ALCTS Non-English Access Working Group on Romanization: Open Discussion Forum

Posts and comments

June 24-Dec. 1, 2009

Model A and Model B

by Robert Rendall on Wed, 06/24/2009 - 2:43pm

The report of the Task Force on Non-English Access says:

"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?"

This was incorporated into the Working Group's charge as:

"Can model A & B records coexist in library systems? If so, should guidelines for usage be adopted?"

Before we start considering the ramifications of having Model A and Model B in the same environment, I'd like to ask:

Why and by whom has the use of Model A been questioned?

and

Where is Model B currently in use?

Please share any thoughts or information you have.

Examples:

Model A

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,
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.

Model B

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,
500 ## Description based on: 1992年8月15日; title from caption.

For more information on Model A and Model B see the this page on the MARC 21 site.

Model A and Model B comments

Joyce Bell
Mon, 06/29/2009 - 9:28am

I, for one, am questioning the use of model A. In the card catalog days we had progressed to the point of the card version of model B. We had to retreat when we moved to online catalogs. The recent advent of Unicode allowed us to regain that lost ground, although we have done so by using a different model. I question why we continue to romanize purely descriptive data. The cataloging rules for many years have had a rule (1.0E1) preferring transcription in the language and script in which they appear for certain elements. Romanizing takes time, and romanization introduces errors.

I have heard similar talk from others, although I am unsure if there is an organized movement. --Joyce

More Model A and Model B comments

Keiko Suzuki
Tue, 06/30/2009 - 2:57pm

(I don't know what's the best way to continue: "reply" to Joyce's comments or "Add new comment"??? I'm "replying" for now, but Robert, please let us know if there's a preference.)

So my comment: I assume, for this post, we are mainly discussing about bibliographic records, not authority records because I would like to think separately, at least the beginning, for bib records and auth records on this topic. Then think about it, since the recent addition of non-Latin script references, we do use Model B for our authority records (no 880s but another 4xxs or 5xxs) although I don't think of any other databases use Model B for bib records.

Yes, I've heard the talks about questioning Model A, and I myself have mixed feeling about it. Model B is simple, much less time-consuming so that suitable for efficient cataloging. Yet, I have to point out that, for Japanese language materials, transcribing with romanization/transliteration is added value to include *reading* for library users. Because we have so many varieations of reading each Chinese character, it's impossible to make automatic transliteration program for Japanese. I would guess there are other languages/scripts like it, especially those without sufficient auto-transliteration software. Also I have to say CJK has some scripts not covered by even Unicode, it's not yet perfectly transcribe of original scripts.

On the other hand, as Joyce mentioned, libraries had Model B card catalogs before. Thus, end-users could survive without it? I mean "adding a redaing" is value (it really is for certain occations since some readings are extremely trickey for even a native speaker), but is it really a libraraian/cataloger's job? I don't know.

Anyway, I would like to hear other language people who have simmilar issue, and/or eventually would like to hear non-catalogers (public services including ILL) and end-users as well.

There are other things, but I stop here now.

Background information

Beth Camden
Thu, 07/02/2009 - 9:24am

There is a document on the ALCTS website which includes some background on past discussions of the Model A & B question and romanization in general. See: ALCTS Task Force on Non-English Access: Comments Received and Their Disposition http://www.ala.org/ala/mgrps/divs/alcts/ianda/nonenglish/07rptcomments.pdf

In particular note pp. 8-9 and 27-30

I feel that the romanization question is not as simple as I once thought. Certainly for users, the original scripts are what they need.

But library systems are used by many others who may not be able to read languages in non-roman scripts. How do staff handle ILL or storage retrieval requests or acquistion functions? Are the software packages that libraries provide for creating bibliographies able to deal with Unicode characters?

If some of our bibl. records have parallel script fields and some have just romanization and some have just script, what will this do to searching?

Sarah S. Elman, Head of

Sarah Elman
Sun, 07/19/2009 - 9:55pm

Sarah S. Elman, Head of Technical Services, C.V. Starr East Asian Library, Columbia University (212) 854-2579, sse2109@columbia.edu

To answer Robert's question about where Model B is currently in used: Although it is not called the same thing, but the equivalent of Model B is used at least in East Asian countries. I suspect other countries where non-Roman scripts are used might be the same.

Romanization is certainly not needed by those who know the language and can read the script. However, in a global environment, Romanization adds value to information retrieval because of the following factors (some of them were mentioned in earlier messages):

1. It provides additional access points for those who prefer to search by it, and it helps library staff who does not read the original scripts handle the materials more easily.
2. Before bibliographic utilities and local systems could handle non-Roman scripts in early 1980's, a large number of bibliographic records were already cataloged using Romanization only. To convert those records will require extensive resources.
3. There are countless Western language works containing Romanized non-Latin words in their titles. No non-Latin scripts will be added to those records.
4. LC Subject Headings are only in Romanized or English form so far.
5. Many languages use different scripts to write, for example, traditional vs. simplified Chinese, Japanese kana vs. Kanji and Korean Hangul vs. Hanja, etc. Unless mapping tables are utilized, Romanization is the best way to bring them together.

As for the choice of Model A or B, assuming that all systems are capable of handling both models simultaneously in the future (all systems have to be Unicode-based and the indexing of non-Latin scripts has to be modified and improved), I think Model B is probably the better choice since it's structurally

simpler than Model A. However, I do not recommend that we give up romanization entirely for the reasons mentioned above. Instead, I think the following scenario might work:

1. The description portion is transcribed in the original language/script of the work while additional access points for key data in Romanized form can be added. For example, a Romanized Chinese title can be entered in 246 as a variant title instead of 880. Name headings can be collocated by authority records.
2. Automatic transliteration software is utilized to reduce time needed to create the romanization from scratch. Of course, catalogers will still have to proof-read it.
3. Mapping tables are devised to link different Romanization systems as well as scripts within the same language.
4. Institutions can choose to suppress the romanization from display to suit local needs if necessary.

Thanks.

Sarah

Sarah S. Elman
Head of Technical Services
C.V. Starr East Asian Library
Columbia University
(212) 854-2579
sse2109@columbia.edu

Some comments on the comments from Joyce, Keiko, Beth and Sarah

Glenn Patton
Tue, 07/21/2009 - 8:13am

One of the pieces of "conventional wisdom" about romanization is a thread in these comment. We've heard for years that having the romanized text of the description is of help to some users (particularly, other library staff) who may not read the script. I'm wondering if that's really true. If I'm a staff member who checks in serial issues and I don't read [fill in the name of a script], does the romanization in the bib record for that serial help me match the issue that's in my hand with the bib record and associated item records in the library's system? Obviously, the romanized text doesn't appear on the issue. Siminarly, does the romanization really help me if I'm working with an ILL request? Could we find a way to solicit "real life" experiences from our colleagues.

When I look at romanization from a global perspective, the usefulness seems similarly unclear. In the US, we're accustomed to using the ALA-LC Romanization Tables but libraries elsewhere in the world are more likely to use the various ISO romanization standards or a national standard. Often, the results of applying different standards results in very different romanized strings that may, at best, look "wierd" (and, at worst, not be recognizable) to a user that is accustomed to what's done in another country. In my own experience, it also can wreak havoc with attempts to match records. And, MARC21 (unlike UNIMARC) has no way of indicating in the bib record which romanization practice has been applied.

