

Expanding Non-Latin script cataloguing in Libraries Australia — Cyrillic, Greek and Tamil

Following the announcement on 25 August 2006 of support for Thai script cataloguing in Libraries Australia, Libraries Australia has investigated support for other non-Latin scripts. We are now pleased to announce that Libraries Australia can fully support Cyrillic, Greek and Tamil script cataloguing.

Currently the Australian National Bibliographic Database (ANBD) has just 250 Cyrillic script records, which were originally from the Library of Congress, OCLC and RLIN and have been added to ANBD through Libraries Australia's Record Import Service. No records have been found in ANBD with Greek and Tamil scripts. Now Cyrillic, Greek and Tamil script records can be created online in Libraries Australia.

Libraries Australia will continue phasing in support for other non-Latin script cataloguing. Target languages on the horizon are Arabic and Hebrew. These may provide some challenges, as Arabic and Hebrew characters are written from right to left.

Some information about Cyrillic, Greek and Tamil script cataloguing

1. Script display

You need to install the required fonts on your PC. We recommend using the *Arial Unicode MS* font. The font is available as part of Microsoft Office and should already be installed on your PC. You can view all non-Latin scripts in both the Libraries Australia Search (LAS) or in the Libraries Australia Cataloguing Client (LACC).

Unfortunately Tamil digits, numerics, and symbols cannot be fully displayed using the *Arial Unicode MS* font.

2. Script input

If the default language of your PC is not the language you need for non-Latin script cataloguing (i.e. Cyrillic, Greek or Tamil), or if you do not already have an input method for the language set up on your PC, we recommend that you install input languages and input devices. Windows provides such devices: the Input Method Editors (IMEs) and the On-Screen Keyboard or the Microsoft Visual Keyboard.

Information about IMEs can be found in *Libraries Australia Cataloguing Client Manual. Appendix 9: Keyboard layouts and Input Method Editors*. A short article entitled *Windows Internationalization Tips* (http://www.ficorp.com/intl_xp.html) provides useful step-by-step instructions on how to add other input languages to your PC.

You can follow the document entitled *Installing the Classical Greek Keyboard* (<http://www.jcu.edu/language/lc/keyboard-setup-greek.htm>) if you need to handle classical Greek characters.

While going through the “add language” step in the IME installation process, you will see only one Tamil or one Greek language entry, but quite a few varieties of Cyrillic listed - Azeri Cyrillic, Kyrgyz Cyrillic, Mongolian Cyrillic, Serbian Cyrillic, Uzbek Cyrillic and Russian. These country-specific versions of Cyrillic reflect dialectical differences that might be enabled on the keyboard. Since not all letters in the Cyrillic alphabet are used in every Cyrillic language, it is necessary to select the appropriate language as the input language, e.g. Russian, Serbian Cyrillic or Mongolian Cyrillic.

After installing these input languages, e.g. Tamil, Russian or Greek, you can then select one of these languages by changing the active keyboard on the Taskbar from English to either Tamil, Russian or Greek (click on the blue “EN” on the Taskbar to see the languages that you have installed, and then select “TA” for Tamil, “RU” for Russian, “EL” for Greek). You can then use the keyboard to enter one of these languages.

For those who are not confident about using the existing keyboard to enter other non-Latin scripts, the On-Screen Keyboard or the Microsoft Visual Keyboard can be an effective alternative.

The On-Screen Keyboard should have already been installed in your PC. To open it, you just click on the Windows **Start** button > **Programs** > **Accessories** > **Accessibility** > **On-Screen Keyboard**.

The Microsoft Visual Keyboard is free software. Download and installation information can be found on the Microsoft web site (<http://www.microsoft.com/downloads/details...&displaylang=EN>).

We have tested the two keyboards and found the On-Screen Keyboard slightly easier to use.

