Catalog using Tamil script

Overview

Use Tamil script data for cataloging items in the Tamil language. Use Tamil script data the same way you use other non-Latin script data in the client.

See Work with international records and Guidelines for contributing non-Latin script bibliographic records to WorldCat for details specific to non-Latin scripts. See also general procedures describing how to:

- Search WorldCat
- Create records
- Edit records
- Use constant data
- Save records and search save files
- Report errors in records
- Export records
- Import records

Tools for using non-Latin scripts

- **Link/unlink** (Edit > Linking Fields > Link [or Unlink]) - Visually link or unlink non-Latin script data fields with equivalent Latin script (romanized) data fields (bibliographic records only)
- **Export options for data fields** (Tools > Options > International) - Determine (for bibliographic records only):
  - Whether to export both equivalent Latin script (romanized) data and non-Latin script data or only one or the other
  - Position of data if you export both Latin and non-Latin script data
  - Sort order

Caution: MARC-8 character verification (Edit > MARC-8 Characters > Verify) is not appropriate for verifying Tamil characters. There is no MARC-8 character set for Tamil. Using this command for Tamil results in marking all Tamil characters as invalid. The OCLC system validates Tamil characters when you validate a record.

UTF-8 Unicode export and import required for Tamil records

Because Tamil script is not included in MARC-8 character sets, you must export and import records in the UTF-8 Unicode character set (settings for export are in Tools > Options > Export, click Record Characteristics, and settings for
import are in File > Import Records, click Record Characteristics). If you export or import using the MARC-8 character set, non-MARC-8 characters are retained in Numeric Character Reference (NCR) notation only.

About Unicode

Unicode is the universal character encoding scheme for written characters and text. It defines a consistent way of encoding multi-script text that enables the exchange of text data internationally.

Unicode provides for three encoding forms: a 32-bit form (UTF-32), a 16-bit form (UTF-16), and an 8-bit form (UTF-8, designed for use with ASCII-based systems).

Connexion client began supporting Tamil script with Unicode version 4.0.0.

Tamil script entry and character set

Script entry method

If your system default language is not Tamil, you can install the Tamil language in Windows. When you install Tamil, Windows provides an input keyboard for entering Tamil script. See more about input methods for languages that use non-Latin scripts.

Character set supported

Tamil characters are defined in Unicode 4.0 (coded in the range U+0B80 to U+0BFF).

Unicode 4.0 coding for Tamil characters does not include most glyphs that are formed when consonants alone (implicit "a" suppressed) are combined with independent vowels.

Caution: Microsoft Arial Unicode MS font, which is recommended for general use in the client, does not support the Tamil digits, numerics, and symbols that are coded in Unicode 4.0.

Script identifier in records

The client adds the following data to ‡c of field 066 in Tamil records to indicate the presence of Tamil characters:

- Tamil

Romanized data

See the ALA-LC Romanization Table for Tamil on the Library of Congress website.

Indexing for Tamil script searches
Notes on searching

• Use word or phrase search indexes and browse indexes.
• Word searches find the data string you enter anywhere in the indexed field. Phrase searches find the data string starting with the first character in a field or subfield and including each character in exact order. Browsing scans an index for the closest match to the character string followed by any other data.
• If you use qualifiers to limit a search, type them in Latin script.
• Do not use derived searching.
• Do not truncate searches (asterisk (*) at the end of a search term). You can use browsing for automatic truncation (enter only as many characters as needed for a match without using an asterisk).
• If you want to retrieve all Tamil script records or see sample records, use the "character sets present" WorldCat search index (label vp:) with the assigned code tam.
  ◦ To find all Tamil script records, enter \texttt{vp:tam} as a command line search in the Search WorldCat window (Cataloging > Search > WorldCat).
  Note: If a search for all Tamil script records alone retrieves too many WorldCat records (limit 1,500 records), you must limit the search and try again (e.g., \texttt{vp:tam/1991-} (qualified by years of publication); \texttt{vp:tam and mt:bks} (limited to records in the Continuing resources format); etc.).

See general procedures and search techniques for searching WorldCat.

Tamil character indexing specifics

Tamil independent vowels are characters that stand on their own. Each has a unique Unicode code value.

Tamil consonants contain an implicit “a” vowel sound. A modifier called Virama (or Pulli) added to the consonant glyph represents the consonant alone with no vowel. If the consonant alone is combined with an independent vowel (not implicit “a”), the vowel becomes dependent and the visual form changes.

The OCLC system indexes each Tamil consonant separately in all three of the following forms based on text:

• Consonant with the implicit "a"
• Consonant alone (indicated by Virama or Pulli)
• Consonant combined with any independent vowel (indicated by the addition of a vowel ligature)

Example

• க [ka] = Tamil consonant KA with implicit a.
• க [k] = consonant KA alone (with Virama or Pulli that looks like a diacritical mark).
• க [ki] = consonant KA combined with Virama and independent vowel ப [i]. The glyph for KA has the added ligature for ப (where the dotted circle represents the consonant to which the ligature is attached). Unicode does not provide a unique code for the transformed glyph. For example, it does not code கப. It codes only the dependent vowel ligature (e.g., கப).
Based on Unicode Tamil character-rendering rules, the OCLC system stores the Unicode values in a sequence that represents a new glyph used in Tamil text writing and indexes accordingly. For the example here, the system indexes the glyphs with the following Unicode values:

- க [ka] as 0B95
- க [k] as 0B95 + -BCD, where Virama (0BDC) is indexed
- க [ki] as 0B95 + 0BBF

Note: Tamil Unicode 4.0 codes are not in collating order. The default sort order for search results, alphabetical sorting by Latin script, is recommended if romanized (Latin-equivalent) data is included in the record. The sort order option is in Tools > Options > International or <Alt><T><O>).