EncTEX

The Extension of TEX For Input Re-encoding

6. 9. 1997, 3. 1. 2003, 12. 1. 2003

Petr Olšák

The original version of the encTEX documentation (in Czech language) is in `encdoc.tex` file.

Původní česká dokumentace je v souboru `encdoc.tex`.

1. The basic information

The encTEX package is a little extension of TEX. You can install it from source files of TEX by changing the `tex.ch` file in your distribution. The patch to `tex.ch` file for `web2c` distribution is supported.

The encTEX is backward compatible with the original TEX. It adds eight new primitives by which you can set or read the conversion tables used by input processor of TEX or used during output to the terminal, log and `\write` files. These tables are stored to the format files thus, they are reinitialized to the same state as in time of `\dump` command when the format file is read.

This extension is fully tested and it passes the TRIP test with only two differences:

- The banner is different
- The number of "multiletter control sequences" is greater by eight.

1.1. The installation

For install instructions of encTEX – read the `INSTALL.eng` file.

1.2. Versions

I released the first version of encTEX in 1997. This version was able to do the byte to byte conversion only by xord and xchr vector and to assign the characters as "printable" (the `\xordcode`, `\xchrcode` and `\xprncode` primitives).

I incorporated the possibility to multi-byte to one byte or control sequence conversion in December 2002 and January 2003. This version is called "Jan. 2003" and it adds five new primitives: `\mubyte`, `\endmubyte`, `\mubytein`, `\mubyteout` and `\specialout`. They give a possibility to set the conversion from UTF-8 encoded files.

1.3. The conflict with TCX tables

The both encTEX and TCX tables manipulates with the same data (xord and xchr vectors). It means that the conflict may be occur. This was a reason why I took back the encTEX after the TCX tables renovation was released in 1998. But TCX are not able to convert from UTF-8 so I upgraded my encTEX in 2002 and I am propagate it again.

Attention: If you are using encTEX then the TCX tables are ignored even if the `*.tcx` file is given in `-translate-file` option. The encTEX is implemented only in `tex.ch` file but TCX are in additional C sources. We are working on possibility of cooperating of the encTEX and TCX tables.

1.4. The TEX license

The encTEX adds the new primitives into TEX so, we cannot call the resulting program by name TEX. On the other hand D. Knuth assumed that TEX internals are filtered always from system dependences. This was a reason why he implemented xord/xchr vectors in TEX. D. Knuth assumed that the parameters of filter from system dependences is set at source code level. EncTEX only moves this setting from source code level to the runtime level. This is nothing new: the TEX memory parameters are possible to set at runtime in modern TEX distributions too. You can set the conversion tables depend on your system. Then you can say `\let\xordcode=\undefined` etc. (the same for other encTEX primitives) and you can do `\dump`. The format has the conversion tables stored by the system specifications and the user cannot do any more changes. The using of this format acts the same as the using the original TEX.

I think that the second line on the terminal and log file is sufficient information about the fact that the program is a modified version of TEX. I think that if the UTF-8 encoding will be used more common then there is no another way than to modify the input processor of TEX otherwise the 8bit TEX will dead in short time.

It is important to say that encTEX has the same default behavior as the original TEX if the new primitives are never used.

IMHO, the new web2c TEX is not exactly the TEX too because you can change its behavior by writing `%&` at the first line of the document. This feature is not documented in *Computers & Typesetting* series.

2. The byte per byte conversion

2.1. The xord and xchr vectors

All text inputs into TEX are mapped by xord vector in input preprocessor (the eyes in TEXbook terminology). If the character has the code x in your system, the same character has the code $y = \text{xord}[x]$ in TEX.

All text outputs from TEX to terminal, log file and files managed by `\write` primitive are filtered by xchr vector and by "printability" feature of the character. If the character with code y is not "printable", then it outputs by `^^code` notation (documented in TEXbook, page 45). If the character with code y is "printable" then the output code of this character on terminal and text files is $z = \text{xchr}[y]$.

2.2. The new primitives with the access to the xord and xchr vectors

The encTEX extension introduces three new primitives with the same syntax as `\lccode`:

- `\xordcode` i ... is xord$[i]$
- `\xchrcode` i ... is xchr$[i]$
- the character with the code i is "printable" (not `^notation` on terminal and the log is used) iff (`\xprncode` $i > 0$) or ($i \in \{32, ..., 126\}$).