When you have activated the On-Screen keyboard and have changed “EN” to “TA” “RU” or “EL”, you will see the English letters in the On-Screen keyboard instantly change to Tamil, Russian or Greek. Now you can enter Tamil, Russian or Greek characters by pressing the appropriate keys.

3. Indexing and searching

If Cyrillic, Greek and Tamil script characters are present in indexed fields, they will be searchable in both LACC and LAS. All LACC and LAS search indexes are applicable.

However, because of the complexity of non-Latin script indexing and searching, we recommend that you use standard Romanization to search. For more details see *Libraries Australia Cataloguing Client Manual. 3.2.16 - Scripts or Romanised searching*

4. Cataloguing

A. Creating records

You can create script records in your local system (and upload them to the ANBD), use the LACC to catalogue directly online, or import records from external databases via Libraries Australia Search. The Libraries Australia web input form does not support non-Latin script cataloguing.

B. Standards

Records created in LACC or in your local system must meet MARC standards, no matter what script you use to enter the data. You must catalogue according to AACR2 and use headings either in or conforming to LCNA and LCSH. For more information, see: *Libraries Australia Cataloguing Standards* (<http://www.nla.gov.au/librariesaustralia/standards.html>)

C. MARC formats

Three MARC formats may be used to create bibliographical records in LACC and in local systems that have multiscript support (eg. Latin and Cyrillic, Tamil or Greek). They include:

Model A: PICA MARC format. It is used only in LACC.

Mode B: MARC 21 format for library systems that support Unicode. As more local library systems in Australia move to full Unicode support, this model should be used more widely at the local level.

Model C: MARC 21 format for library systems that support non-Latin scripts in the MARC-8 repertoire. We believe only a handful of local libraries have this support. This model is commonly used for CJK script cataloguing but has a very limited use for Cyrillic or Greek script cataloguing.

See: *Appendix A: MARC formats* for details for these three models.

D. Romanized data

In MARC 21, data is recorded in the appropriate MARC fields using Romanized words. There are different Romanization schemes for different scripts. Libraries in the English speaking world must follow the guidelines provided by the American Library Association and the Library of Congress. Romanization tables for Cyrillic, Greek and Tamil can be found in the *ALA-LC Romanization Tables* on the Library of Congress website (<http://www.loc.gov/catdir/cpsd/roman.html>)

E. Mandatory script data fields

You can decide what level of record you would like to create. For minimum level records, you must adhere to the *Libraries Australia Minimum Record Standard* (<http://www.nla.gov.au/librariesaustralia/network.html#minimum>). In addition, script data must be present in at least the 245 field.

To create higher level records, please consult *Appendix B: In which MARC fields should you add non-Latin script?*

F. Sources for copy cataloguing

Libraries in the western world with Cyrillic, Greek and Tamil holdings mostly use Romanization in their catalogues. The Library of Congress, OCLC and RLIN have only recently provided Cyrillic script support, and Tamil script cataloguing was only implemented in OCLC's Connexion in July this year. Naturally records that have Cyrillic, Greek and Tamil scripts can be found in the home country libraries.

We have found some Cyrillic, Greek and Tamil records in the following databases:

Greek records in:

Hellenic Academic Libraries (http://argo.ekt.gr/Opac2_3/zConnectENU.html)

Hellenic Public Libraries (http://argo.ekt.gr/Opac2_4/zConnectENU.html)

Cyrillic records in:

Russian Library Consortia Resources

(<http://www.ruslan.ru:8001/rcls/resources/>)

Tamil records in:

SILAS (Singapore Union Catalogue) Available via Libraries Australia Search

Libraries Australia will investigate the availability of databases which allow importing of records for copy cataloguing.

5. Data exchange between Libraries Australia and local systems

Two character encoding schemes, MARC-8 and Unicode, are used in bibliographic and authority MARC records. Cyrillic and Greek can be found in either MARC-8 or Unicode repertoires, but Tamil can only be found in

Unicode. Libraries Australia supports Unicode but it also provides the choice of Unicode or MARC-8 as data formats for record import or export in MARC 21 format.