--Glenn

Some thoughts.

Jan Wang (non-member)
Tue, 07/21/2009 - 11:15am

Model B works for library users who:

1) have access to a computer which offers entry of original scripts and a library catalogue which supports searching and display in original scripts; or

2) know how to translate a title in original scripts into English.

Most of my library experience is in public library world, where multiscript entry and searching may not be facilitated by library catalogue and computers, and some library users not English proficient. In this case, Romanization or Model A, as is mentioned, offers an additional access point. I guess Model B is preferred more by academic libraries, where catalogue and computers offer more support to searching in original scripts and library users have better English language skills.

"Conventional wisdom" about romanization

Faye Leibowitz
Mon, 07/27/2009 - 12:01pm

Glenn--

I agree that romanization is of limited use to library staff unfamiliar with a non-roman script. My experience has been that these staff will look for an ISSN, ISBN, etc. to match up a book or serial issue with a bib record rather than trying to transliterate a non-roman script.

Most of the materials that I catalog are in Western European languages, but since I have some knowledge of the Hebrew and Arabic scripts, I'm the person responsible for cataloging anything in these scripts as well. I am by no means an "expert" in these languages. I subscribe to the HebNACO and Mideastcat discussion lists to learn about new developments and trends relating to cataloging these types of materials. My impression from monitoring these lists is that even "experts" in these languages differ relating to the romanization of many words. This is particularly true because Hebrew and Arabic are generally printed without vowels in the vernacular, so there is a certain degree of uncertainty in romanizing many words, even after consulting the standard dictionaries.

That's why I'm a big fan of model B. But model A should be sustained until all libraries are technologically equipped for a totally unicode environment-- maybe model A and model B must coexist for a long time to come. But the fact that some libraries can't support model B should not deter libraries that are able to use it from doing so.

Are there lots of libraries out there that have large non-roman script collections that are incapable of adopting model B? I would think that libraries with large non-roman script collections would require unicode-compliant systems as a priority. I'm not trying to minimize the importance of having non-roman script materials in ALL libraries to promote diversity-- I'm just saying that this question might have more impact on libraries with large non-roman script collections.

Perhaps a decision might need to be based primarily on the requirements of the libraries that are most affected by the problem.

Best wishes,

Faye Leibowitz

University of Pittsburgh

frleibo@pitt.edu

Continuing this facinating discussion ...

Keiko Suzuki
Mon, 07/27/2009 - 1:52pm

Transliteration is, in general, not the original information on items, and it could be slight variations from system to system. Thus, it sometime gives additional complexity in some positive way (as I said for the case of Japanese, we would have extra information of reading the original characters) and negative way (on the other hand, sometime Japanese catalogers need to spend much time to find the correct "readings") or the case of Arabic and Hebrew to add vowels might be both ways? At the same time, some pointed out that we could give extra access points for some users who are not proficient to the original scripts. I had a chance to talk to our public services librarian and she thinks it could do disservice to the users to stop transliterating, especially undergraduate or beginners, or researchers whose major are not exactly in those languages/areas but to do need to do some research. It might be the case that people learn some languages not through its original scripts but transliteration, thus, easier with it? So I would like to hear more from reference librarians, too.

Glenn, actually I haven't checked myself, how OCLC index all those scripts records in culturally appropriate manner? I assume some scripts are no problem with unicode point order because their alphabets are orderly in the unicode. However, for the CJK, and assume some other scripts, the order of unicode point is not exactly meaningful. In one way to deal with it is the order of radical strokes, and the other is Latin-alphabetical order of transliteration. However, if we stop transliterating ... then that's not possible.

- Keiko Suzuki

Continuing the discussion ...

David Reser
Mon, 07/27/2009 - 3:12pm

I'm not sure I have anything unique to add to the discussion-- I often hear folks here describe romanization as a necessary evil. Glenn has suggested that some of the oft-heard reasons may be debunked if we look at them very closely (I know one reason used here in the past has been related to our deck attendants retrieving from the stacks, and yes, we transcribe a romanized form of a title and enum/chron on the t.p. so they can match against a record (at least we used to)).

As others have suggested, we aren't all in a perfect system world, technologically speaking, and that abandonment of Model A may pose problems. It was rather shocking to us when our Cataloging Distribution Service ran a survey in 2007 related to character sets in MARC records-- we had hoped we

were to the point that we didn't have to distribute MARC-8 records anymore, but a large portion of our subscriber base was not yet able to handle UTF-8 records. Needless to say, CDS subscribers are **not** small public libraries of the type you might not expect to have the newest system capabilities, so we were a little surprised.

Some other recent experiences and infelicities have pointed out that romanized text may be the only consistent lingua franca for remote searchers of our catalog. We're trying to cope with situations where searching by script alone is completely unsatisfactory because of flaws (sometimes major) in the Microsoft IMEs that we use to input records or users use to input searches. We also have a system that is currently incabable of advanced indexing techniques-- for example, our system cannot normalize Chinese ideographs for both simplified and traditional characters into a single index, and the story gets even worse for 'compatible' characters covered by the CJK Compatibility Database on our website. We've been working on some OPAC help screens to alert users to some of these issues related to searching scripts, but the results won't be very good (even if you can get people to look at your help screens!).

We're also very concerned about languages/scripts outside of the MARC-8 repertoire of UTF-8 (that is, other than Chinese, Japanese, Korean, Arabic, Persian, Hebrew, Yiddish, Cyrillic, Greek)-- very few of the scripts we may wish to expand to (presuming our customers become UTF-8 compliant!) are currently impossible to input into our local system due to a bug that renders many Microsoft IMEs unusable. Romanization, sadly, will be the only way to render those records for some time to come. Couple that with Glenn's point that the romanization schemes used outside of the Anglo-American library community are probably different (e.g., ISO) and it is not a rosy picture.

I can't recall at this point whether anyone has mentioned culturally-sensitive sorts in non-Latin languages and scripts. I've heard folks mutter this phrase for years, but have no idea whether the sorts have ever been developed and how they might be implemented (our system does not use them, for sure). Sorting is another of the things we still rely on romanized data for.

Sorry for having blabbed on so long and probably not saying anything new.

Dave Reser

Footnote to Dave Reser's message

Joan Biella (non-member)
Fri, 07/31/2009 - 8:26am

Just a note that the case of Persian is especially bad, though "bad" doesn't seem like a strong enough word. The Microsoft Farsi IME lacks some common and necessary characters, which must be created by workarounds in cataloging and (as far as I know) can't be input at all by searchers--so searching in nonroman will always be largely unsuccessful. And the ALA/LC Persian romanization system is roundly criticized by every Persian-speaker I've ever met or heard of, who say no one who knows the language would ever search by current romanizations.

Neither Model A nor Model B can help this situation! Perhaps a sub-group should be formed just to consider these problems.

