WHITE PAPER : ISSUES RELATED TO NON-LATIN CHARACTERS IN NAME AUTHORITY RECORDS

Non-Latin Characters in Name Authority Records

The major authority record exchange partners (British Library, Library and Archives Canada, Library of Congress, National Library of Medicine, OCLC, and RLG Programs (and predecessors)) have agreed to a basic outline that will allow for the addition of non-Latin characters in references on name authority records distributed as part of the NACO program—no earlier than April 2008. Rather than using 880 fields that parallel 'regular' MARC fields as in bibliographic records, non-Latin script references in authorities will be added following MARC 21's "Model B" for multi-script records. Model B provides for unlinked non-Latin script fields with the same MARC tags used for romanized data, such as authority record 4XX fields.

Although there was initial discussion about choosing a 'preferred' non-Latin script variant form and flagging it as such on the authority record, it was recognized that forging agreements about what constituted the 'preferred' form could seriously delay the addition of non-Latin references. The group preferred a plan with a short-term goal to allow non-Latin references without declaring whether that reference was the preferred form for any particular language or script. As references are allowed to conflict, no attempt is needed to determine if non-Latin references on one record normalize to the same form as other non-Latin references on different records.

Pre-Population of the LC/NACO Authority File

While the record-by-record addition of non-Latin references to new or existing authority records will be optional for NACO participants, OCLC has proposed to pre-populate the LC/NACO Authority File with non-Latin references derived from non-Latin bibliographic heading fields from WorldCat, making use of data-mining techniques developed for the WorldCat Identities product. This approach of harvesting non-Latin heading forms that correspond to entities in the authority file will provide an immediate value for the authority file, based on the significant intellectual work of the many libraries that have provided non-Latin headings on bibliographic records for many years. This project could see the addition of as many as 500,000 non-Latin references to authority records, a significant 're-use' of existing metadata in new contexts.

Because there have not been uniform practices in the use and form of non-Latin headings in bibliographic records, this lack of uniformity will be reflected in the pre-population of the authority records. Once the references have been added to the authority records, catalogers will be better able to observe the past practices related to non-Latin headings, and should be in a better position to recommend future 'best' practices for the LC/NACO Authority file. This review, and the development of recommendations of best

practices, will ideally be addressed as a community in the first six months following the pre-population (April-October 2008). The rest of this paper provides background information on the varieties of existing practices, and raises many of the issues that will need to be considered by the community in the development of best practices.

Background on Practices for Non-Latin Headings on Bibliographic Records : the LC Experience

Policies for allowing non-Latin references in authority records need to be established. As a starting point, the existing practices for supplying parallel 880 fields for headings on bibliographic records are discussed here to identify the relevant issues and to determine the degree to which existing bibliographic 880 practices might inform policies for non-Latin references in authority records. This discussion will also bring to light the mix of practices that will be observed with the pre-population of the authority record references from bibliographic records headings.

While it would be difficult to succinctly characterize practices for non-Latin access points in bibliographic databases, the experiences of the Library of Congress in supplying such access points may be typical of other libraries, thus the LC experience is discussed here as background.

Library of Congress practices for supplying non-Latin access points in bibliographic 880 fields vary considerably from language group to language group. These variations were developed and nurtured for two decades; the lack of conformity is based in part on a consensus view at LC, developed in 1987, that non-Latin access points were not intended to perform a 'controlling' or 'collocating' function, but rather an 'identification' function. The original LC Nonroman Cataloging Committee concluded that LC did not have the resources to support two simultaneous accessing systems following the traditional conventions of authority systems. Instead, a controlled system using the *romanized* forms of names was adopted: the romanized heading is represented by an authority record 1XX form, and is the 'official' heading for LC's English-language based cataloging data, subject to full authority control. The Latin script or romanized heading is established per AACR2, with the inclusion of appropriate cataloger-added additions as specified in AACR2 and LCRIs. All appropriate references specified in AACR2 or LCRIs are also required to be in romanized form.

Since the Latin script or romanized forms were performing the 'controlling' function, the function of non-Latin strings in bibliographic 880 fields was seen primarily as a complementary one—providing uncontrolled access to the catalog by presenting a non-Latin string in addition to the Latin script or romanized heading. To accomplish this, no research need be performed to establish the 'authorized' non-Latin form or to break conflicts among different entities using the same name. Catalogers' additions were generally added or not added depending on team practices, even when found

in the authorized romanized forms in the work being cataloged (this latter point has led to many variations between language groups).

