ALCTS Non-English Access Working Group on Romanization

Report

Dec. 15, 2009

Contents

A. Introduction

The ALCTS Non-English Access Working Group on Romanization was established by the ALCTS Non-English Access Steering Committee to implement Recommendation 10 of the report of the ALCTS Task Force on Non-English Access:

> 10. Examine the use of romanized data in bibliographic and authority records. Explore the following issues (including costs and benefits):

(1) Alternative models (Model A and Model B) for multiscript records are specified in the MARC 21 formats. The continuing use of 880 fields (that is, Model A records) has been questioned, but some libraries may need to continue to use Model A records. What issues does using both Model A and Model B cause for LC, utilities, and vendors?

(2) Requirements for access using non-Roman scripts (in general terms -- defining requirements for specific scripts falls under Recommendation 2)

(3) Requirements for access using romanization

The Steering Committee charged the Working Group as follows:

Reporting to the ALCTS Non-English Access Steering Committee, the Task Force on Romanization will examine the current use of romanized data in bibliographic and authority records, and make recommendations for best practices.

In particular, the Task Force will review Model A (*Vernacular and transliteration*) and Model B (*Simple multiscript records*) for multiscript data in MARC records (http://www.loc.gov/marc/bibliographic/ecbdmulti.html) and how these models are currently used in library systems and catalogs, including the Library of Congress catalog and OCLC WorldCat. The Task Force should consider the needs of library users for search and retrieval of items and the impact that romanized data have on searches. The recent addition of non-Roman data to authority records and how library systems are using these records should also be considered.

The impact on library staff, including acquisitions, cataloging, circulation and interlibrary loan, should also be considered, particularly in situations where staff who are not language experts may need to process materials and requests.

The task force should address the following questions:

- Is romanization still needed in bibliographic records, and if so, in which situations and/or for which access points? Should best or different levels of practices be adopted for romanization?
- Can model A & B records coexist in library systems? If so, should guidelines for usage be adopted?

Time frame: The task force should complete a report by: December 15, 2009.

The Working Group has discussed whether to recommend continuing the use of Model A indefinitely, adopting Model B now, or adopting Model B at some point in the future when certain conditions are met. Related questions include whether catalogers could stop adding romanized parallel fields for some scripts but not others, and whether some libraries could stop adding them for some or all scripts while others working in shared databases continue to do so.

The Working Group released a draft of this report on Nov. 24, 2009 and received comments from librarians at a wide variety of institutions in North America and overseas. Many of their suggestions have been incorporated into the final report.

B. Model A and Model B

Two different models for multi-script bibliographic records can be followed in MARC 21: Model A (vernacular and transliteration) and Model B (simple multiscript records). In Model A, original-script fields are paired with corresponding romanized fields. Any Roman-script text in these paired fields is repeated in both of them. The original-script fields are coded as 880 fields at the end of the bibliographic record, but in public display, and sometimes in staff display, they can be displayed next to the corresponding romanized field. (This type of display is used in the examples below.) In Model B, all transcribed text is entered only in the script in which it appears.

Model A	Model B
245 00 Татарская кухня : ǂb будни и праздники. 245 00 Tatarskai͡a͡ kukhni͡a͡ : ǂb budni i prazdniki. 260 ## Москва : ǂb Лабиринт-Пресс, ǂc 2005. 260 ## Moskva : ǂb Labirint-Press, ǂc 2005. 300 ## 511 p. ; ǂc 21 cm. 490 1# Золотая коллекция 490 1# Zolotai͡a͡ kollekt͡s͡ii͡a͡	245 00 Татарская кухня : ǂb будни и праздники. 260 ## Москва : ǂb Лабиринт-Пресс, ǂc 2005. 300 ## 511 p. ; ǂc 21 cm. 490 1# Золотая коллекция
245 00 भारतीय गौरव के बाल नाटक / ǂc संपादक, गिरिराजशरण अग्रवाल. 245 00 Bhāratīya gaurava ke bāla nāṭaka / ǂc sampādaka, Girirājaśaraṇa Agravāla. 246 1# ǂi Title on t.p. verso in roman: ǂa Bharatiya gaurav ke baal natak	245 00 भारतीय गौरव के बाल नाटक / ǂc संपादक, गिरिराजशरण अग्रवाल. 246 1# ǂi Title on t.p. verso in roman: ǂa Bharatiya gaurav ke baal natak 260 ## नई दिल्ली : ǂb डायमंड पॉकेट बुक्स, ǂc 2007. 300 ## 163 p. : ǂb ill. ; ǂc 22 cm.