Joan Biella

Joan Biella
Israel/Judaica Section
Library of Congress

future automation developments

Joyce Bell
Tue, 08/04/2009 - 2:32pm

Sarah mentions "automatic transliteration software". These new sorts of technology are a really interesting development, but they differ by script. It sounds like Sarah is saying that CJK can be romanized automatically based on original script. Arabic and Persian can have original script reconstructed based on romanized text. Technology needs to advance much further for Middle Eastern languages to be romanized based on original script.

I do wonder, however, if we might foresee future developments and anticipate them in making some of our decisions. Just as RDA is forging into new areas more or less assuming that structures will build up around it, having a group planning for the future of romanization, can we not forge ahead anticipating a future whose outline is only just visible at present?

In an environment with shrinking staffs and production pressures, can we afford the duplicative effort to provide both transliteration and romanization in areas where it does not directly affect access or usability? Just as you can plug a web page into a translator site on the web and get a reasonable translation, can we assume a future in which conversions to the American transliteration system or the French transliteration system for a script are performed as needed by some sort of plug-in?

I am a fan of Model B, or some sort of hybrid along the lines Sarah has mentioned.

transliteration at LC

Joan Biella (non-member)
Wed, 08/05/2009 - 6:39am

Another FYI--at the Library of Congress, we now use software which, given Chinese, Arabic, Cyrillic, or (I think) Korean ALA/LC romanization, can spit out good nonroman parallel fields equipped with all needed indicators and subfielding. (Naturally, a small amount of massage is often needed after the initial spitting-out.) The needed Unicode Formatting Characters for right-to-left display of Arabic script are also provided automatically, which is a great boon to catalogers who need them.

A similar program which can spit out Hebrew script given Hebrew ALA/LC romanization is now in the works at LC--I'm told it's already better than the much-maligned OCLC service, which I haven't had to work with. Unfortunately, Hebrew is so romanized (with many possible clues omitted) that zapping Hebrew script from it is a much harder task, with accordingly severe massaging still needed at this point.

Joan

Joan Biella
Israel/Judaica Section
Library of Congress

Unfortunately, as I said

Keiko Suzuki
Wed, 08/05/2009 - 7:35am

Unfortunately, as I said before for the case of Japanese, I don't think it's easy to create a feasible automatic transliteration program, and haven't seen or heard anything like that yet. If you do, please let me know!!

I agree with Joyce that we might need to come up with short-term and long-term recommendations since we are really in transition. RDA might really change the structures of records to more FRBR-ized one and we might move out of MARC although I would think that will still take a long time.

I also agree with the efficiency of Non-Latin scripts cataloging would improve dramatically if we stop doing transliteration, completely or partially. However, I still not convinced where are "the duplicative effort to provide both transliteration and romanization in areas where it does not directly affect access or usability". I have started to ask around my East Asia non-cataloging colleagues, but so far, they seem to think it's a disservice not having transliteration. I'll keep asking more.

- Keiko

Some comments on the comments on Model A and Model B

Sandra Nugraha (non-member)
Fri, 09/04/2009 - 5:53am

I prefer to use Model A as shown in the example posted by Robert.
Romanization or pinyin for CJK is absolutely needed by the users who do not read the language. But the model should be without 880 field. So, it could be simple. The original language could either come first or later but it should be in parallel just exactly the same as :

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,
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.

The parallel works should also be added for the headings or added entry.

Automatic transliteration software mentioned by Sarah is certainly a great idea. I know there is a soft ware for automatic transliteration, which is for some one who learn the language. But it could be applied also to the library system. So, once we type the original language, the original language and the Romanization/ pinyin will appear at the same time.

By adding Romanization/pinyin it could help other librarian who does not read the language to find the title. It is right that the staff can search from ISBN or ISSN, but remember, some titles do not have ISBN/ ISSN.

Who needs romanization?

by Robert Rendall on Tue, 08/11/2009 - 4:56pm

In our discussion so far various needs for romanization have been mentioned. This is a summary and an invitation to comment further on these issues.

Users who can't read the original script

Romanization can help staff who can't read the original script work with library materials for various purposes (acquisitions, ILL requests, storage retrieval requests).

Question: Is romanization really necessary for this, in addition to the ISBN/ISSN and the call number? If someone actually writes the romanized title on the title page at the time of cataloging that would give staff something to match with a romanized record later, but if it's not on the t.p. what use is the romanization in the record? Specific examples would be helpful. Can we get input from our public services colleagues on this?

Collocation of forms romanized the same way

Romanization provides collocation when the same word can be written differently in the original language. 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. Our systems are not yet sophisticated enough to interfile these original-script forms in their indexing. But by searching for the romanized form you get both variants.

Question: Is this comparable to the difference between "labor" and "labour" in English? We usually expect patrons to know to search both forms. Or is this a bigger problem than that?

Sorting

Doing a browse search for romanized text gets you an alphabetical list of results in the OPAC that you can profitably scroll through. Greek and Cyrillic search results also seem to be more or less correctly alphabetized in OCLC and in my OPAC, with a few exceptions. But CJK characters don't seem to be sorted in a meaningful way, as far as I can tell.

Question: Is this important enough to be a deciding issue?

Headings

All headings in the LC/NACO file and in LCSH are in romanized form, so they must appear in romanized form in bibliographic records. Name headings may now have original-script references in the authority record, but subject headings do not.

Question: Until/unless that changes, the question isn't whether romanization should be used for headings in bib. records, but rather: should there also be original-script parallel fields for these headings in bib. records? Why? And if so, what form should the original-script heading be in?

Added value

In the process of romanization, catalogers provide information not available in the original script. For example, the Arabic word تاريخهم ("their history") is pronounced - and is therefore romanized - tārīkhuhum or tārīkhihim in different grammatical contexts. And romanization requires the cataloger to determine and indicate which of the many possible readings of a Japanese character is correct in the case being transcribed. (Note that this produces the opposite of the collocation mentioned above – words or characters that appear identical in the original script are indexed differently when searched in romanization.)

Question: Wouldn't it be simpler just to transcribe what you see in the original script and be done with it? Is this necessary for access? (And isn't the cataloger just guessing anyway, sometimes?)

Systems that can't handle non-Roman script

Records with non-Roman script only are useless in systems that can't handle non-Roman script. And we've heard that those systems are still more common than you might think.

Question: Is this enough of a reason for everyone to go on romanizing all records for the time being? For how much longer?

Other

I've probably left some out, or there are others that haven't been mentioned yet – let us know!

Who needs romanization comments

Joyce Bell
Tue, 09/01/2009 - 7:31am

Robert, you have done a great job in summarizing these points.

Users who can't read the original script:

--I have heard from a colleague in the CJK area that citations are often given in romanized form in publications. Users come with those citations looking for help finding the material cited. With romanized records public service staff can help them even if they don't read the script themselves. I bring this up to be fair to an alternate viewpoint, but I don't agree. A public services staff person who doesn't know the script can do little to help such a patron beyond simply typing the data in as it appears, which the patron could easily do themselves. This argument holds no weight in the HAPY area where there is not a widely accepted romanization system--there are almost as many romanization systems as there are publications!