Two additional factors influenced either the original practice, and/or the evolution of language-specific variations over time: 1) non-Latin data was neither searchable nor viewable in the LC online catalog until fairly recently, thus any discrepancies between language groups or between Latin script or romanized and non-Latin headings were not easily perceived; and 2) the difficulties of keying multidirectional data necessitated that the number of fields that used both Latin and non-Latin scripts should be kept to an absolute minimum. Needless to say, this latter point was of critical import primarily for catalogers of right-to-left script languages (Hebrew, Arabic, Persian, Yiddish (HAPY)). Thus, the HAPY catalogers follow a practice of formulating non-Latin headings that reflect forms found on the publication, without dates or other cataloger-added additions specified by the rules or LC Rule Interpretations[1]. Because of this, the strings used in non-Latin headings for a particular entity may fluctuate considerably from bibliographic record to bibliographic record.

Since there were fewer difficulties for mixing roman script data and Chinese, Japanese, or Korean (CJK) data in the same field, the CJK catalogers generally follow a practice of creating non-Latin headings in bibliographic 880 fields that more closely parallel the Latin script/authorized form of the heading rather than simply replicating the form found on the item being cataloged. See Appendix 1 for an explanation of what we believe to be current LC practice.

Note that LC does not currently provide non-Latin script cataloging for Cyrillic and Greek (outside of a few serial bibliographic records). Unlike the JACKPHY languages where longstanding practices have developed, there does not appear to be a consensus practice in bibliographic 880 fields for headings in Cyrillic or Greek script at this time, although as left-to-right scripts, these languages are unlikely to encounter the same difficulties as the HAPY languages.

While these issues have formed LC's experience with non-Latin characters in bibliographic records, it is clear from records created by other libraries found in various databases that there are even more divergent practices in play. Different practices, particularly for fields with mixed roman and right-to-left script, are commonplace.

[1] Note that even if roman script additions (e.g., dates, qualifiers) were made to right-to-left script headings, there is still disagreement within the library community as to how the roman script data should display (e.g., date spans in logical or visual order).

Possible Guideline Approaches

Even with the less-stringent short-term goal of providing non-Latin script references without specifying a 'preferred' form, some basic guidelines on adding non-Latin references are seen as desirable, and will be an issue that the community must address in the first six months after pre-population of the authority file (April-October 2008). This desire is particularly important for script/language groups that have yet to develop a consensus approach to headings on bibliographic records (e.g., Cyrillic, Greek). The question at hand is which approach the guidelines should endorse:

Approach 1. Add 4XX fields to authorities that reflect the form found on publications without catalogers' additions specified by the rules (i.e., similar to the LC HAPY approach for bibliographic 880 fields).

Approach 2. Add 4XX fields to authorities that closely parallel the authorized roman form, including catalogers' additions (i.e., similar to the LC CJK approach for bibliographic 880 fields).

Approach 3. A mix of approaches, as currently found in bibliographic 880 fields; this could range from a 'cataloger judgement' policy, to suggested 'best practices' based on languages or scripts (e.g., possibly following Approach 1 for right-to-left scripts, and Approach 2 for left-to-right scripts).

Pros and cons of the three approaches follow.

Pros: Approach 1	Cons: Approach 1
• Easy to transcribe the form of name to 4XX field, even in right-to-left scripts (generally without switching to left-to-right script keyboards) • Not necessary to develop consensus as to how mixed roman and right-to-left non-Latin scripts should be displayed • Variant forms of names found on publications will be searchable as references in authority systems	• Different forms found in different publications may lead to many different 4XX forms on authority records that may under current guidelines be seen as 'variants of variants' • LCRI instructions to formulate references the same as headings vis a vis most cataloger-added additions would need to be changed to allow variation for non-Latin script references • Browsing references without cataloger additions such as dates may make it harder to identify/select entities

Pros: Approach 2	Cons: Approach 2
• References would be constructed as headings vis a vis catalogers' additions, generally following rules/LCRIs	• Additions requiring combinations of right-to-left and left-to-right scripts may be more difficult and time

• Browsing references with cataloger additions such as dates may make it easier to identify/select entities	consuming to input in some systems, and will require agreements on the use of Unicode Formatting Characters by NACO nodes • Disagreement in the cataloging community as to whether roman additions in right-to-left script headings should display in logical or visual order (start-end, or end-start) would need to be resolved, as would issues related to use of dates from non-Gregorian calendars and using non-Western characters for numbers • Roman additions to non-Latin strings may look "odd" to some users; note, however, that these records are intended for use in English language catalogs

Pros: Approach 3	Cons: Approach 3
• See above • Systematic pre-population of authority record reference fields from bibliographic 880 fields will already represent this mix of practices	• See above • Mix of practices on records with different scripts may prove difficult for catalogers who work in multiple non-Latin scripts

Cataloger Additions to Non-Latin References: Script Choices

If approaches 2 or 3 above are accepted and non-Latin references need to "match" the roman or romanized heading in certain aspects, there likely will need to be agreement on the form of cataloger additions to headings. AACR2 instructs (or implies) that some additions to personal names should be "in the vernacular,"[2] or in the form found with the name on items being cataloged or in reference sources. Examples of allowing

[2] In certain cataloging communities, 'vernacular' has become equivalent to 'non-Latin' in common parlance; while others see it as a pejorative or offensive term. AACR2, however, implies a much broader definition: the standard native language of a country or locality (which, could even be English or French). This paper will use 'native language/script' as a friendly, but possibly imperfect, substitute for vernacular.

additions in the native language/script include 22.5F, 22.6, 22.12, 22.16D, etc.