260 ## नई दिल्ली : ǂb डायमंड पॉकेट बुक्स, ǂc 2007. 260 ## Naī Dillī : ǂb Ḍāyamaṇḍa Pôkeṭa Buksa, ǂc 2007. 300 ## 163 p. : ǂb ill. ; ǂc 22 cm.	
245 00 香港經濟日報 = ǂb H.K. economic times. 245 00 Xianggang jing ji ri bao = ǂb H.K. economic times. 246 31 H.K. economic times 260 ## 香港 : ǂb 港經日報有限公司, 260 ## Xianggang : ǂb Gang jing ri bao you xian gong si, 300 ## v. : ǂb ill. ; ǂc 58 cm. 500 ## Description based on: 1992年8月15日; title from caption. 500 ## Description based on: 1992 nian 8 yue 15 ri; title from caption.	245 00 香港經濟日報 = ǂb H.K. economic times. 246 31 H.K. economic times 260 ## 香港 : ǂb 港經日報有限公司, 300 ## v. : ǂb ill. ; ǂc 58 cm. 500 ## Description based on: 1992年8月15日; title from caption.

In addition to descriptive fields, headings may also appear in paired fields in Model A. In Model B, they appear only in their authorized, romanized form.

Model A	Model B
100 0# พิมพวัลค์ เสถบุตร, ǂd 1926- 100 0# Phimphawan Sētthabut, ǂd 1926-	100 0# Phimphawan Sētthabut, ǂd 1926-
610 24 인천 국제 공항 ǂx History. 610 20 Inch'ŏn Kukche Konghang ǂx History.	610 20 Inch'ŏn Kukche Konghang ǂx History.
700 1# Βενιζελος, Ελευθεριος, ǂd 1864-1936. 700 1# Venizelos, Eleutherios, ǂd 1864-1936.	700 1# Venizelos, Eleutherios, ǂd 1864-1936.
830 #0 平安文學叢刊 ; ǂv 4. 830 #0 Heian bungaku sōkan ; ǂv 4.	830 #0 Heian bungaku sōkan ; ǂv 4.

A system similar to Model B was used in North American card catalogs. Non-Roman descriptive elements were transcribed in their original script, and a "Title transliterated" (pre-AACR) or "Title romanized" (AACR) note was added at the bottom of the card, with a romanization of the title proper only.

When North American library catalogs were first automated, only Roman script could be used, so both descriptive and access fields had to be entered in romanization only. In the 1980s OCLC and RLIN began to introduce character sets for major non-Roman scripts, enabling catalogers to transcribe bibliographic data as it appears on the piece in hand. Since then libraries have cataloged material in available scripts with full romanization and varying amounts of non-Roman data in parallel fields (Model A). The amount of non-Roman data appearing in these records varies, but an attempt at standardization is now in progress, as a task force put together by the Program for Cooperative Cataloging is working on new draft PCC Guidelines for Creating Bibliographic Records in Multiple Character Sets. However, for scripts not yet implemented in OCLC, such as Tibetan or Ethiopic, romanization remains the only option. Even CJK (Chinese, Japanese and Korean script) has some characters not yet included in Unicode, so it is still not possible to transcribe original script in every case.

Model B is not yet widely used in North America, but it is used in East Asian online catalogs, i.e. no attempt is made to "transliterate" English or French text into Korean or Japanese script. However, Roman script is much more widely known and used in East Asia than CJK scripts are in North America, so the use of Model B for Roman-script publications there does not have the same implications that the use of Model B for CJK publications would have here.