Collocation of forms romanized the same way:

--This is not an issue with the HAPY languages.

Sorting:

--This sounds like an important issue for CJK. It isn't a problem with HAPY languages. Does CJK have a standard sorting order in other contexts that users would expect to find in an opac?

Headings:

--Yes, I think there should be parallel fields for headings in bib records. Partly for KW or left-anchored searching on script names. Partly for generating a complete display in script of the basic bibliographic description. Such fields should "match" the authorized heading.

Added value:

--Yes, it would be simpler with HAPY languages just to romanize what is actually there. But ... nobody actually does this. Of all the different romanizations I've seen of Arabic, I have very seldom run across any where they leave the vowels out. And yes, the cataloger is sometimes just guessing what the vowels are. Not frequently, but on occasion. This probably explains why I am pretty firmly in the camp of relying more on original script and less on romanization.

Systems that can't handle non-roman script:

--Do we have any idea of how many institutions with catalogs which can't handle non-roman script have collections of non-roman script material?

--JEB Sept. 1

Added value

Robert Rendall
Tue, 09/01/2009 - 3:41pm

Thanks for these comments! Just a clarification on "added value": I didn't mean to question whether vowels should be included in romanization. Romanization allows/forces the cataloger to provide information about the standard pronunciation of script forms that are pronounced differently in different contexts. That could be seen as an argument in favor of romanizing (as opposed to not romanizing, which was the alternative I had in mind). Providing romanization tells users whether in a given instance تاريخهم is supposed to be pronounced tārīkhuhum or tārīkhihim (kind of a silly example), or whether 中 is pronounced naka or chū (maybe a better example).

Added value additional comments

Joyce Bell
Thu, 09/17/2009 - 9:20am

Robert, you have put your finger on some REALLY sore points in the Arabic/Persian arena. Romanization can often be a very hotly-disputed topic for us.

Your sentence "romanization allows/forces the cataloger to provide information about the standard pronunciation of script forms that are pronounced differerently in different contexts" gets a bit more complicated for Arabic than your example. It is not just a question of pronouncing tarikhuhum vs. tarikhihim depending on grammatical context. It may be a case of pronouncing naft vs. nift depending on dialectal variations--we don't necessarily *have* a standard pronunciation.

The really sore spot is with the Persian romanization table. It is felt very strongly by some that the table is too influenced by Arabic, and it does not reflect pronounciation of Persian adequately at all. Deciding what the "correct" vocalization would be to match Persian pronounciation, however, introduces disagreements because of Persian dialectal variations.

The result is that rather than providing added value, romanization is playing favorites. It values one legitimate pronounciation over many other equally legitimate pronounciations. It can also limit search results by forcing users to account for grammatical differences (-uhum vs. -ihim) that aren't present in the original script.

needs assessment by script/language?

Joan Biella (non-member)
Thu, 09/24/2009 - 7:46am

This is a request that those on the committee or participants in the discussion with a special interest in Hebrew and Yiddish contact me at jbie@loc.gov. Perhaps we can come up with a statement of romanization pros and cons for these two languages and discern some common ground with other scripts and languages.

I look forward to hearing from you!

Joan

Joan Biella
Israel/Judaica Section
Library of Congress

RE: Systems that can't handle non-Roman script

Wing Kau Mak
Thu, 09/24/2009 - 7:52am

Sometime it is not whether a system can handle non-roman script or not, the issue may be the campus IT dept. doesn't want to support non-roman script input in public terminals. Take my workplace as an example, I can input CJK characters since I have administrative privilege to install CJK input keyboard in my PC (I am working in a Windows environment). However, patrons who use public terminals can't input CJK characters because they don't have the same administrative privilege to do the installation. Or they can install the input keyboard but the system requires a reboot after the installation. Since all newly written data to the hard drive will be wiped out once a public terminal is restarted, the patron is now back to square one. You may argue that we can ask campus IT to include input keyboard of different languages into the image that they use for public terminals. Then the question is how many kinds of input keyboard campus IT has to include in the image. For Simplified Chinese, there are four different kinds of keyboard available in Windows. For Traditional Chinese, there are eight.

Lucas Mak

Metadata & Catalog Librarian

Michigan State University Libraries

Model A and Model B: summary

by Robert Rendall on Wed, 09/16/2009 - 2:45pm

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). This refers primarily to the entry of descriptive fields. Headings present special issues of their own (see below).

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 transliteration of the title proper only.

When library catalogs were computerized, at first only Roman script could be used, so both descriptive and access fields had to be entered in romanization only. As the character set that could be used in online systems expanded, catalogers began to add original script to records as parallel fields, resulting in the development of Model A.

The adoption of Model B would result in simpler bibliographic records and more efficient cataloging. Romanization systems vary from country to country, and even the standard romanization systems we are supposed to use in North America are difficult to apply consistently, unfamiliar to native speakers, and sometimes controversial (Persian, Greek).

Model B is currently used in East Asian online catalogs, i.e. no attempt is made to “transliterate” English or French text into Korean or Japanese script. But Latin 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 Latin-script publications there does not have the same implications that it would have for CJK publications here.

The Working Group will need to decide 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. Continuing use of Model A would have to be justified by demonstrating the ongoing benefits of romanization; deferring adoption of Model B would have to be justified by indicating what the current obstacles are and how we expect them to be overcome.

Related questions are whether we could stop adding romanized parallel fields for some scripts and 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.

Please review the post on "Who needs romanization?" and add your comments. This will help us to determine what the consensus is on the questions raised there within the Working Group and among the members of the open discussion forum. Add any comments on this post at the bottom here. And if there are any other issues you feel we need to consider at this point, please initiate discussion of them in a comment here or in a separate post on this site.

Note on headings:

Since headings are established in romanized form in the LC/NAF, they need to be entered in bibliographic records in romanization. Current practice allows the addition of parallel heading fields in original script. 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). So original-script headings, unlike romanized ones, are never completely consistent, and result in split indexes in the catalog. Is this something we can live with indefinitely?

Model A vs. Model B

Heidi Lerner (non-member)
Thu, 09/17/2009 - 8:36pm

We are aware of the radical shift in opinion in the North American library community as to whether or not vernacular scripts belong in the catalog record. The resounding opinion is that cataloging data of materials in a particular language should be in the script or writing system in which the item is published.

In the early 1980s, after years of minimal romanization appearing in catalog cards produced by LC, LC began online cataloging of materials in non-Latin script languages, necessitating the complete romanization of these materials if they were to be cataloged in an online environment. This scenario changed in the latter part of the 1980s as both OCLC and RLIN introduced character sets for what became known as the JACKPHY languages enabling catalogers to transcribe bibliographic data as it appears on the piece in hand. As we know, over the next two decades more and more libraries began cataloging their non-Latin script materials with varying amounts of non-Latin script data according to what MARC21 calls Model A. The amount of non-script data appearing in these varies from language to language, institution to institution and cataloger to cataloger. An attempt at standardization is now in progress as a task force put together by the Program for Cooperative Cataloging is working on some draft-PCC Guidelines for Creating Bibliographic Records in Multiple Character Sets.

