What is MARC?

- MAnchine-Readable Cataloging
- MARC is an exchange format
- Focus on MARC 21 exchange format An implementation may disguise certain features of MARC
  - Parallel fields in RLIN are the regular field and its 880 equivalent in MARC
MARC-8
The Individual Character
Sets of MARC 21

Scholarship in all languages
למדוותר בכל שפה
Наука на Всех Языках
علم في كل لغة
所有語言的學術信息
世町全言語による学術情報
학문을 모든 언어로

plus Greek, Superscripts, Subscripts, and Greek Symbols

The Alternative!
MARC-8 vs. Unicode

245 00 <HBR>qcx dbcd yl
tgq<LAT> =‡b<CYR>aGA
DA [EL pesah]<LAT> =
Hebrew-Russian Hagga
dah for Passover.<EOF>
EOF is End Of Field

 Eggs Haggadah for Passover.

Encoding Forms of Unicode

• UTF-8
  – Characters encoded as ASCII-compatible sequences of 8-bit bytes
  – Chosen for MARC 21 record exchange
• UTF-16
  – Characters encoded in 16 bits
  – Surrogate technique for characters beyond Basic Multilingual Plane (BMP)
• UTF-32
  – 1-to-1 relationship between encoded character and code unit
• ALL are Unicode!
• ALL Unicode characters can be represented in every encoding form!
UTF-8

<table>
<thead>
<tr>
<th>Size in Octets</th>
<th>Coverage</th>
</tr>
</thead>
</table>
| 1             | ASCII repertoire  
A-Z, a-z, 0-9, punctuation, # % & * + < > @ … |
| 2             | Latin script extensions, other European and Middle Eastern scripts, IPA extensions, combining diacritical marks, … |
| 3             | South, Southeast and East Asian scripts, common symbols, … |
| 4             | Historic scripts, symbols for music and math, additional ideographs, … |

“octet” = 8-bits

MARC 21 Alternatives

MARC-8
- Leader 09 blank
- Discrete 8-bit character sets
- “Announcers” for character sets
- Non-default sets listed in 066 field
- 880 fields for “alternate graphic representation”
  - Linked / unlinked
- Non-spacing marks before base character
MARC 21 Alternatives

**MARC-8**
- Leader 09 blank
- Discrete 8-bit character sets
- “Announcers” for character sets
- Non-default sets listed in 066 field
- 880 fields for “alternate graphic representation”
  - Linked / unlinked
- Non-spacing marks **before** base character

**Unicode**
- Leader 09 = “a”
- UTF-8 encoding throughout
- No “announcers”
- No 066 field
- 880 fields for “alternate graphic representation”
  - Linked / unlinked
- Non-spacing marks **after** base character

RLIN Differences

**RLIN® ITPS**
- Field pairs
- Linked / unlinked
  - “Core fields”
- Non-spacing marks before base character
- Numeric data in “visual order”
- Proprietary fonts and keyboards
- See all scripts or only romanization
- Abbreviations on displays for scripts in record
**RLIN Differences**

**RLIN® ITPS**
- Field pairs
- Linked / unlinked
  - “Core fields”
- Non-spacing marks before base character
- Numeric data in “visual order”
- Proprietary fonts and keyboards
- See true scripts or romanization
- Abbreviations on displays for scripts in record

**RLIN21™ client**
- Field pairs
- Linked / unlinked
  - “Core fields”
- Non-spacing marks after base character
- Regular order
  - Highest no. first
- Regular fonts, RLG keyboard
- See all scripts
  - No toggling
- “scripts” indicates non-Roman data in record

**Non-Spacing Marks**

- Combining characters come after the base character in Unicode
  - May affect non-filing indicator count
  - L’êtoile (3) versus L’e `toile (2)
- Specified order in Unicode for combining characters
  - Affects multiple diacritics that interact
  - Do you ever include nekudot?
Non-Filing Indicators