C. Questioning Model A

In the days of the card catalog, North American catalogers were able to enter original script in catalog records (Model B). That option was temporarily lost after the move to online catalogs, but catalogers have now resumed entering non-Roman script in catalog records, although they do so using a different model and retaining full romanization as well (Model A). It can now be questioned why we continue to romanize purely descriptive data. The cataloging rules for many years have had a rule (AACR2 1.0E1) preferring transcription in the language and script in which they appear for certain elements. The adoption of Model B would result in simpler bibliographic records and more efficient cataloging. Romanizing takes time and can introduce errors. Romanization systems vary from country to country, and even the standard romanization systems we are supposed to use in North America can be difficult to apply consistently, unfamiliar to native speakers, and sometimes controversial (Persian, Greek, Ethiopic). Adopting Model B would not preclude the addition of, for example, variant titles in standard or nonstandard romanization to bibliographic records in 246 fields, but romanization would no longer be a routine part of the creation of every record.

C.1. Different romanization standards

Romanization is problematic when viewed from a global perspective. In North America, the ALA-LC Romanization Tables are an established standard for library cataloging, but libraries elsewhere in the world are more likely to use the various ISO romanization standards or a national standard. Often, different standards result in very different romanized strings that may, at best, look strange (and, at worst, not be recognizable) to a user accustomed to what is done in another country. They can also wreak havoc with attempts to match records. And MARC 21, unlike UNIMARC, has no way of indicating in the bibliographic record which romanization practice has been applied. (MODS, the XML schema based loosely on MARC 21, does have a type attribute for indicating both script and transliteration that can be added to any element. While a MARBI proposal to do so in MARC 21 may prove that it is too difficult to implement this in MARC 21 in ISO 2709, it may be possible in MARCXML.)

C.2. Romanizing unvocalized scripts

For some languages, even experts can differ on the correct romanization of individual words. In Hebrew and Arabic script, many vocalization marks are omitted in ordinary writing and printing, so there is a degree of uncertainty in romanizing many words. In principle, standardized romanizations are selected by consulting specified dictionaries, but even standard forms that can be easily determined this way may seem arbitrary or controversial. For the Arabic word نفـــط, the standard romanization used by the Library of Congress is *nafṭ*, but many Arabic speakers might prefer *nift*. Romanization is, in a sense, playing favorites. It values one legitimate pronunciation over other apparently equally legitimate pronunciations.

Romanization errors can easily occur if the cataloger misinterprets the romanization rules or is not deeply versed in the grammar of the language. Patrons attempting to use these romanization systems to search the catalog will face similar difficulties. Personal names and nonstandard dialect words are particularly problematic when unwritten vowels must be supplied, and it can be difficult or impossible to find an authoritative source – or any source at all – for a "correct" romanization for these. Forcing catalogers to guess in cases like these slows down the cataloging process and potentially provides incorrect information to users.

An additional complication with Hebrew and Arabic script is provided by "partially vocalized" title pages, where the publisher has provided the vocalization marks usually seen only in sacred texts or works for children. These marks are not normally included in original-script fields in cataloging records, but vowels must be included in the corresponding romanizations. The vowels provided on Hebrew materials are usually accurate, but those on Arabic materials often do not correspond to the vocalization recommended by standard sources. The Arabic word for "index," فهـرس, is

often vocalized as *fahras* on title pages, but the standard romanization is *fihris*. Current practice in Arabic cataloging is to normalize the vocalization and use the standard form rather than transliterating the vowels actually indicated on the piece.

Entire romanization systems can be problematic. The ALA-LC Persian romanization table (now under review) is frequently criticized by Persian speakers who object that no one who knows the language would ever search by current romanizations. Romanizing Persian with the same three-vowel system used for Arabic ensures that most Persian words borrowed from Arabic are romanized in the same way as they are for Arabic text, facilitating romanized searches across languages, but this vowel system does not reflect the actual pronunciation of Persian in a way acceptable to most Persian speakers.

