted in other copies of the file.

– Ed Glazier, RLG

Script Capability and Authority Records

<table>
<thead>
<tr>
<th>Script</th>
<th>LC</th>
<th>RLG</th>
<th>OCLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CJK</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cyrillic</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
MARC 21 Format for Authority Records

- Data elements for non-Roman scripts
  - Same as in Format for Bibliographic Records
    - Alternate graphic representation (880 field)
    - Character sets specified (066 field)
    - Subfield 6 for field linkage, directionality
  - Assumption of 1:1 field equivalence
  - Not yet implemented in NAF, SAF

- Linkage between established headings
  - 7XX Heading Linking Entry fields

Excerpt from NAF Record

<table>
<thead>
<tr>
<th>Field</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>110 20 ‡a United Nations</td>
<td>English</td>
</tr>
<tr>
<td>410 20 ‡a Nations Unies</td>
<td>French</td>
</tr>
<tr>
<td>410 20 ‡a Naciones Unidas</td>
<td>Spanish</td>
</tr>
<tr>
<td>410 20 ‡a Organizatsia Ob edinennykh Natsii</td>
<td>Russian (ALA-LC romanization)</td>
</tr>
<tr>
<td>410 20 ‡a Lien he guo</td>
<td>Chinese in pinyin (ALA-LC romanization)</td>
</tr>
<tr>
<td>410 20 ‡a Lien ho kuo</td>
<td>Chinese (Wade-Giles romanization)</td>
</tr>
<tr>
<td>410 20 ‡a Umam al-Muttaḥidah</td>
<td>Arabic (ALA-LC romanization)</td>
</tr>
</tbody>
</table>
Excerpt with True Scripts

110 20 ‡a United Nations
410 20 ‡a Nations Unies
410 20 ‡a Naciones Unidas
410 20 ‡a Organizatsii Ob’edinennykh Natsii
410 20 ‡a Lien he guo
410 20 ‡a Lien ho kuo
410 20 ‡a Umam al-Muttaḥidah
880 20 ‡6 410-00 ‡a Организация Объединенных Наций
880 20 ‡6 410-00 ‡a 组织和联合
880 20 ‡6 410-00/r ‡a الأمم المتحدة
Linkage within Authority Records

- Romanized/non-Roman field pairs?
  - Need to see all data in authority record
  - Only one source of name or subject authority for 1XX
  - 1:many , many:1 relationships in authorities

→ UNLINKED 880 FIELDS ←
Linkage within Authority Records

- Romanized/non-Roman field pairs?
  - Need to see all data in authority record
  - Only one source of name or subject authority for 1XX
  - 1:many , many:1 relationships in authorities
    → UNLINKED 880 FIELDS ←

- Any need to use 880 fields?
  - Substitution of data as for bibliographic display does not apply
  - To simplify checking of field contents?
  - To simplify implementation?

Characters in Authority Records

Rules for 13 Characters as of 01/06/2001

<table>
<thead>
<tr>
<th>Character</th>
<th>Rule for Use in Authority Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spacing circumflex</td>
<td>Substitute %5E for now</td>
</tr>
<tr>
<td>Spacing underscore</td>
<td>Substitute %5F for now</td>
</tr>
<tr>
<td>Spacing grave</td>
<td>Should not occur</td>
</tr>
<tr>
<td>Curly brackets (open &amp; close)</td>
<td>Use</td>
</tr>
<tr>
<td>Spacing tilde</td>
<td>Substitute %7E for now</td>
</tr>
<tr>
<td>Degree sign</td>
<td>Substitute superscript 0 for now</td>
</tr>
<tr>
<td>Lowercase script l</td>
<td>Pass through; do not actively supply</td>
</tr>
<tr>
<td>Phono copyright mark</td>
<td>Pass through; do not actively supply</td>
</tr>
<tr>
<td>Copyright mark</td>
<td>Pass through; do not actively supply</td>
</tr>
<tr>
<td>Sharp</td>
<td>Substitute number sign for now</td>
</tr>
<tr>
<td>Inverted question mark</td>
<td>Use when available</td>
</tr>
<tr>
<td>Inverted exclamation mark</td>
<td>Use when available</td>
</tr>
</tbody>
</table>
### Reality Check

**“Unicode” is not a magic word**

- Shorthand for The Unicode Standard
- *Potential* for multiscr ipt support
- Unicode is not software
  - Underlying software comes from OEMs
- Unicode is the foundation, not the whole edifice
  - Many parts to a system
Corporate Members

Adobe Systems; Agence Intergouvernementale de la Francophonie; Apple Computer; Basis Technology; Hewlett-Packard; IBM; India (Ministry of Information Technology); Justsystem; Microsoft; Oracle; Pakistan (National Language Authority); PeopleSoft; RLG; SAP; Sun; Sybase

Associate Members

33 Associate Members including:
The Church of Jesus Christ of Latter-day Saints; Columbia University; Endeavor Information Systems; Ex Libris; Innovative Interfaces; The Library Corporation; OCLC; SIL International; SIRSI; VTLS