Other instructions require (or LC has interpreted the rule to imply) that the addition should be supplied in English, such as 22.11A (addition to name, such as "(Writer)"), 22.13 (Saint), 22.14 (Spirit), 22.16B (Pope), etc., although some instructions prefer such an English addition only if there is an adequate English equivalent, such as 22.16A (Royalty). There are indeed cases where a heading might require both an addition in the native language/script and one in English.

Geographic names may also have additions that use English or native language/script, depending on the outcome of the application of AACR2 23.2A-B (Geographic names). Additions for corporate bodies are also mixed—English additions must be used in some instances (24.4B1 (additions to convey corporate-ness), 24.4C7 (other designations), 24.10 (Churches), 24.11A (Radio and television stations), etc.), while place names added as qualifiers could be in English or native language/script per 23.2 (Geographic names), and institution names used as qualifiers would be in the form and language used for that institution as a heading. Conferences, congresses, meetings, etc., also get qualifiers for number (in ordinal English form), date (English implied), local place (English or native language/script per 23.2), or other location (form found on the item). Needless to say, additions to uniform titles may also be required to be in English, e.g. "Laws, etc.", or may be in the native language/script. It should be noted that some cataloging communities already supply additions in the native language/script, even where the rules call for English language forms.

Several basic questions must be answered in determining the form of qualifiers:

Question 1. In those instances where AACR2 specifies that an English form of a qualifier must be used in the authorized heading, should the same English qualifier be used as part of an addition to a non-Latin reference (where applicable[3]) given that the mix of scripts may be less usable in the future for non-English displays?

Question 2. Should English language collective uniform titles specified in AACR2 (e.g., Works, Selections, Laws, etc.) used in authorized headings also be used in the non-Latin reference in this English form?

Question 3. Should abbreviations specified by the rules like "ca.," "b." and "d.," and "fl." be supplied using non-English and/or non-Latin forms?

Question 4. In those instances where native language/script forms of additions are specified by the rules, should the additions be in romanized or non-Latin script forms?

Question 5. For name/title authority records, should non-Latin references reflect non-Latin script forms for both the name and title portion

[3] Although references are generally constructed in the same form in which they would be constructed if they were headings, there are some instructions in the LCRIs that allow some additions to differ from the heading, or to omit some additions from references.

of the reference, or should the name portion follow the authorized heading? Likewise, when establishing or adding a non-Latin reference for a sub-body entered indirectly, should the form of the higher body also be in non-Latin script, or follow the authorized form?

While the focus of this paper has been primarily on authority record references using non-Latin characters, it should also be recognized that 670 (Source citation) and other notes fields found in authority records will also have to be included in authority records. This will present the same bi-directional challenges for fields that combine right-to-left and left-to-right scripts in a single citation.

Appendix 1: Current LC Practice for Personal Name Headings in Non-Latin script 880 Fields

Caveat: LC practices have evolved over time; older records may not necessarily reflect the 'current' practices noted here. In addition, LC catalogers using copy from other institutions may or may not adjust otherwise acceptable copy to conform to LC's practices in all cases.

Basic name portion of heading

Hebrew, Yiddish: transcribe the form found on the item being cataloged, without attempting to 'match' the established heading.
Arabic, Persian (?): construct the heading to 'match' some additions to the established heading, even when not found on the item being cataloged (at least for personal names)
Chinese, Japanese, Korean: match the established romanized heading

Additions: titles and other words associated with a name (X00, $c)

Hebrew, Yiddish: Provide $c, in the non-Latin form, when the term associated with a name consisting only of forenames or in accordance with rule 22.12A when the term appears in the item. Do not supply $c in romanized or English forms due to bi-directionality issues.
Arabic, Persian: Give $c only rarely, only when the established heading includes $c. Do not supply $c additions to non-Latin headings in either romanized or English forms due to bi-directionality issues.
Chinese, Japanese, Korean: Give $c to match the established romanized heading. Korean catalogers provide it in the non-Latin form, Chinese repeat the form used in the romanized heading, even if that form is in English.

Additions: dates added to personal names (X00 $d)

Hebrew, Arabic, Persian, Yiddish: Provide $d only if present in copy cataloging record. [in what script, Hindi numerals or western-style?]
Chinese, Japanese, Korean: Match the romanized form.