D. Benefits of Model A

The prospect of adopting Model B raises several concerns. A number of advantages of retaining Model A and romanization have been proposed.

D.1. Users who cannot read the original script

It is often suggested that romanization can help staff and patrons who cannot read non-Roman script work with library materials in these scripts for various purposes (acquisitions, ILL requests, storage retrieval requests, assembling bibliographies). Many public libraries collect material in a wide variety of languages to serve linguistically diverse user communities, but are unable to employ specialists in all these languages. Even large research libraries are unlikely to have staff in every department who are able to interpret all the scripts used in the material they need to process.

D.1.a. Technical services

Romanization may seem to be of limited use to library staff unfamiliar with a non-Roman script. If a staff member is handling an item in non-Roman script and cannot read it, how does the romanization in the bibliographic record help the staff member match the item in hand with the record? The romanized text in the record will not appear on the piece. These staff will be more likely to look for an ISSN, ISBN or call number to match a book or serial issue with a bibliographic record, rather than trying to use tables to transliterate a non-Roman script they do not know (and even that would be impossible for unvocalized or non-alphabetic scripts).

However, the presence of romanized titles and other romanized information in bibliographic records does provide staff who cannot read the original script with text that they can easily reproduce in writing and even attempt to pronounce, for example when communicating with patrons about recalled items or fines, or in a phone conversation with a vendor. Items in non-Roman script may also arrive from the vendor with information in ALA-LC romanization attached, allowing, for example, serials check-in staff to match new serials items to the correct bibliographic record in their catalog. At some institutions the romanized title (and romanized enumeration/chronology if present) is written on the title page as part of the cataloging process, so in the case of items that are already cataloged, staff can retrieve them from the stacks and match the romanization in the bibliographic record against the form on the title page to confirm that they have the right piece.

With romanized text in bibliographic records, library staff who cannot read the original script have the opportunity to develop a limited knowledge of the language that may be useful in specific situations. If a media librarian working with video recordings has learned that the Chinese phrase *dao yan* ("director") usually appears next to the director's name, he or she will be able to identify the director of a Chinese film by looking at the statement of responsibility in the bibliographic record and finding the romanized name that appears with *dao yan*. This would be impossible if this same phrase 導演 and the director's name appeared only in Chinese script. If Yiddish text is romanized, much of it becomes comprehensible to those who know German, and while Билл Клинтон is opaque to anyone unfamiliar with the Cyrillic alphabet, its romanization *Bill Klinton* is easier to recognize.

D.1.b. Library patrons and public services

Many users search library catalogs to find material cited in bibliographies in Roman-script sources, or to find works on people mentioned in Roman-script newspapers or other publications. Most of these bibliographies and other publications do not include original script, and no newspapers do. For many languages the original-script form can be easily determined from any romanization, but for CJK this can be difficult or completely impossible, since even an expert would not know the original characters for an unfamiliar transliterated name. Having romanized data indexed in the catalog allows users to search these romanized citations and names without first having to determine the original-script form from another reference source.

Romanized records enable public service staff to help patrons with romanized citations even if they do not read the original script themselves. Of course, a public services staff member who does not know the script will just type the data in as it appears, which patrons could easily do

themselves. But being able to assist with these searches, if only in a simple way, gives librarians an opportunity to become directly involved in the patron's research that they would not have otherwise, and may allow them to go on to assist the patron in other more substantial ways.

Searches using romanized citations found in outside sources may not be equally successful for all scripts and languages. While the Chinese or Japanese romanization systems used in libraries are widely used in non-library contexts as well, for other scripts like Cyrillic or Thai there is no single broadly accepted system, and romanized data provided by a patron may not be in the system used by the library.

Even when the original-script form of a citation or name is known, standard romanization provides additional access points for those who are familiar with the system used in the catalog and might prefer to search using it. For Chinese or Japanese, some catalog users may be non-native speakers who can read the original script to a limited extent but are more comfortable with romanization. In some contexts and for some scripts a romanized search may be easier to input than an original-script one, even for users who can read the original script (see section E.).