All setting to `\xordcode`, `\xchrcode` and `\xprncode` are possible in 0...255 range and are *global* every time. It means that the setting inside group are global and it is irrelevant if you write `\global` prefix or you do not.

The initial values at iniTEX state of the mentioned vectors are:

- `\xordcode` $i = i$ for $i \in \{128...255\}$,
- `\xchrcode` $i = i$ for $i \in \{128...255\}$,
- `\xprncode` $i = 0$ for $i \in \{0...31, 127...255\}$,
- `\xprncode` $i = 1$ for $i \in \{32...126\}$.

The `\xordcode` i and `\xchrcode` i for $i \in \{0...127\}$ are system dependent, but on systems with ASCII encoding holds: `\xordcode` $i = i$, `\xchrcode` $i = i$.

3. The multi-byte conversion

Since version Dec 2002, the encTEX is able to convert more bytes to one byte or control sequence on input processor level. This "one byte" is converted back to the original "more bytes" when `\write` is processed or TEX outputs to the terminal or log file. The main reason of this extension of TEX is to serve to work with the UTF-8 encoded input files: we need to assign the `\catcode`s, `\uccode`s etc. to the letters in our alphabet but some letters are encoded in two bytes in UTF-8. The encTEX is able to map other codes from UTF-8 to control sequences thus, the number of UTF-8 codes from input file examined by TEX is unlimited.

There are four new primitives to manage the conversion: `\mubytein`, `\mubyteout`, `\mubyte`, `\endmubyte`. The `\mubytein` and `\mubytout` are integer registers with zero value by default: it means that no conversion is processed even if the conversion table (created by `\mubyte`, `\endmubyte`) is non empty. If `\mubytein` is positive then the conversion on input processor level is performed by the conversion table. If `\mubyteout` is positive then the conversion for output to the `\write` files, the log file and the terminal is activated by the same conversion table. The conversion table is empty by default and you can add the new line into this table by the couple of `\mubyte`, `\endmubyte` primitives:

`\mubyte` ⟨*first_token*⟩⟨*one_optional_space*⟩⟨*optional_prefix*⟩⟨*byte_sequence*⟩`\endmubyte`

Each ⟨*byte_sequence*⟩ will be converted to the ⟨*first_token*⟩ at input processor level. There are two possibilities for ⟨*first_token*⟩: it may be a character or a control sequence. If the ⟨*first_token*⟩ is a character then the catcode of it is ignored and the ⟨*first_token*⟩ is interpreted as a ⟨*byte*⟩. This ⟨*byte*⟩ is converted back to the ⟨*byte_sequence*⟩ in `\write` files, log file and terminal.

If the ⟨*first_token*⟩ is a control sequence then the ⟨*byte_sequence*⟩ is converted to this control sequence at input processor level to the "one token" form. It means that the token processor never changes this

control sequence. The token processor stays in middle line state after this control sequence is scanned. If `\mubyteout<2` then the output is not converted back to the ⟨*byte_sequence*⟩ and the control sequence is expanded as usual. If `\mubyteout>=2` then the control sequence declared by `\mubyte` is converted back to the ⟨*byte_sequence*⟩ in `\write` parameters. This works only if the control sequence is not expanded. It means that the control sequence have to be non expandable or it have to be marked by `\noexpand`. If `\mubyteout>=3` then encTEX suppresses the expansion of control sequences declared by `\mubyte` automatically. See section 3.7 for more details.

The syntax and meaning of ⟨*optional_prefix*⟩ will be explained in section 3.4.

3.1. The conversion table manipulation

The data are stored into conversion table as a global assignment. On the other hand the assignment to `\mubytein` and `\mubyteout` registers are local as usually.

The `\mubyte`, `\endmubyte` primitives work very similar as a well known `\csname`, `\endcsname` pair. The difference is that the ⟨*first_token*⟩ is not expanded and that this token can be followed by ⟨*one_optional_space*⟩ (after expansion). The ⟨*byte_sequence*⟩ is scanned with the full expansion. If the other non expandable control sequence than `\endmubyte` occurs during this process then the error message is printed:

```
! Missing \endmubyte inserted.
\begtt

The "\mubyte" is not performed on the expand processor level: it is a
assign primitive. If you write

\begtt
\edef\a{\mubyte X ABC\endmubyte}
```

then the macro `\a` includes the `\mubyte X ABC\endmubyte` tokens.