We are now being asked to decide whether or not we should continue to produce catalog records for non-Latin script materials according to Model A, or whether we should return to the older practice of pre-online cataloging and offer minimal romanization as per what MARC21 calls Model B.

I believe that to impose Model B on North American libraries at the present time would create havoc at best. For elimination of romanization to be practical, all library systems would need to be able to accommodate non-Latin scripts for both display and searching. Our experience at Stanford has shown that this is an expensive and complicated overhaul, necessitating many, many hours of staff time. If we found it cumbersome, what about those libraries whose resources are limited at best?

I also believe that dropping romanization will take away another layer of searching options. Once we take away something we rarely get it back. Generations of patrons and librarians have become used to searching for materials using the Latin alphabet and keyboard. I will not repeat the imperfections of romanization as that has already been eloquently done by many of our colleagues (variant standards, inconsistent application of standards, difficulty to apply romanization to languages such as Hebrew, Arabic and Persian … [et al.]), but having both options in the catalog record makes it easier for patrons and librarians to work with and access the materials that a library holds and collects.

Another question that arises if we immediately adopt Model B is what happens to those records that have romanized data only, or have the parallel fields of Model A. Automatic transliteration tools are

used successfully for some languages: some from Latin-script to non-Latin script and vice versa, others only work in one direction. But these systems remain largely imperfect and for some languages such as Hebrew woefully inadequate. So how do we get romanized only records converted to non-Latin script and what do we do with the roman script fields in records that follow Model A. I am sure that we do not want split catalogs for our non-Latin script materials.

If economics of adding both roman and non-Latin data were the only issue I would argue that it is only of minimal cost and time for catalogers to add both scripts. And with the addition of catalog records from libraries around the world that do catalog in their native scripts contributing records to OCLC, our work is even more reduced.

In sum, I believe that at some point we may very well transition to Model B but the time is not now.

Some problems with "nonroman only" (Model B)

Joan Biella (non-member)
Mon, 09/21/2009 - 12:11pm

Topic 1. Regarding nonroman parallel fields for headings:

As I remember the sequence of events, a very important selling point for adding nonroman script references to authority records was the carrot held out that, if they were added, it would no longer be necessary to provide nonroman parallel access points in bibliographic records. When the rubber met the road, however, and the project was undertaken to for prepopulate the LC/NACO Authority file with nonroman headings from OCLC, OCLC made it a requirement that parallel nonroman access must still be provided in bib records because OCLC has no mechanism for matching nonroman headings in bib records to established roman forms in the Authority File. Until such a mechanism is developed and everyone employs it, it seems the work in both files must continue … a considerable disappointment to my cataloging colleagues. (If my understanding of the reason for the disappointment is incorrect, please correct me!)

Topic 2. Shouldn't we just transcribe what we see in the original script and be done with it? Is a roman equivalent necessary for access? (And isn't the cataloger just guessing anyway, sometimes?)

Thoughts on Arabic and Hebrew in relation to these questions:

1a) Shouldn't we just transcribe what we see in the original script and be done with it? Arabic is simpler: maybe yes for Arabic. For Hebrew, the situation is more complicated, as most words have at least two possible orthographic representations, and there's not way to predict which a publisher will choose to employ on a given title page. One orthographic system provides extra consonantal y's, w's, and sometimes alefs to flesh out the normal lack of vowel representation, and the other doesn't give such help, or doesn't give so much of it. Without the item in hand, a librarian or patron can't guess how many y's, w's, alefs to include in his nonroman search, and if the phrase to be searched includes several words which can be written more or less fully, the number of roman searches needed to cover all possibilities can be quite high. Should catalogers "normalize" the nonroman spelling to the fuller or less full orthography, and trust patrons to know which system has been chosen? Should catalogers provide

double nonroman fields, one fuller and one less full, neither or which may correspond to the actual spelling on the item? Should catalogers provide triple nonroman fields, one fuller, one less full, and one accurate? The existence of a romanized field which "matches" all possible orthographies is a boon in searching—it is, in fact, the desired "normalization" of the competing orthographies.

1b) The spelling of personal names, especially surnames, is particularly difficult to guess, and may include use of yet another "helping vowel," 'ayin, along with w, y, and alef. Does the author spell his name Rwz'nb'rg, R'z'nb'rg, Rwznbrg, Rwzynbrg, or any of a number of other possibilities? Will "Jerusalem" be spelled Yrwšlym or Yrwšlm? And what is the Hebrew (or Yiddish) for "Lakewood, New Jersey"? Even a traditional Bible forename like Aharon may appear sometimes as 'hrn, sometimes as 'hrwn.

2) Isn't the cataloger just guessing, anyway? Not usually—hardly ever, in Arabic and Hebrew. To be sure, misromanizations sometimes occur when the cataloger is not deeply versed in the romanization rules, but I find mistakes in interpretation to be rare, except when the cataloger is not deeply versed in the grammar of the language.

3) What's to be done when a title page is "partially vocalized"—that is, when the publisher has provided the little marks usually omitted which actually nail down the romanization of each word—marks that are usually seen only in sacred texts (to prevent mispronunciation of holy words) or works for children who are just learning to read? Why don't we transcribe them, if they're present, in out nonroman fields? In my experience, the vowels provided on Hebrew materials are usually accurate, while those on Arabic materials are inaccurate, according to the "schoolbook grammar" of each language. Arabic publishers like to beautify their title pages with a lot decorative jots and tittles, some perhaps intended as vowel marks, some only resembling them, and others (we used to call them "birds") just there for fun. For just one common example, the normal word for "index," "fihris," is maddeningly almost always vocalized as "fahras" on title pages. What's the use?

Topic 3. Are all nonroman languages alike?

I note that in Robert's excellent presentation of topics already raised, his comments may often be paraphrased, "But is this consideration regarding romanization for [e.g., Japanese] sufficient to require everyone dealing with a nonroman language to provide searchable parallel fields?" Your answer to this question will probably be colored by the depth of your own experience in romanizing whichever language it is—e.g., Japanese. I doubt that any consideration for or against Japanese romanization would seriously influence my ideas about Arabic or Hebrew romanization.

I agree with the messages we've received that the provision of nonroman script fields for as many languages as possible as soon as possible is desirable, even needed. The tricky decisions concern whether searchable roman fields are ALSO needed in catalog records, and whether they're needed in ALL catalog records for nonroman languages.

As even the initial discussion in this group has shown, opinions vary greatly as to whether any fields need roman script equivalents, and if they do, which ones. It's also very clear already that catalogers of different scripts feel different needs, and that even within a single script (Arabic/Persian, for example), different languages require different treatments.

For some languages (Arabic perhaps?) with romanization schemes which match their scripts nearly character by character, automatic transliteration tools can be designed which make few errors. The situation for Hebrew is entirely different (as described above), since not only does the romanization system contain many ambiguous signs, but Hebrew orthography is not fixed. No matter how accurate in other ways a Hebrew transliteration system may be, it will not be able to deduce from a romanization whether the publisher chose a “full” or a “defective” orthography. In a retrospective project to add nonroman script to romanized Hebrew records, accurate transcription can be achieved only if every physical item is examined. Alternatively, a “normalized” Hebrew orthography could be imposed on such records—and librarians and patrons alike would have to be taught this system and its idiosyncrasies (I’m sure it would have some!) as they are now “taught” romanization. Hebrew romanization, with all its faults, already acts as a “normalizer” and this is one of its most useful features.

