



POD Translation
by *pod2pdf*

ajf@afco.demon.co.uk

Charset.pm

Table of Contents

Charset.pm

NAME	1
SYNOPSIS	1
DESCRIPTION	1
Latin (Basic/Extended + Greek Symbols, Subscripts and Superscripts)	1
Hebrew	1
Cyrillic (Basic + Extended)	1
Arabic (Basic + Extended)	1
Greek	1
East Asian Characters	1
METHODS	1
new()	1
to_utf8()	1
to_marc8()	2
g0()	2
g1()	2
TODO	2
to_marc8()	2
SEE ALSO	2
MARC::Charset::Generic	2
MARC::Charset::ASCII	2
MARC::Charset::Ansel	2
MARC::Charset::ArabicBasic	2
MARC::Charset::ArabicExtended	2
MARC::Charset::Controls	2
MARC::Charset::CyrillicBasic	2
MARC::Charset::CyrillicExtended	2
MARC::Charset::EastAsian	2
MARC::Charset::Greek	2
MARC::Charset::GreekSymbols	2
MARC::Charset::Hebrew	2
MARC::Charset::Subscripts	2
MARC::Charset::Superscripts	2
AUTHORS	2
Ed Summers <ehs@pobox.com>	2

NAME

MARC::Charset - A module for doing MARC-8/UTF8 translation

SYNOPSIS

```

use MARC::Charset;

## create a MARC::Charset object
my $charset = MARC::Charset->new();

## a string containing the Ansel value for a copyright symbol
my $ansel = chr(0xC3) . ' copyright 1969'.

## the same string, but now encoded in UTF8!
my $utf8 = $charset->to_utf8($extLatin);

```

DESCRIPTION

MARC::Charset is a package that allows you to easily convert between the MARC-8 character encodings and Unicode (UTF-8). The Library of Congress maintains some essential mapping tables and information about the MARC-8 and Unicode environments at:

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

MARC::Charset is essentially a Perl implementation of the specifications found at LC, and supports the following character sets:

- Latin (Basic/Extended + Greek Symbols, Subscripts and Superscripts)
- Hebrew
- Cyrillic (Basic + Extended)
- Arabic (Basic + Extended)
- Greek
- East Asian Characters
 - Includes 13,478 "han" characters, Japanese Hiragana and Katakana (172 characters), Korean Hangul (2,028 characters), East Asian Punctuation Marks (25 characters), "Component Input Method" Characters (35 characters)

METHODS

new()

The constructor which will return MARC::Charset object. If you like you can pass in the default G0 and G1 charsets (using the g0 and g1 parameters, but if you don't ASCII/Ansel will be assumed.

```

## for standard characters sets: ASCII and Ansel
my $cs = MARC::Charset->new();

## or if you want to specify Arabic Basic + Extended as the G0/G1 character
## sets.
my $cs = MARC::Charset->new(
    g0 => MARC::Charset::ArabicBasic->new(),
    g1 => MARC::Charset::ArabicExtended->new()
);

```

If you would like diagnostics turned on pass in the DIAGNOSTICS parameter and set it to a value that will evaluate to true (eg. 1).

```

my $cs = MARC::Charset->new( diagnostics => 1 );

```

to_utf8()

Pass to_utf8() a string of MARC8 encoded characters and get back a string of UTF8 characters. to_utf8() will handle escape sequences within the string that change the working character sets to Greek, Hebrew, Arabic (Basic + Extended), Cyrillic (Basic + Extended), and East Asian.

to_marc8()

When you pass this method a UTF8 string you will be returned a MARC8 encoded string. to_marc8() handles creating the appropriate character escapes.

g0()

Returns an object representing the character set that is being used as the first graphic character set (G0). If you pass in a MARC::Charset:* object you will set the G0 character set, and as a side effect you'll get the previous G0 value returned to you. You probably don't ever need to call this since character set changes are handled when you call to_utf8(), but it's here if you want it.

```
## set the G0 character set to Greek
my $charset = MARC::Charset->new();
$charset->g0( MARC::Charset::Greek->new() );
```

g1()

Same as g0() above, but operates on the second graphic set that is available.

TODO

- to_marc8()
A function for going from Unicode to MARC-8 character encodings.

SEE ALSO

[MARC::Charset::Generic](#)
[MARC::Charset::ASCII](#)
[MARC::Charset::Ansel](#)
[MARC::Charset::ArabicBasic](#)
[MARC::Charset::ArabicExtended](#)
[MARC::Charset::Controls](#)
[MARC::Charset::CyrillicBasic](#)
[MARC::Charset::CyrillicExtended](#)
[MARC::Charset::EastAsian](#)
[MARC::Charset::Greek](#)
[MARC::Charset::GreekSymbols](#)
[MARC::Charset::Hebrew](#)
[MARC::Charset::Subscripts](#)
[MARC::Charset::Superscripts](#)

AUTHORS

Ed Summers <ehs@pobox.com>