- Affected by different order of base character and non-spacing marks
  - Known case of article in Greek
- Romanized alif and ayn
  - Symbols representing letters
  - Treated as non-spacing diacritics
    - Different non-filing counts for romanized Hebrew or Yiddish and the same text in original script
  - OCLC and RLG objections to current treatment not accepted by LC
  - Currently disregarded in searching throughout text

European Numbers

“Arabic” numbers (= from the Arabs)

**RLIN Terminal for Windows**

- Basic Latin character set
  - Most significant digit first
    - 2, 0, 0, 4
- Hebrew character set
  - Least significant digit first
    - 4, 0, 0, 2

**RLIN21 Editor & RLIN21 searching**

- Most significant digit first
  - 2, 0, 0, 4
Traditional Arabic Numbers

“Hindi” numbers (= from the Indians)

RLIN Terminal for Windows

• Basic Arabic character set
  – Least significant digit first
    • ١, ٢, ٣, ٤

RLIN21 Editor & RLIN21 searching

• Most significant digit first
  • ٤, ٣, ٢, ١

Date Spans

• Multi-digit numbers entered most significant digit first
• In RTL field, earlier date is on the right
  – RTL overall, with dates LTR
    • Example: 1967-1948
  – Same with Latin hyphen or with Hebrew makef
• For reverse order, special Unicode formatting characters required
  • Example: 1948-1967
  – Not preserved in current RLG Union Catalog
  – Input record using RTFW if this order required
  – Will be supported post-migration
Headings

• Name headings established under AACR2
  – Communally created file
  – Hebrew funnel
• LCSH
  – External contributions via SACO
• Names, subjects in Hebrew script languages
  – Uncontrolled
  – Locally controlled

Authority Record Options

• AACR2 record with cross-references from ALL forms of name in other scripts
• An authority file for each language/source of authority with authorized headings for the same entity providing connections to other authority files
  – VIAF model
  – YIVO card catalog conceptually
MARC Authority Format

- Hospitable to either option
- Alternate graphic representation
  - Specified
  - Not implemented in NAF
  - Hong Kong has authority file of Chinese names written in ideographs; uses MARC 21

- No linking of romanized and 880 fields!!
Gazing into the Future

• Will Alternate Graphic Representation persist?
  – Needed while Roman-only systems exist

Gazing into the Future

• Hebrew headings in bibliographic records?
  – Are Hebrew cross-references in NAF sufficient for access? (Only AACR2 headings in records)
  – If not, why not?
Gazing into the Future

• A Hebrew NAF?
  – Language, not script, is fundamental: Hebrew language file, Yiddish language file, …
  – What is the source of cataloging authority for each file?
  – Coordination with Israel and other centers of Hebraica for Hebrew language file?
  – Decision about alternative orthographies (Hebrew) or requirement for use of standard orthography (Yiddish)

Gazing into the Future

• There is a true script catalog in your future …
  – What you need to know
  – What you need to consider
Planning for a Script

• Adequate character repertoire?
• Need for a rendering algorithm?
  – Unicode Bidirectional Algorithm
• Word definition in languages?
• Special requirements for indexing?
• Defined sort order?
• Adequate fonts?

Making the Case

• Why should management spend money on your script?
• “Isn’t romanization good enough?”
  – Examples of problems
“Conformance”

• A very special meaning in the context of Unicode
  – Specified in Chapter 3
• Not a general synonym for “makes use of Unicode”
• Pay attention to how it’s used

Testing
(The Skill of the Black Thumb)

• Work from planned test scripts
• Verify every character, not just some
  – Check treatment of spacing “marks” in conversion to Unicode (they are not diacritics!)
• Don’t test just the obvious
• Test negative conditions as well as positive ones
• Stress the system with evil “corner cases”
Reality Check

• Unicode provides the potential for multiscrut support
• Unicode is not software
  – Underlying software comes from OEMs
• Unicode is the foundation, not the whole building
  – Many parts to a system

http://www.unicode.org