Heidi Lerner wrote, at the end of her recent message, “I believe that at some point we should and will [go] to a Model B but the time is not now.”

I agree with Heidi in this, but I can’t believe that the transition will ever be easy or that it could or should be identically planned for every script or language. Each language/script has unique problems. For some, romanization is useful, even very useful, as earlier comments in this group have shown; for others it is less useful or even not useful at all; and for still others no satisfactory and/or noncontroversial romanization scheme has yet been proposed or perhaps ever will be possible.

I suggest that one of the best things members of this group could do is pool their expertise on individual languages and prepare a document for each listing the pros and cons of discarding, retaining, or modifying its present romanization scheme and the value of continuing to provide searchable nonroman parallel fields in bibliographic records for works in the particular language. I’d like to volunteer to work on such documents for Hebrew and Yiddish, and I have some familiarity with Arabic problems as well.

When this has been done, perhaps it will be easier for us to discern if there are general recommendations we can make for treatment of all languages, or if individualized recommendations are unavoidable.

Joan Biella
Israel/Judaica Section
Library of Congress

Re: Some problems with "nonroman only" (Model B)

Robert Rendall
Mon, 09/21/2009 - 1:59pm

Thanks for these comments! Some quick responses:

Topic 1: parallel fields for headings

Does OCLC really require catalogers to add parallel non-Roman fields for headings in any formal way? Certainly it's a good idea when the romanized form is ambiguous (e.g. most Chinese names!). Current PCC documentation is contradictory: for Arabic CONSER records with parallel descriptive fields, parallel

heading fields are required, but for CJK CONSER records with parallel descriptive fields, parallel heading fields are optional (see the CONSER Editing Guide appendices E and O). The new general draft guidelines for PCC released recently make these fields optional.

Topic 2: why not just transcribe what you see?

“Guessing” at the correct romanization is common enough for Arabic script, I think. From a previous job I remember Middle Eastern catalogers around the corner from me agonizing over what vowels to supply for unvocalized Sudanese names of non-Semitic origin, and recently on the MidEastCat list there have been several postings asking for help romanizing unusual Moroccan names or vocabulary. This sort of dilemma certainly slows down cataloging when it happens, and romanizations resulting from guesswork are probably not very useful for any purpose…

Topic 3: are all nonroman languages alike?

You’re right that discussion so far has made it clear that different languages and scripts raise very different issues, and making a list of romanization pros and cons by language/script could be a useful way to move our discussion forward. But as for modifying present romanization schemes, I think making specific recommendations on that might be out of scope for us. If catalogers of Persian or Greek haven’t yet been able to agree whether or how to change the systems they currently use, I don’t think there’s much that a group like ours composed mostly of outsiders can do to help them.

comments on comments on "problems with nonroman only"

Joan Biella (non-member)
Tue, 09/22/2009 - 8:52am

Does OCLC really require catalogers to add parallel non-Roman fields for headings in any formal way?

I imagine (does anyone here really know?) that OCLC's action was along the lines of refusing to absolve catalogers in a formal way from the need to add parallel nonroman fields for headings.

Current PCC documentation is contradictory [from language to language about whether to include parallel headings].

Let's not think of this variation as "contradictory." How 'bout phrasing it as "Present CONSER practices vary from language to language and/or script to script"? In my own (pessimistic?) view, this will always have to be true.

“Guessing” at the correct romanization is common enough for Arabic script, I think.

Personal names and nonstandard dialect words are problematic in every language, I think. (And in many cases the patron will share the cataloger's bafflement.) It would be a pity to make their existence a showstopper for romanization if romanization also offers advantages.

As for modifying present romanization schemes, I think making specific recommendations on that might be out of scope for us.

I didn't mean to suggest that this group offer proposals for modified romanization schemes. I meant that we might document controversies and proposed solutions that already exist in the relevant communities. I'd like to see a relatively cut and dried description of the problem(s) with Persian romanization, for

example. Sometimes it sounds to me as if the problem in very large part boils down to dissatisfaction with the suffix "-ah" as opposed to "-eh."

Joan Biella
Israel/Judaica Section
Library of Congress

Does OCLC really require parallel fields?

Glenn Patton
Thu, 09/24/2009 - 1:19pm

To answer Joan's question ("Does anyone here really know?"), I certainly do know. OCLC does not require that there be parallel sets of romanized and non-Latin script fields (headings or otherwise). Years ago, OCLC systems did indeed require that, if there was a non-Latin script field, there had to be a romanized field linked to it. But, about the time that we introduced support for Arabic script (and added the capability to export a record containing only non-Latin script fields), we removed that restriction.

One further comment about the relationship between non-Latin references in authority records and parallel fields for headings in bibliographic records: I'm afraid that my memory of the discussions within the NACO partners that preceded the addition of non-Latin script references to authority records differs. The concern I expressed on OCLC's behalf during the discussions about whether, on completion of that project, catalogers could stop including parallel fields for headings is one that has already been expressed (I think) in this group's discussion: providing users with full access to records without parallel fields for headings can work only if authority data is fully integrated into the searching process. If the system in use does not fully integrate authority data (and OCLC's various interfaces to WorldCat are among many systems that do not), then access is lost if parallel fields are not maintained.

--Glenn

New article in Cataloging & Classification Quarterly

Glenn Patton
Tue, 09/29/2009 - 7:56am

I thought working group members may be interested in an article in the most recent issue of *Cataloging & Classification Quarterly*:

Seikel, Michele (2009) 'No More Romanizing: The Attempt to Be Less Anglocentric in RDA', Cataloging & Classification Quarterly, 47:8, 741- 748.

--Glenn

It was an interesting read,

Keiko Suzuki
Wed, 09/30/2009 - 3:24pm

It was an interesting read, but I have to go back to check with RDA Nov 2008 draft and read it again since my comparison of AACR2 and RDA had a bit different conclusion. In any case, the part particularly interesting for me was Agenbroad's quote at the beginning of the conclusion: "To provide equal and effective access, a library with a multiscript collection for readers of different scripts needs a multiscript

catalog based on cataloging rules that specify access in the original script. It may also need access via the script most used by most of its readers." I kind of agree with him. I often think about "who are our users/readers?" In my academic/research library setting, most likely my library users for Japanese language materials I catalog are non-native speakers but they can read/write original scripts. Still, many of them are more comfortable or easier with transliteration. And then the larger library users are English speakers. So for me as a Japanese cataloger, it's very important to supply transliteration for key descriptions and access points.

I agree with others that, although Model B is simpler, doesn't mean we would give up transliteration all together, and eventually we are moving to the direction, this is not the time yet. I feel now is really transition period. We might move to RDA, and that might lead to non-MARC environment in future. I would like to wait until the test result of RDA comes out to start talking about how Model B works for us. Also I don't think we could overcome some technical issues to have completely original script description records (for some languages/scripts?) before that, anyway. OR maybe we could do more extensive community survey to our colleagues and users, by languages/scripts, meanwhile?