The MARC-8 repertoire includes the following scripts: Latin, Arabic, Chinese, Cyrillic, Greek, Japanese, Korean, Persian, Hebrew and Yiddish. Tamil script is outside the MARC8 repertoire. The Unicode repertoire covers many more Latin and non-Latin scripts, and theoretically covers all languages in the world.

The MARC-8 character encodings for Cyrillic and Greek scripts can be found in *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media: Code Tables* at <http://www.loc.gov/marc/specifications/specchartables.html>

The Unicode character encodings for Cyrillic, Greek and Tamil can be found at the Unicode website: *Unicode Character Code Charts by Script* in Unicode Web at <http://www.unicode.org/charts/>

Libraries Australia supports full Unicode. Some local library systems also support full Unicode, but the majority of libraries in Australia are MARC-8 based and support Latin script only. We believe only a handful of Australian libraries are MARC-8 based but also support non-Latin scripts.

If your library system is a Unicode-based system, you will be able to create non-Latin script records in your system, and import or export non-Latin script records between your local system and Libraries Australia using Unicode as the data exchange format.

If your library system is a MARC-8 based system that only supports Latin script, firstly you are unable to create non-Latin script records. Secondly, if you import non-Latin script records from Libraries Australia, no matter what data format you are using, the non-Latin characters of the records imported to your local system will be unintelligible.

If your library system is a MARC-8 based system that supports non-Latin scripts, you can create records containing the non-Latin scripts that your system supports. You can export those non-Latin script records from your system to Libraries Australia using MARC-8 as the data exchange format. You can import Libraries Australia Unicode script records mapped to MARC-8 to your local system by using MARC-8 as the data exchange format. However, you cannot import records that have scripts which are not included in the MARC-8 repertoire, as the characters in the records imported to your local system would be unintelligible.

Appendix A: MARC formats

A. Bibliographical records

Model A

In this example the record is created in LACC by the National Library. The language of cataloguing is English (Latin script) and the language of the bibliographic item is Russian (Cyrillic script). The cataloguer transliterates the Cyrillic script data into the Latin script. Notes, subjects, etc. are in English.

```

040 ## $aANL$beng$cANL
100 0# $601$7ba$a(Name in Latin that is found in LCNA or
      created according to LCNA)
100 0# $601$7ca$a(Name in Cyrillic)
245 1# $602$7ba$a@(Title proper in Latin$cStatement of
      responsibility in Latin)
245 1# $602$7ca$a@(Title proper in Cyrillic$cStatement of
      responsibility in Latin)
250 ## $603$7ba$a(Edition in Latin)
250 ## $603$7ca$a(Edition in Cyrillic)
260 ## $604$7ba$a(Place of publication in Latin) :$b(Name
      of publisher in Latin),$c1992
260 ## $604$7ca$a(Place of publication in Cyrillic):[$bName
      of publisher in Cyrillic,$c1992
300 ## $a86 p. ;$c19 cm.
500 ## $aBibliography: p. 167-[168].
650 #0 $a(LCSH heading if applicable)
651 #0 $a(LCSH heading if applicable)
710 2# $605$7ba$a(Corporate name in Latin that is found in
      LCNA or created according to LCNA)
710 2# $605$7ca$a(Corporate name in Cyrillic)

```

Notes

1. Leader character position 9: Character encoding scheme defaults to “a” which indicates that the record uses Unicode for its character encoding. More information about Leader character position 9 can be found in: (<http://www.loc.gov/marc/bibliographic/ecbdldrd.html>)
2. Some of the regular fields (100, 245, 250, 260) are in pairs. One is in Latin script and the other in Cyrillic script. They are linked to each other by the linkage subfield \$6 plus the occurrence number such as 01, 02, 03, 04 and 05.
3. \$7 is the script subfield, “ba” and “ca” are script codes for Latin and Cyrillic respectively.

