Valid character sets for supported scripts

Discover the valid character sets for supported non-Latin scripts in Connexion client.

**Arabic, CJK, Cyrillic, Greek, and Hebrew**

Character sets for these scripts are listed in [MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media, Code Tables](https://www.loc.gov/marc/specifications/). These MARC-8 character sets are subsets of Unicode characters that are approved for use in MARC 21 cataloging.

Scripts defined by MARC-8 character sets are supported for bibliographic records and for variant name headings in authority records.

The following list defines the scope of valid characters in the Connexion client for Arabic (including Persian), CJK, Cyrillic, Greek, and Hebrew scripts:

- Basic Arabic = 33(hex) [ASCII graphic: 3]
- Extended Arabic = 34(hex) [ASCII graphic: 4]
- Chinese, Japanese, Korean (EACC) = 31(hex) [ASCII graphic: 1]
- Basic Cyrillic = 4E(hex) [ASCII graphic: N]
- Extended Cyrillic = 51(hex) [ASCII graphic: Q]
- Basic Greek = 53(hex) [ASCII graphic: S]
- Basic Hebrew = 32(hex) [ASCII graphic: 2]
  
  Note: In bibliographic records, the client inserts the notation (3, (4, $1, (N, (Q, (S, or (2, respectively, into field 066 to indicate which script(s) are used in a record. If multiple scripts are used, the notations are inserted individually, each in a separate subfield c.

**Armenian, Bengali, Devanagari, Ethiopic, Syriac, Tamil, and Thai**

These scripts are supported for bibliographic records only.

There are no defined MARC-8 character sets for Armenian, Bengali, Devanagari, Ethiopic, Syriac, Tamil or Thai. In addition, Connexion Client also supports Cyrillic characters outside the MARC-8 character set. OCLC implemented the following script identification codes for these scripts based on [ISO 15924 Code Lists](https://www.loc.gov/marc/codes/).

The following list shows the ranges of UTF-8 Unicode characters that define valid characters for these scripts in the Connexion client:

- Armn = Armenian (character range U+0530 to U+058F)
• Beng = Bengali (character range U+0980 to U+09FF)
• Cyrl = Cyrillic character set (outside the MARC-8 character set)
• Deva = Devanagari (character range U+0900 to U+097F)
• Ethi = Ethiopic (character range U+1200 to U+1399, U+2D80 to U+2DDF, U+AB00 to U+AB2F)
• Syrc = Syriac (character range U+0700 to U+074F)
• Tamil = Tamil (character range U+0B80 to U+0BFF)
• Thai = Thai (character range U+0E00 to U+0E7F)

Note: The client inserts Armn, Beng, Cyrl, Deva, Ethi, Syrc, Tamil, or Thai, respectively, in field 066 of a bibliographic record to indicate that the script is used. If multiple scripts are used, the notations are inserted individually, each in a separate subfield c.

**Limitations on using Armenian, Bengali, Cyrillic (outside the MARC-8 character set), Devanagari, Ethiopic, Syriac, Tamil, and Thai scripts**

• To export or import records containing Armenian, Bengali, Cyrillic (outside the MARC-8 character set), Devanagari, Ethiopic, Syriac, Tamil, and Thai scripts, you must select the UTF-8 Unicode character set option in:
  ◦ To export: Tools > Options > Export; click Record Characteristics and select UTF-8 Unicode in the Character Set list under Bibliographic Records.
  ◦ To import: File > Import Records; click Record Characteristics and select UTF-8 Unicode in the Character Set list under Bibliographic Records.

Because Armenian, Bengali, Cyrillic (outside the MARC-8 character set), Devanagari, Ethiopic, Syriac, Tamil, and Thai scripts are not part of MARC-8 characters, you cannot export or import these scripts using the MARC-8 character set option.

Because MARC-8 characters are part of UTF-8 Unicode, you can safely export or import Arabic, CJK, Cyrillic, Cyrillic (outside of the MARC-8 character set), Greek, and Hebrew records using either the MARC-8 or the UTF-8 Unicode character set option.

• Armenian, Cyrillic (outside the MARC-8 character set), Bengali, Devanagari, Ethiopic, Syriac, Tamil, and Thai scripts are not supported for variant name headings in authority records.

**Invalid characters in Connexion client**

Any characters that are not included in the above lists of defined characters or that cannot be inserted via Edit > Enter Diacritics (or <Ctrl><Shift><I> or <Ctrl><E>) are invalid in the client. To include non-Latin characters that you need but that are invalid in Connexion client, you can:

• Enter the character in the record, export the record to your local system using Unicode export format, and then remove the character before processing the record in WorldCat.

  Or

• Enter the name of the character within square brackets, using the Unicode standard if available, (e.g., enter [schwa]), or for CJK characters, enter the reading of the character (e.g., enter [yin]).
For reference, see the Unicode charts, which has a character name index.

Note: Z39.50 access to WorldCat records also supports MARC-8 and Unicode UTF-8 character sets. See Z39.50 Cataloging for information on non-Latin script support in Z39.50.

**Multiscripts in a single record are valid**

Use as many supported non-Latin scripts as you need anywhere in a record, including within the same field.