General summary and possible recommendations

by Robert Rendall on Fri, 10/16/2009 - 3:33pm

This is an attempt to summarize our discussions in a single text and present some possible conclusions. If we think this covers all the right topics, it could be developed into the first draft of our report. Obviously a lot of this is just cut and pasted from your comments and the draft would need further editing and better organization. Offers to rewrite whole sections and suggestions about how to organize all this differently are welcome. This is just what I've managed to pull together before going on vacation to England for the next week and a half!

I would like to publicize our draft report in November on the [nonenglishaccess] list or elsewhere before we produce a final version. We need wider input, particularly on the perceived advantages of having romanization in records (including language/script-specific issues) and on how common systems that cannot handle non-Roman script are. At a minimum we could ask people to respond to the first draft of our report when we post it; we could also compose a specific (short) questionnaire to go along with it if we think that would be useful.

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. Continuing use of Model A would have to be justified by demonstrating the ongoing benefits of romanization; deferring adoption of Model B would have to be justified by indicating what the current obstacles are and how we expect them to be overcome.

Related questions are 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.

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 transliterated fields. These are coded as 880 fields at the end of the bibliographic record, but in public display (and sometimes in staff display, as in OCLC Connexion) they display next to the corresponding transliterated field.

Model A

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,
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.

Model B

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,
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.

700 1# Βενιζελος, Ελευθεριος, ǂd 1864-1936.
700 1# Venizelos, Eleutherios, ǂd 1864-1936.

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 transliteration of the title proper only.

When library catalogs were computerized, at first only Roman script could be used, so both descriptive and access fields had to be entered in romanization only. This scenario changed in the 1980s when both OCLC and RLIN began to introduce character sets for what became known as the JACKPHY languages enabling catalogers to transcribe bibliographic data as it appears on the piece in hand. Since then libraries have cataloged their non-Roman script materials with full romanization and varying amounts of non-Roman script data in parallel fields (Model A). The amount of non-Roman script data appearing in these varies from language to language, institution to institution and cataloger to cataloger. 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.

Model B is currently used in East Asian online catalogs, i.e. no attempt is made to “transliterate” English or French text into Korean or Japanese script. But Latin 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 Latin-script publications there does not have the same implications that the use of Model B for CJK publications would have here.

Questioning Model A

In the days of the card catalog, catalogers were able to enter original script in catalog card records (Model B). That option was temporarily lost when we moved to online catalogs, but the advent of Unicode has allowed us to resume use of non-Roman script in catalog records, although we have done 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

(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 romanization introduces errors. Romanization systems vary from country to country, and even the standard romanization systems we are supposed to use in North America are difficult to apply consistently, unfamiliar to native speakers, and sometimes controversial (Persian, Greek).

Problems with 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, MARC21 (unlike UNIMARC) has no way of indicating in the bib record which romanization practice has been applied.

For many languages, even experts differ on the correct romanization of many words. Hebrew and Arabic are generally printed without vowels in the vernacular, so there is a certain 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 may seem arbitrary or controversial. For the Arabic word نفط, the standard romanization used by LC is nafṭ, but many Arabic speakers might prefer nifṭ. Romanization is in a sense playing favorites. It values one legitimate pronounciation over other equally legitimate pronounciations.

An additional complication with Hebrew and Arabic script is provided by “partially vocalized” title pages, where the publisher has provided the vowel marks usually seen only in sacred texts or works for children who are just learning to read. 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, while those on Arabic materials often do not correspond to the vocalization recommended by standard sources. The Arabic word for “index,” فهرس, is often vocalized as “fahras” on title pages, but the standard romanization is “fihris.” So standard romanization can require catalogers to use vowels different from those explicitly indicated on the piece.

Romanization errors can occur when the cataloger is not deeply versed in the romanization rules, or when the cataloger is not deeply versed in the grammar of the language. In addition, 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 serves no clearly useful purpose.

Entire romanization systems can be problematic. The ALA/LC Persian romanization system is frequently criticized by Persian speakers who say no one who knows the language would ever search by current romanizations. It is felt very strongly by some that the table is too influenced by Arabic. 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.

Advantages of romanization

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

1. Users who can't read the original script

It is often suggested that romanization can help staff and patrons who can't read non-Roman script work with library materials in these scripts for various purposes (acquisitions, ILL requests, storage retrieval requests, assembling bibliographies).

In principle, romanization seems 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 can't read the original script, how does the romanization in the bib record help the staff member match the item in hand with the bib 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 up a book or serial issue with a bib record rather than trying to use tables to transliterate a non-Roman script they don't know (and even that would be impossible for non-alphabetic scripts). However, not all titles have an ISBN or ISSN, and items not yet cataloged do not have a call number.

We have heard that 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 for items that are already cataloged staff can retrieve them from the stacks and match the romanization in the bib. record against the form on the title page to confirm that they have the right piece.

It has been reported that in the CJK area citations are often given in romanized form in publications. Users come to the library with those citations looking for help finding the material cited. With romanized records public service staff can help them even if they don't read the script themselves. But a public services staff person who doesn't know the script can do little to help such a patron beyond simply typing the data in as it appears, which the patron could easily do themselves. And while the Chinese or Japanese transliteration systems used in libraries may be widely used in non-library contexts as well, in the HAPY area there is no widely accepted romanization system and any romanized data provided by a patron is unlikely to be in the system used by the library.

Romanization provides additional access points for those who might prefer to use them. 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 transliteration. And in some cases a romanized search may be easier to input that an original-script one, even for users who can read the original script (see problems with input below).

2. Collocation of forms romanized the same way

Romanization provides collocation when the same word can be written in different ways in the original language. 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-MARC21 characters in users' searches with the equivalent MARC21 forms (as given in LC's CJK Compatibility Database) that catalogers have to use to represent them in bibliographic records. But 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 can't 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 roman searches needed to cover all possibilities can be quite high. The family name transliterated "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 "normalized" spelling so that all variants are retrieved when a romanized search is performed.

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), where sorting by Unicode code point generally produces a list in alphabetical order. But culturally-sensitive sorts have not yet been developed in library systems for non-alphabetic languages and scripts. For CJK, sorting by code point (the current effect of a browse search) does not produce acceptable results. The sorting orders that would be meaningful to native speakers are by radical and stroke number, or by Latin transliteration. The former would be difficult to implement; romanization provides the latter.

Is this important enough to be a deciding issue?

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 example, romanization requires the cataloger to determine and indicate which of the many possible readings of a Japanese 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>.)

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 lot of time to determine 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 transliterating, especially for undergraduates or beginners, or researchers whose are not experts in these languages but need to work with materials written in them and have some ability to do so.

But it is not clear whether providing this sort of information should be seen as an essential function for a cataloger. It would certainly be simpler just to transcribe the original script as it appears on the piece. And users who search using romanization 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.

Systems that can’t handle non-Roman script (storage, display)

Records with non-Roman script only are useless in systems that can’t handle non-Roman script. And those systems are still quite common.