Model B

In this example the record is created by the City of Sydney Library in their local Unicode system. The language of cataloguing is English (Latin script) and the language of the bibliographic item is Russian (Cyrillic script). The

cataloguer transliterates the Cyrillic script data into the Latin script. Notes, subjects, etc. are in English.

```
000      01499cam a2200337 a 4500
040    _   $aN$PL$beng$cNSPL
100    0_   $6880-01$a(Name in Latin that is found in LCNA or
        created according to LCNA)
245    10   $6880-02$a(Title proper in Latin$cStatement of
        responsibility in Latin)
250    _   $6880-03$a(Edition in Latin)
260    _   $6880-04$a(Place of publication in Cyrillic):[$bName
        of publisher in Cyrillic,$c1992
300    _   $a86 p. ;$c19 cm.
500    _   $aBibliography: p. 167-[168].
650    _0   $ a(LCSH heading if applicable)
651    _0   $ a(LCSH heading if applicable)
710    2_   $6880-05$a(Corporate name in Latin that is found in
        LCNA or created according to LCNA)
880    0_   $6100-01$a(Name in Cyrillic)
880    10   $6245-02$a(Title proper in Cyrillic$cStatement of
        responsibility in Latin)
880    _   $6250-03$a(Edition in Cyrillic)
880    _   $6260-04$a(Place of publication in Cyrillic):[$bName
        of publisher in Cyrillic),$c1992
880    2_   $6710-05$a(Corporate name in Cyrillic)
```

Notes:

1. Leader character position 9 is coded as "a", which is the default value in most Unicode-based systems .
2. Some of the regular fields (100, 245, 250, 260, and 710) which contain standard Romanized data are linked to the 880 fields by the linkage subfield 6 plus the occurrence number.
3. The 880 linked fields which contain the Cyrillic script are linked back to the regular fields by the linkage subfield 6, plus the same occurrence number. Information about the 880 fields (Alternate graphic representation) can be found in the following link:
(<http://www.loc.gov/marc/bibliographic/ecbdhold.html#mrcb880>)

Model C

In this example the record is created by the University of Sydney in their local system, which supports non-Latin scripts in the MARC-8 repertoire. The language of cataloguing is English (Latin script) and the language of the bibliographic item is Russian (Cyrillic script). The cataloguer transliterates the Cyrillic script data into the Latin script. Notes, subjects, etc. are in English.

000 01499cam 2200337 a 4500
040 — \$aNU\$beng\$cNU
066 \$c(N
100 0_ \$6880-01\$a(Name in Latin that is found in LCNA or
created according to LCNA)
245 10 \$6880-02\$a(Title proper in Latin\$cStatement of
responsibility in Latin)
250 — \$6880-03\$a(Edition in Latin)
260 — \$6880-04\$ a(Place of publication in Cyrillic):[\$bName
of publisher in Cyrillic,\$c1992
300 — \$a86 p. ;\$c19 cm.
500 — \$aBibliography: p. 167-[168]..
650 _0 \$ a(LCSH heading if applicable)
651 _0 \$ a(LCSH heading if applicable)
710 2_ \$6880-05\$a(Corporate name in Latin that is found in
LCNA or created according to LCNA)
880 0_ \$6100-01/(N\$a(Name in Cyrillic)
880 10 \$6245-02/(N\$a(Title proper in Cyrillic\$cStatement of
responsibility in Latin)
880 — \$6250-03/(N\$ a(Edition in Cyrillic)
880 — \$6260-04/(N\$ a(Place of publication in
Cyrillic):[\$bName of publisher in Cyrillic),\$c1992
880 2_ \$6710-05/(N\$ a(Corporate name in Cyrillic)

Notes

1. Leader character position 9 is coded as blank to indicate the character encoding is not Unicode.
2. 066 contains the code of the character set present:

066 \$c(N for basic Cyrillic
066 \$c(Q for extended Cyrillic
066 \$c(S for Greek

For more information about 066 see:

<http://www.loc.gov/marc/bibliographic/ecbdclas.html#mrcb066>

For other language codes see:

<http://www.loc.gov/marc/specifications/speccharmac8.html>

3. Some of the regular fields (100, 245, 250, 260, and 710) which contain standard Romanized data are linked to the 880 fields by the linkage subfield 6 plus the occurrence number.
4. The 880 linked fields contain the Cyrillic linked back to the regular fields by the linkage subfield 6, plus the same occurrence number and the language code e.g. 6100-01/(N\$a(Name in Cyrillic)

B. Authority records

We recommend following the practice that is used for CJK authority records. The practice is established according to MARC21 format for authority data which is to record the preferred form in the 1XX fields, and the script form in the 4XX fields as non-preferred form e.g.

```
100 1 $a(Name that is found in LCNA or created according to LCNA)$d1961-  
400 1 $a(Name in original script)$d1961
```

Currently there are more than 1,400 CJK authority records in Libraries Australia with script data in the 4XX fields.

Appendix B: In which MARC fields should you add non-Latin scripts?

Even though Unicode-based systems enable users to enter non-Latin scripts anywhere in a MARC record, there are some common rules to be followed. They can be found in the following two useful documents:

- Program for Cooperative Cataloging Non-Roman Core Record Task Group Final Report (<http://www.loc.gov/catdir/pcc/archive/jackphy.html>)
- CONSER Editing Guide Appendix O. Creating Records with Data in Non-Roman Script for Chinese, Japanese, and Korean Serials (<http://www.itsmarc.com/crs/edit7839.htm>)

In Libraries Australia you should follow this guideline:

| Element | Latin script | Non-Latin scripts |
|-------------------------|--------------|-------------------|
| Fixed length data (0XX) | M | NA (1) |
| Main entry (1XX) | MA | MA (2) |
| Uniform title (240) | MA | O (3) |
| Description (245-300) | M | M (4) |
| Series statement (4XX) | MA | MA |
| Notes (5XX) | O | O (5) |
| Subject access (6XX) | MA | MA (6) |
| Added entries (7XX) | MA | MA (2) |

Key

- M* *Mandatory*
MA *Mandatory if applicable*
O *Optional*
NA *Not applicable*

Notes

1. Fixed length data

If a non-Latin ISBN is to be entered it must be entered in subfield z.

2. Main entry and added entries

Name entries may have non-Latin script entries if the Latin entry uses standard ALA/LC Romanization of the script heading. Script headings are required if the Latin entry represents standard ALA/LC systematic Romanization.

We believe if non-Latin script data are included in authority records, there should be no need to supply parallel headings for access points in bibliographic records. We encourage cataloguers to end this redundant practice.

Do not supply parallel non-Latin script fields for conventional headings that do not conform to standard ALA/LC systematic Romanization.

Give non-Latin script fields in the form found in the publication if it equates to the established form.

For personal names given in a parallel non-Latin script field, do not follow a surname with a comma.

3. Uniform title

LACC cataloguers may provide parallel fields containing non-Latin script data for variant titles, if the variant title provides a useful access point.

4. Description

In the non-Latin script title (245) field, all information must be transcribed, including title proper, parallel titles, and statement of responsibility as found on the chief source.

5. Notes

Non-Latin script general note (500) fields should not be constructed by the cataloguer. Non-Latin script data may be represented in original script. The general note (500) field may be present and entered either in Romanization only or in the non-Latin script only or in parallel fields. Content (505) fields may be present and may be entered either in Romanization only or in the non-Latin script only or in parallel fields.

6. Subject access

Do not supply non-Latin script data for topical subject headings (650) even though there may be a one-to-one equivalent between non-Latin script characters and the established form.