D.2. Collocation of forms romanized the same way

Romanization provides collocation when the same word can be written in different ways in the original script. For example, *Han'guksa* ("history of Korea") can be written 韓國史 and 한국사 in Korean; *Zhongguo yi shu* ("Chinese art") can be either 中國藝術 or 中国艺术 in Chinese. Many of our systems are not yet sophisticated enough to treat these original-script forms as equivalent in their indexing (although WorldCat uses CJK mapping tables that allow traditional-character Chinese data to be retrieved when simplified characters are searched, and vice versa). And no system can automatically replace non-MARC 21 characters in users' searches with the equivalent MARC 21 forms (as given in LC's CJK Compatibility Database) that catalogers have to use to represent them in bibliographic records. A search for the romanized form retrieves all these variants.

In Hebrew, many words can be written either with extra consonantal letters to flesh out the normal lack of vowel representation (full orthography), or without them (defective orthography). Without the item in hand, a librarian or patron cannot guess how many consonantal letters to include in a non-Roman search, and if the phrase to be searched includes several words which can be written more or less fully, the number of non-Roman searches needed to cover all possibilities can be quite high. The family name romanized *Rozenberg* may appear as רוזנברג, ראזענבערג, רוזענבערג, or any of a number of other possibilities. *Yerushalayim* ("Jerusalem") may be spelled ירושלים or ירושלם, and the name *Aharon* may appear as אהרן or as אהרון. The Hebrew or Yiddish spelling of

a foreign name like “Lakewood, New Jersey” is even harder to predict. Catalogers transcribe these in original-script fields as they appear; they do not “normalize” the non-Roman spelling to one system, or enter multiple variants to account for possible spellings other than the one actually used. The presence of a romanized field which corresponds to all possible original-script orthographies provides a single “normalized” spelling so that all variants are retrieved when a romanized search is performed. (But sometimes romanization has the opposite effect; see the end of section D.4. below.)

In Arabic and Hebrew, prepositions and the definite article are prefixed to nouns. The combination is presented as a single word in non-Roman script. In ALA-LC romanization, these prefixed elements are separated from their nouns by hyphens which have no equivalent in non-Roman script. Thus a romanized search for the Arabic word *taqrīr* (“report”) will retrieve both records containing this word without an article (romanized as *taqrīr*) and records containing it with an article (romanized as *al-taqrīr*). The corresponding non-Roman forms (تقـــرير and التقـــرير) are indexed as single words and have to be searched separately. Of course, this problem is not limited to non-Roman scripts, and articles like French or Italian *l′* can cause similar problems in some systems.

D.3. Sorting

Doing a browse search for romanized text produces an alphabetical list of results in the OPAC that the user can scroll through with the expectation that specific results, if present, will be in predictable locations. Browse searches also appear to work well in most systems for the major non-Roman alphabetic scripts (Cyrillic, Greek, Arabic, Hebrew). But culturally-sensitive sorts are not yet commonly available in library catalogs for non-alphabetic languages and scripts. For CJK, sorting by code point (the current effect of a browse search in many systems) does not produce acceptable results. The sorting orders that would be meaningful to native speakers are by radical and stroke number, or alphabetical by romanization. The former is not yet supported in most systems; romanization provides the latter. However, many systems now provide the option of relevance ranking for some search results, where these sorting issues play no role.

D.4. Added value

For some languages, romanization requires the cataloger to provide information about the standard pronunciation of script forms that are pronounced differently in different contexts. For Japanese, romanization requires the cataloger to determine and indicate which of the many possible readings of a character is correct in the case being transcribed, for example whether 中 is pronounced *naka* or *chū* in a given context. (Japanese online catalogs such as NACSIS also

indicate pronunciation, although they use Japanese syllabic characters rather than romanization to do this. In the NACSIS record for the title 日本漢学文芸史研究, the title proper is followed by its pronunciation spelled out in angle brackets: <ニホン カンガク ブンゲイシ ケンキュウ>, as is the corporate name in the added entry for the issuing body: 東京教育大学文学部 <トウキョウ キョウイク ダイガク ブンガクブ>. A library adopting Model B in North America could potentially use alternative original-script data like this rather than romanization to indicate pronunciation.)

