Unicode™ Implementation at the Library of Congress
Cataloging Policy Position
Latest revision: March 16, 2006

The Current Situation

- The Library of Congress has added non-Latin scripts to bibliographic records using specialized cataloging systems (RLIN, OCLC) for many years.
- The non-Latin script data resides in MARC field 880 (Alternate Graphic Representation) of bibliographic records and these fields are generally provided in parallel to transliterated data in Latin script following Model A (Vernacular and Transliteration) of Appendix D (Multiscript records) of the MARC 21 Format for Bibliographic Data.
- These multiscript records are distributed via the MARC Distribution Service in the following products (total record counts as of 10/1/2005): BOOKS Arabic (49,972 records), BOOKS CJK (437,329 records), BOOKS Hebrew (51,438 records), and the MDS Serials product (over 21,000 records with non-Latin scripts). There are currently 6 subscribers for MDS-Books CJK, 2 for Books Hebrew, 1 for Books Arabic and 20 for MDS-Serials.
- LC generally provides non-Latin script data for textual monographs and integrating resources in Japanese, Chinese, Korean, Arabic, Persian, Hebrew, and Yiddish, as well as serials in Chinese, Japanese, Korean, Arabic, and Persian.
- For monographs that contain a mixture of scripts, including one non-Latin script from the JACKPHY set, DCM B5 App. 6 is used to determine whether the quantity of non-Latin script data qualifies for a multiscript record to be created in RLIN. If it is determined that a multiscript record is not warranted, the record is input in the ILS with all non-Latin script data transliterated to Latin script, whether it is a single letter, a word, several words, or perhaps an entire area.

Voyager with Unicode™ Implementation

In November 2005 the Library of Congress implemented an upgrade to its local integrated library system (Voyager) that provides for Unicode support. The primary immediate benefit of the "Voyager with Unicode" release is that users of the Library's catalog have the ability to search for and correctly display the non-Latin data previously input in 880 fields in RLIN and OCLC. The Voyager with Unicode release also provides the necessary platform for input of non-Latin data in the Voyager cataloging module. The system technically allows for the use of any Unicode character using UTF-8 character encodings; however, several protections have been built into Voyager to assure that all records are "standards compliant" within the MARC 21 environment. For example, Voyager can limit UTF-8 characters to the MARC-8 repertoire (i.e., those records that can successfully be converted from UTF-8 to MARC-8), and can automatically "decompose" pre-composed UTF-8 characters not allowed by MARC 21. The system allows for the export of records in either MARC-8 or UTF-8, although it is understood that the majority of CDS customers will require MARC-8 encodings for the near future.

Given the complexity of the changing environment, LC has already indicated that it does not intend to significantly expand the use of non-Latin data using its local system for the first year of our implementation (i.e., calendar 2006). In summary, this means that for the first year of implementation:

- Only staff that input non-Latin data prior to implementation will input non-Latin data after implementation, and only in the scripts currently used.
- LC will maintain a commitment to distribute records with MARC-8 character encodings; the only scripts that can currently be encoded in MARC-8 are those scripts within the MARC-8 repertoire of Unicode: the JACKPHY scripts, Greek, and Cyrillic.
- RLIN and OCLC will continue to be the only venues for non-Latin script data entry.
- Staff who do original and copy cataloging in the ILS using Latin scripts will continue to transliterate non-Latin data, as appropriate, even in copy cataloging records (however, this generally should not involve the removal of non-Latin script characters). For this reason, the use of IMEs will not be necessary for the input of non-Latin scripts in Voyager with Unicode during this time period.
- No non-Latin scripts will be used in MARC 21 Holdings records.
- Explorations into using non-Latin scripts for MARC 21 Authority records will be conducted during this first year.

Opportunities for the Future

The Library of Congress will spend this first year of implementation testing the system capabilities and options for expanding the inclusion of non-Latin script data in its bibliographic and authority records-- these tests should inform any decisions on changes to the current cataloging policies and practices. In addition, during 2006 the Library anticipates moving to a new operating system for client workstations, Windows XP. This is likely to bring improved tools for the Unicode environment. The Library will, of course, continue to coordinate these discussions and decisions with the NACO nodes (RLG, OCLC, British Library, NLM), its other cooperative cataloging projects, and with subscribers to the MARC Distribution Services. The Library will also be an interested participant in the expansion of the MARC repertoire of Unicode to scripts that are not part of that subset today.

