UNICODE

By Ron Kaplan

This document was last edited in July 2023.

The UNICODE library package defines external file formats that enable Medley to read and write files where 16 bit character codes are represented as UTF-8 byte sequences or big-endian UTF16 byte-pairs. It also provides for character codes to be converted (on reading) from Unicode codes to equivalent codes in the Medley-internal Xerox Character Code Standard (XCCS) and (on writing) from XCCS codes to equivalent Unicode codes.

UTF-8 External formats

Four external formats are defined when the package is loaded:

:UTF-8	codes are represented as UTF-8 byte sequences and XCCS/Unicode character conversion takes place.
:UTF-16BE	codes are represented as 2-byte pairs, with the high order byte appearing first in the file, and characters are converted.

The two other external formats translate byte sequences into codes, but do not translate the codes. These allow Medley to see and process characters in their native encoding.

:UTF-8-RAW	codes are represented as UTF-8 byte sequences, but character conversion does not take place.
:UTF-16BE-RAW	codes are represented as big-ending 2-byte pairs but there is no conversion.

These formats all define the end-of-line convention (mostly for writing) for the external files according to the variable EXTERNALEOL (LF, CR, CRLF), with LF the default.

The external format can be specified as a parameter when a stream is opened:

```
(OPENSTREAM 'foo.txt 'INPUT 'OLD '((EXTERNALFORMAT :UTF-8)))

(CL:OPEN 'foo.txt :DIRECTION :INPUT :EXTERNAL-FORMAT :UTF-8)
```

The function STREAMPROP obtains or changes the external format of an open stream:

```
(STREAMPROP stream 'EXTERNALFORMAT) -> :XCCS

(STREAMPROP stream 'EXTERNALFORMAT :UTF-8) -> :XCCS
```

In the latter case, the stream's format is changed to :UTF-8 and the previous value is returned, in this example it is Medley's historical default format :XCCS.

Entries can be placed on the variable *DEFAULT-EXTERNALFORMATS* to change the external format that is set by default when a file is opened on a particular device. Loading UNICODE executes

```
(PUSH *DEFAULT-EXTERNALFORMATS* '(UNIX :UTF-8))
```

so that all files opened (by OPENSTREAM, CL:OPEN, etc.) on the UNIX file device will be initialized with :UTF-8. Note that the UNIX and DSK file devices reference the same files (although some caution is needed because {UNIX} does not simulate Medley versioning), but the device name in a file name ({UNIX}/Users/... vs. {DSK}/Users/...) selects the particular device. The default setting above applies only to files specified with {UNIX}; a separate default entry for DSK must be established to change its default from :XCCS.

The user can also specify the external format on a per-stream basis by putting a function on the list STREAM-AFTER-OPEN-FNS. After OPENSTREAM opens a stream and just before it is returned to the calling function, the functions on that list are applied in order to arguments STREAM, ACCESS, PARAMETERS. They can examine and/or change the properties of the stream, in particular, by calling STREAMPROP to change the external format from its device default.

Mapping between Unicode and XCCS character codes

The XCCS/Unicode mapping tables are defined by the code-mapping files for particular XCCS character sets. These are typically located in the Library> sister directory

```
..>Unicode>Xerox>
```

and the variable UNICODEDIRECTORIES is initialized with a globally valid reference to that path. The global reference is constructed by prepending the value of the Unix environment-variable "MEDLEYDIR" to the suffix >Unicode>Xerox>.

The mapping files have conventional names of the form XCCS-[charsetnum]=[charsetname].TXT, for example, XCCS-0=LATIN.TXT, XCCS-357=RSYMBOLS4.TXT. The translations used by the external formats are read from these files by the function

```
(READ-UNICODE-MAPPING FILESPEC NOPRINT NOERROR)                    [Function]
```

where FILESPEC can be a list of files, charset octal strings ("0" "357"), or XCCS charset names (LATIN EXTENDED-LATIN). Reading will be silent if NOPRINT, and the process will not abort if an error occurs and NOERROR. The value is a flat list of the mappings for all the character sets, with elements of the form (XCCC-code Unicode-code).

When UNICODE is loaded the mappings for the character sets specified in the variable DEFAULT-XCCS-CHARSETS are installed. This is initialized to

```
(LATIN SYMBOLS1 SYMBOLS2 EXTENDED-LATIN FORMS SYMBOLS3 SYMBOLS4 ACCENTED-
  LATIN GREEK)
```

but DEFAULT-XCCS-CHARSETS can be set to a different collection before UNICODE is loaded.

The internal translation tables used by the external formats are constructed from a list of correspondence pairs by the function

```
(MAKE-UNICODE-TRANSLATION-TABLES MAPPING [FROM-XCCS-VAR]
  [TO-XCCS-VAR])                                                   [Function]
```

This returns a list of two arrays (XCCS-to-Unicode Unicode-to-XCCS)containing the relevant translation information organized for rapid access. If the optional from/to-variables arguments are provide, they are the names of variables whose top-level values will be set to these arrays, for convenience. For the external formats defined above, these variables are *XCCSTOUNICODE* and *UNICODETOXCCS*.

The macro

(UNICODE.TRANSLATE CODE TRANSLATION-TABLE) [Macro]

is used by the external formats to perform the mappings described by the translation-tables.

The following utilities are provided for lower-level manipulation of codes and strings.

(XTOUCODE XCCSCODE) -> corresponding Unicode

(UTOXCODE UNICODE) -> corresponding XCCS code

(NUTF8CODEBYTES N) -> number of bytes in the UTF-8 representation of N.

(NUTF8STRINGBYTES STRING RAWFLG) -> number of bytes in the UTF-8 representation of STRING, translating XCCS to Unicode unless RAWFLG.

(XTOUSTRING XCCSSTRING RAWFLG) -> The string of bytes in the UTF-8 representation of the characters in XCCSSTRING (= the bytes in its UTF-8 file encoding).

(HEXSTRING N WIDTH) -> the hex string for N, padded to WIDTH

The UNICODE file also contains a function for writing a mapping file given a list of mapping pairs. The function

(WRITE-TRANSLATION-TABLE MAPPING [INCLUDEDCHARSETS] [FILE])

produces one or more mapping files for the mapping-pairs in mapping. If the optional FILE argument is provided, then a single file with that name will be produced and contain all the mappings for all the character sets in MAPPING. If FILE and INCLUDEDCHARSETS are not provided, then all of the mappings will again go to a single file with a composite name XCCS-csn1,csn2,csn3.TXT. Each cs may be a single charset number, or a range of adjacent charset numbers. For example, if the mappings contain entries for characters in charset LATIN, SYMBOLS1, SYMBOLS2, and SYMBOLS3, the file name will be XCCS-0,41-43.TXT.

If INCLUDEDCHARSETS is provided, it specifies possibly a subset of the mappings in MAPPING for which files should be produced. This provides an implicit subsetting capability.

Finally, if FILE is not provided and INCLUDEDCHARSETS is T, then a separate file will be produced for each of the character sets, essentially a way of splitting a collection of character-set mappings into separate canonically named files (e.g. XCCS-357=SYMBOLS4.TXT).