A 2007 Cataloging Distribution Service survey related to character sets in MARC records found that a large portion of their subscriber base was not yet able to handle UTF-8 records.

In addition, many languages and scripts outside of the MARC-8 repertoire of UTF-8 (that is, other than Chinese, Japanese, Korean, Arabic, Persian, Hebrew, Yiddish, Cyrillic, Greek) 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 to enter and distribute those records for some time to come.

Even CJK has some characters not covered by Unicode, so it is not yet possible to transcribe original script in every case.

Another concern is whether the software packages that libraries provide for creating bibliographies are able to deal with Unicode characters.

For elimination of romanization to be practical, all library systems would need to be able to accommodate non-Roman scripts for both display and searching. This can be an expensive and complicated overhaul, necessitating many hours of staff time. If this has been a difficult process for large libraries, what about those libraries whose resources are limited at best?

But: is this enough of a reason for everyone to go on romanizing all records for the time being? For how much longer? Would we eventually reach a critical mass of libraries that can handle non-Roman script, with a few others left behind?

Systems that can’t handle non-Roman script (input)

Another technical problem is not whether a system can handle non-Roman script, but whether public library terminals (or users’ personal computers) allow non-Roman script input. Even if input is supported, different users might need a variety of keyboards, depending on what input method they are used to. 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 available, even on the library's own terminals.

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

Headings

Since headings are established in romanized form in the LC/NAF, they need to be entered in bibliographic records in romanization. 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 in original script. For example, in current PCC documentation (now under revision), 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. Are these parallel heading fields still needed, now that the same non-Roman forms can be added to authority records? And if so, what form should the original-script heading be in?

Parallel fields for headings in bib records are still necessary for keyword or left-anchored searching on script names. 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.

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. Is this something we can live with indefinitely?

One of the perceived advantages of adding non-Roman script 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 for 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 fields for headings can work only if authority data is fully integrated into the searching process. If the system in use does not fully integrate authority data (and many systems that do not), then access is lost if parallel fields are not maintained.

Automation of romanization

The effort required to provide romanization in bibliographic records could theoretically be reduced by automation, although some human checking for minor variations will always be necessary. The extent to which romanization can actually be automated varies by script. Conversion from original script to transliteration could be automated fairly easily for Cyrillic and (with the exception of the rough breathing) for Greek. Chinese could also be romanized automatically based on original script, but it would be very difficult to automate romanization of Japanese (because of the many possible readings for each character). Korean might also be problematic because of the complexity of the romanization rules.

Because vowels are not indicated in normal orthography, HAPY languages cannot be automatically romanized from original script. However, Arabic and Persian can have original script automatically reconstructed if the romanization is entered first. Since their romanization schemes match their scripts nearly character by character, automatic tools can be designed which make few errors. But this is not possible for Hebrew, since not only does the romanization system contain many ambiguous signs, but Hebrew orthography is not fixed. No matter how accurate in other ways a Hebrew transliteration system may be, it will not be able to deduce from a romanization whether the publisher chose a "full" or a "defective" orthography.

Model A & B in one catalog

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 Western language 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 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. We would have permanently split catalogs for our non-Roman script materials. Would this be unacceptable?

POSSIBLE CONCLUSIONS:

[These recommendations would be for libraries contributing cataloging in a shared environment. Locally they could make other decisions.]

1. Model B should be adopted soon:

We should foresee future developments and anticipate them in making some of our decisions. Just as RDA is forging into new areas more or less assuming that structures will build up around it, having a group planning for the future of romanization, can we not forge ahead anticipating a future whose outline is only just visible at present? In an environment with shrinking staffs and production pressures, can we afford the duplicative effort to provide both transliteration and romanization in areas where it does not directly affect access or usability? Just as you can plug a web page into a translator site on the web and get a reasonable translation, can we assume a future in which conversions to the American transliteration system or the French transliteration system for a script are performed as needed by some sort of plug-in?

2. Model A should be maintained for the time being:

We may need to come up with short-term and long-term recommendations, since we are in a period of transition. RDA and a possible move away from MARC may change the structure of our records. We should wait until the (presumed) implementation of RDA before considering this kind of change. And the current technical issues (sorting, indexing, unsupported characters and scripts, input problems) are significant enough to make a shift to Model B premature. At some point we may very well transition to Model B but the time is not now.

3. Model A should be maintained indefinitely:

Dropping romanization will take away an important layer of searching options. Once we take away something we rarely get it back. Generations of patrons and librarians have become used to searching for materials using the Latin alphabet and keyboard. Having both options in the catalog record makes it easier for patrons and librarians to work with and access the materials that a library holds and collects. In most cases, adding romanization to a Model A record does not require excessive time or effort. Specific problems and inconsistencies in current romanization practice are not significant enough to outweigh the more general advantages of having romanized access points in records.

4. Model A could be retained but with fewer romanized fields:

[If all systems are capable of handling non-Roman script in the future (all systems Unicode-based; improved indexing of non-Roman scripts; mapping tables to link different romanization systems and writing variants within the same language; authority data fully integrated into the searching process):]

We could reduce the amount of romanization in records. We could enter all descriptive fields in the original language/script of the work while adding additional access points for key data in romanized form. For example, a romanized Chinese title could be entered in 246 as a variant title instead of 880. Automatic transliteration software could be utilized to reduce time needed to create the romanization, when possible. Institutions could choose to suppress the romanization from display if desired.

5. Some libraries could move to Model B while others continue to use Model A:

Model A and model B may coexist for a long time to come. But the fact that some libraries can't support Model B should not deter other libraries that are able to use it from doing so. In that case, Model A libraries using copy created by Model B libraries would have to add romanized fields to the records as they encounter them.

6. Some language/script cataloging communities could move to Model B while others continue to use Model A:

Different languages and scripts raise very different issues. For some, romanization is useful, even very useful; for others it is less useful or even not useful at all; and for still others no satisfactory and/or uncontroversial romanization scheme has yet been proposed or perhaps ever will be possible. Opinions vary greatly as to whether any fields need roman script equivalents, and if they do, which ones. Catalogers of different scripts feel different needs, and even within a single script (Arabic/Persian, for example), different languages require different treatments. It is unreasonable to expect the romanization issues of e.g. Japanese to affect the decisions made by catalogers of Arabic or Hebrew. We may gradually move to Model B, but the transition will not be easy and it should not be identically planned for every script or language.

[I don't think this group can recommend that e.g. Arabic romanization should stop and Hebrew romanization should continue. But we could recommend that romanization should no longer be generally required for all scripts, and that it would be acceptable for individual cataloging communities to make different best-practice decisions. Then the actual decisions would be up to them.]

7. Different "levels" of romanization practice should be adopted:

This idea is mentioned in our charge, but it hasn't come up in our discussions so far, unless you want to call the two previous options different "levels" of romanization.

I would like to add about

Sandra Nugraha (non-member)
Tue, 12/01/2009 - 10:12pm

I would like to add about automatic transliteration. We have been using the transliterator to add Chinese characters. It has no problem at all.