A number of issues will be addressed during the first year:

- Is it possible to use non-Latin scripts in regular MARC fields of bibliographic records per “Model B” (Simple Multiscript Records) as outlined in Appendix D (Multiscript Records) of MARC 21 Format for Bibliographic Data, as opposed to the use of 880 fields in parallel to transliterated fields? This is of particular interest for transcription of multi-script identifying data elements among other issues. Preliminary decisions tend towards continuing status quo for Bibliographic records, but exploring new options for Authority records.

- Is it feasible to use the Voyager with Unicode cataloging module instead of RLIN for input of primarily non-Latin script records in the JACKPHY scripts? What practices might need to be adjusted to support indexing, searching, display of non-Latin scripts in the Voyager environment?
- Is there a need to standardize policies for the formulation of headings in non-Latin scripts in 880 fields that are currently subject to varying cataloging team practices? Movement in this direction is already taking place.
- Is it feasible to expand the use of non-Latin scripts beyond JACKPHY (i.e., Greek, Cyrillic), and if so, is it feasible to use Voyager with Unicode as the input system? LC has a preliminary prioritized list of languages/scripts to add (based on recommendations from its reading rooms) and will coordinate an agreed target list with the NACO nodes and other stakeholders.
- Is it possible to use the Voyager with Unicode cataloging module for including small quantities of non-Latin data in Latin script records (e.g., single characters, words, phrases). If pursued, determine whether Model A or Model B would be most appropriate.
- Will LC maintain a commitment to distributing records that follow MARC-8 character encodings, i.e., providing non-Latin script data only in the MARC-8 repertoire, or will it create some records for use only in systems capable of accepting UTF-8 character encodings? LC is proposing to continue distribution in MARC-8 as well as offer a Unicode UTF-8 alternative beginning sometime in 2006.
- Should LC consider eliminating the separate MDS-Books CJK, Books Arabic, and Books Hebrew services and distribute those records via MDS-Books All? This would be similar to when LC included the separate MDS-Serials CJK records into MDS-Serials a few years ago. As long as model A is used and the parallel 880s created, that data can easily be stripped out by those subscribers not wanting to utilize non-Latin scripts.
- How best to include non-Latin script data in authority records? As references (4XX) on records for transliterated headings, as separate 7XX fields, or the use of 880s? Does MARBI need to develop methods for coding subfield specific identification of the language and script used (e.g., see the Discussion Paper on Multilingual Authority Records at http://www.loc.gov/marc/marbi/2001/2001-dp05.html). LC has proposed to avoid 880s in authority records for variant forms of headings, but would like to see MARBI provide subfield coding to clarify the specific language/ script/ transliteration scheme used.
- If non-Latin script data are included in authority records, is it still necessary to supply parallel headings for access points in bibliographic records? LC would like to end this redundant practice, if possible.
- If non-Latin script data are input by LC catalogers in Voyager with Unicode, what new tools/applets/macros might need to be developed to facilitate this, e.g., tools for fool-proof generation of $6 pairing information between regular MARC fields and 880s, assistance with transliteration and/or generation of non-Latin scripts

from transliterations or vice/versa. LC has been looking for macros and other tools for "automatic transliteration" and will continue this exploration and testing over this first year. LC has already developed two tools of note: JLGReveal, a tool used to identify unambiguously the underlying characters in a text string (e.g., a field in a bibliographic record) regardless of font, and a CJK Compatibility Database to help CJK catalogers quickly and conveniently replace non-MARC characters with their MARC equivalents.

Although not technically related to Unicode per se, the Voyager with Unicode release provides the technical capability for implementing some of the characters added to the MARC-8 Latin Script set over the last several years that have not yet been implemented by LC and/or its cooperating partners. LC is working with the NACO nodes to implement these characters, and to identify retrospective file maintenance where necessary.