This effect of adding romanization to bibliographic records can be seen positively (providing "added value" by giving extra information about the readings of original characters) or negatively (sometimes Japanese catalogers need to spend a considerable amount of time researching the correct readings before they can enter them). For CJK, providing pronunciation-based access points can be useful for users who know the basics of a language but are not fully proficient in the original script. From a public services perspective, it could be a disservice to users to stop providing romanization, especially for undergraduates, beginners, or researchers who are not experts in these languages but need to work with materials written in them and have some ability to do so.

Although this sort of information is helpful to some users, it is not clear whether providing it should be seen as an essential function for a cataloger. It would certainly be much simpler just to transcribe the original script as it appears on the resource. And users who search using romanization may sometimes have to do separate searches to account for differences in pronunciation in text strings that would be retrieved by a single search done in the original script (for example the Arabic word لغـة, which can be romanized as either *lughah* or *lughat* depending on the grammatical context).

E. Systems that do not support non-Roman script

Records with non-Roman script only are useless in systems that cannot handle non-Roman script at all, and while these are now probably rare in North American academic and research libraries, they are still more common in other types of libraries (public libraries, school libraries) and in other countries. Many systems can handle some scripts but not others. A 2007 Cataloging Distribution Service survey related to character sets in MARC records found that a significant portion of their subscriber base was not yet able to handle UTF-8 records, i.e., they were generally limited to non-Roman scripts that are part of the MARC-8 repertoire (Chinese, Japanese, Korean, Arabic, Persian, Hebrew, Yiddish, Cyrillic, Greek). The limitations vary system by system and may be related only to certain facets of system functionality (e.g., import, export, input, display, indexing).

A separate question is whether Unicode characters are fully supported by other software packages that libraries use or provide: labeling software, openURL resolvers, electronic resource management systems, tools for creating bibliographies or e-mailing bibliographic records, etc.

Many languages and scripts outside of the MARC-8 repertoire of UTF-8 are currently impossible to input into LC's local system due to a bug that renders their Microsoft IMEs unusable. Even if other libraries become UTF-8 compliant, romanization will be the only way for LC to enter and distribute those records for some time to come.

Public library terminals (or users' personal computers) may not always allow non-Roman input. Even if input is supported, different groups of users may need a variety of keyboards, depending on which input method they normally use. For simplified-character Chinese, there are four different kinds of keyboard available in Windows. For traditional-character Chinese, there are eight. Not all of these may be installed on individual machines, even on the library's own terminals.

In some cases, searching by script alone is completely unsatisfactory because of flaws in the Microsoft IMEs used by staff to input records or by users to enter catalog searches. The Microsoft Farsi (Persian) IME lacks some common and necessary characters which must be created by workarounds in cataloging and cannot be input at all by searchers, so non-Roman searches for strings containing these characters will always be largely unsuccessful.

F. Headings and authority records

Since headings are established in romanization in the LC/NACO Authority File, they need to be entered in bibliographic records in the same romanized form. Name headings may now have original-script references in the authority record, but subject headings do not. Current practice allows and sometimes requires the addition of parallel heading fields in bibliographic records that contain original script. For example, in current PCC documentation (now under review), parallel original-script heading fields are required for Arabic CONSER records with original-script descriptive fields, but for CJK CONSER records with parallel descriptive fields, parallel heading fields are optional.

Parallel fields for headings in bibliographic records are still necessary for keyword searching on script names in most current systems, which do not provide keyword searching of references in linked authority records. They are also necessary for a complete display in script of the basic bibliographic description for users who are unfamiliar with the romanization used (for example Cantonese speakers looking at Chinese records, where romanization is based on Mandarin

pronunciation). They are essential when the romanized form is ambiguous, as it is for Chinese names where any romanized form could correspond to multiple names written with entirely different characters.

For some language/script cataloging communities, current guidelines attempt to ensure that headings are entered in a form that "corresponds" to the authorized romanized form, but there are still problems that prevent complete standardization. The same authorized romanized form may correspond to more than one original-script spelling (Ивановъ or Иванов for *Ivanov*; 中國 or 中国 for *Zhongguo*), and different practices exist for cataloger-supplied qualifiers (entered in the authorized romanized form, or in a "corresponding" original-script form, or omitted; this is a particularly difficult problem for right-to-left languages). So original-script headings, unlike romanized ones, are never completely consistent, and result in split indexes in the catalog.

One of the perceived advantages of adding non-Roman references to authority records was the hope that, if they were added, it would no longer be necessary to provide non-Roman parallel access points in bibliographic records. However, when the project was undertaken to prepopulate the LC/NACO Authority File with non-Roman headings from OCLC, it became clear that providing users with full access to records without parallel heading fields is possible only if authority data is fully integrated into the searching process. Ideally, most heading variations should be handled by authority records; if successfully implemented, this approach would resolve many of the difficulties described above and allow redundant use of original-script data in parallel bibliographic headings to be limited to special cases where it might still be necessary. But when the system in use does not fully integrate authority data (and most current systems do not), access is lost if parallel fields are not maintained in all records.

In the current cataloging environment, romanization of author and title is also necessary to allow Roman-script catalogers to catalog translations and history or criticism of non-Roman works. It would be difficult or impossible for them to enter uniform titles and subject headings if the work cataloged is in Roman script and the record for the original is in non-Roman script only. Romanization is also needed to determine the cutter number for a newly cataloged non-Roman work, and subsequently to allow other catalogers to determine shelf arrangement when cuttering other works around that non-Roman title. Works in non-Roman script cannot be viewed in isolation from the other bibliographic records in the catalog, and romanization provides a linguistic crosswalk to make them transparent to staff and users with different language backgrounds.

G. Automation of romanization

The effort required to provide romanization in bibliographic records can be reduced by automation. Conversion from original script to current romanization schemes and back can be automated fairly easily for scripts like Cyrillic and, with the exception of the rough breathing, Greek. For scripts where not all vowels are indicated in normal orthography, text cannot be automatically romanized from the original script. However, Arabic and Persian can have original script automatically reconstructed if the romanization is entered first. Conversion from romanization to original script is also more easily implemented for South Asian scripts than the reverse. Since these romanization schemes match their scripts nearly character by character, automatic tools can be designed which make few errors. Hebrew script presents more problems, since the romanization system contains many ambiguous signs and Hebrew orthography is not fixed. An automated process will not be able to tell from a string of text romanized according to the current ALA-LC Hebrew tables whether the publisher chose a “full” or a “defective” orthography for the original script.

The Library of Congress is using a tool (Transliterator) developed by staff at LC and Northwestern University that currently provides automatic transliteration for several languages and scripts (Chinese, Korean, Arabic, Persian, Urdu, Russian and some other Slavic languages), and is testing a module for Hebrew and Yiddish. While this tool works better for some languages and scripts than others, thorough review of the results by the cataloger is required for all. The Korean module represents a cooperative effort to develop the transliteration dictionary, and LC is interested in pursuing similar cooperation for a Japanese module. Users of OCLC’s Connexion client software have access to a tool that transliterates Latin script to Arabic script for Arabic and Persian languages, and OCLC member libraries have contributed macros to transliterate to and from Cyrillic, Greek and Hebrew. In India, DK Agencies has also begun to develop software to automate romanization of the major South Asian scripts.

Even with some of the shortcomings mentioned above, automated transliteration techniques usually provide a significant benefit to the cataloger, and for many scripts they can greatly increase cataloger productivity.

H. Models A & B in one catalog

Models A and B can coexist in one catalog, and already do in OCLC WorldCat where many records in non-Roman script only have been loaded by vendors and from the National Library of Israel. Non-Roman searches will retrieve records created under both models. But if libraries adopt Model B for future cataloging, their catalogs will still have (in addition to the existing Model A records) a large number of older bibliographic records cataloged using romanization only. In

addition, there are countless records for Roman-script works containing romanized non-Roman words in their descriptive fields and headings. It would be very difficult to add non-Roman script to those records. To convert them manually would require extensive resources, and for many scripts automated conversion without review would not provide even approximately correct results. If the records are left unconverted, original-script searches would not retrieve pre-Model A records, and romanized searches would not retrieve post-Model A records. Libraries would have permanently split catalogs for their non-Roman materials, although here again more sophisticated use of authority data or improvements in automated script conversion in future systems could mitigate some aspects of the problem.

I. Recommendations

1. A majority of the Working Group believes that the factors discussed in this report are significant enough to make a general shift to Model B in bibliographic records premature at this point. Some members of the Working Group feel that having romanized access points in records provides enough added value that their use should be continued indefinitely. Others believe that in an environment of shrinking staffs and production pressures we should anticipate future developments in making our decision and recommend a move to Model B sooner rather than later. However, most believe that although a gradual move towards the use of Model B for current cataloging is probable, we should continue use of Model A for now as we prepare for a potential transition.

2. Further research is needed into the remaining obstacles so that we can identify decision points that will allow us to move beyond the status quo. We recommend that ALCTS sponsor a survey of libraries and library systems, and consult with library system vendors and developers of other tools used in libraries. This will provide us with a better understanding of the current situation and possible future directions from a technical perspective.

3. Automatic transliteration software should be utilized to reduce time needed to create romanization from original script (or original script from romanization), when possible. This will be an option primarily for larger libraries that have significant non-Roman collections and knowledgeable staff able to proofread and correct the results of automated processing.

4. Since different languages and scripts raise very different issues, some language/script cataloging communities may decide to move to Model B sooner than others. A coordinated decision to change practice within each community would be preferable to individual decisions to implement Model B across all scripts in different libraries at different times. Further research into

the needs of specific user communities in different types of libraries should be conducted to guide these decisions.

5. Non-Roman text transcribed in descriptive fields should be entered in its original script whenever possible. Language/script cataloging communities should consider whether the amount of romanization in Model A records could be reduced by limiting it to those fields containing key data for access (such as titles and headings) which provide the benefits of romanization described in this report. Non-Roman text in other fields would then be entered in original script only, not in romanization only.

Appendix: Members of the Working Group

Robert Rendall (chair)
Principal Serials Cataloger
Columbia University Libraries

Joyce Bell
Cataloging and Metadata Services Director
Princeton University

Joseph (Yossi) Galron-Goldschlaeger
Head, Hebraica & Jewish Studies Library
The Ohio State University Libraries

Douglas Hasty
Head, Access Services Department
University Libraries, Florida International University

Heidi G. Lerner
Hebraica/Judaica Cataloger
Stanford University Libraries

Sandra Nugraha
LC Overseas Offices Representative
Library of Congress Jakarta Office

Glenn E. Patton
Director, WorldCat Quality Management
OCLC

Dave Reser
Policy & Standards Division, Library of Congress

Keiko Suzuki
Japanese Catalog Librarian, East Asian Library
Sterling Memorial Library, Yale University

Jian Wang
Coordinator, Multilingual Library Services
Saskatchewan Provincial Library, Ministry of Education

Magda El-Sherbini (liaison to the Steering Committee)
Associate Professor and Head, Cataloging Department
The Ohio State University Libraries

The Working Group is grateful to the following members of its Open Discussion Forum who contributed to the development of the first draft of this report:

Joan Biella, Library of Congress

Beth Camden, University of Pennsylvania

Sarah Elman, Columbia University

Faye Leibowitz, University of Pittsburgh

Lucas Mak, Michigan State University