Examples:

```
\mubyte  ^^c1      ^^c3^^81\endmubyte % \'A
\mubyte  ^^e1      ^^c3^^a1\endmubyte % \'a
% etc. -- the UTF8 implementation

\mubyte  \endash   ^^c4^^f6\endmubyte % the mapping to the control sequence
\mubyte  \integral  INT\endmubyte     % the illustrative example, see below

\mubytein=1  \mubyteout=1  % conversions are activated here

\def\endash {--} % this is good definition for \write files too
\def\integral {\ifmmode \int\else $\int$\fi}
```

We have written more spaces (or tabs) in ⟨*one optional space*⟩ in this example because these characters have the catcode of the space and the token processor converts them to right ⟨*one optional space*⟩.

The word "`INTEGRAL`" is converted to the token `\integral` followed by the letters "`EGRAL`" if the example code is used. The text "`INT something`" is converted to the token `\integral` followed by space and the word "`something`". You can write the following constructions: `\defINT{something}`, `\let INT=\foo`, etc. If the `\integral` is undefined control sequence then the error message is printed if you write the "`INT`". The error message has a peculiar form:

```
! Undefined control sequence.
l.13 this is a INT
                  EGRAL.
```

You can type `\show INT` with the following answer:

```
> \integral=undefined.
l.13 \show INT
```

and `\string INT` expands to the text: `\integral`.

Assume the `INT` declaration from the previous example and assume that you write `\INT`. What happens? Strictly speaking, the empty control sequence (`\csname\endcsname`) followed by `\integral` control sequence would be the output from the token processor. But there is an exception in encTEX because to avoid the confusion with the empty control sequences. The `\INT` produces only the control sequence `\integral`, the backslash is ignored in this situation. The token processor stays in middle state after `\INT` is scanned, the letter can follow immediately.

3.2. The features of the conversion process

The input is converted immediately after `\mubytein` is set to the positive value; it means the conversion may start at the same line where the `\mubytein` setting occurs.

The ⟨*byte_sequence*⟩ is converted only if the whole ⟨*byte_sequence*⟩ is included in the one line. The `\endlinechar` character can be the last part of the ⟨*byte_sequence*⟩.

The sequence `^^c3^^81` is not converted to the letter Á even if the code from the example was used. The reason is that the `^^` conversion is done in token processor after the `\mubyte` conversion.

The `\xordcode` conversion is performed before `\mubyte` conversion in input side and the `\xchrcode` conversion is done after `\mubyte` conversion during output to the files or to the terminal. The following diagram shows the sequence of the conversions:

```
input text -> \xordcode -> \endlinechar appended ->
              \mubyte -> token processor -> expand processor ...
\write argument -> expand processor -> \mubyte -> \xchrcode -> output
```

The converted ⟨*byte_sequence*⟩ is not converted to the `^^` form during output to the file even if the `\xprncode` of the bytes from ⟨*byte_sequence*⟩ is zero. The ⟨*byte_sequence*⟩ is not converted again even if there exist a character in it which is normally converted by another rule in conversion table.

There exists an exception from output conversion process to the log file and to the terminal. If the complete line from input is reprinted to the log or to the terminal and if `\mubytein` is positive then there is no conversion of such line on input side nor back on output side. This line is reprinted byte per byte in the same form. Only `\xchrcode` followed by `\xordcode` conversion is active. The conversion to `^^aa` form is deactivated in this situation even if `\xprncode` is zero. This exception concerns to the error messages where the place of the problem is shown on the terminal and log file by printing the actual line and splitting it to two parts. This exception does not concern to any `\write` or `\message` output. Arguments of these commands are always expanded, translated by `\mubyte` conversion table and translated by `\xchrcode`.

3.3. Controversial records in conversion table

Let exist two or more ⟨*byte_sequences*⟩ in the conversion table which are equal or which have the same begin part and one sequence is a subsequence of the second. Then the precedence has the last occurrence of such ⟨*byte_sequences*⟩ in the conversion table and all others are ignored. Example:

```
\mubyte X ABC\endmubyte
{\mubytein=1 the ABC is converted to X}
\mubyte Y ABCDE\endmubyte
\mubyte W ABFG\endmubyte
{\mubytein=1 now, the ABC stay unchanged and ABCDE is converted to Y
             and ABFG is converted to W}
\mubyte Z AB\endmubyte
{\mubytein=1 now, the ABCDE is converted to ZCDE}
```

This convention servers the possibility of removing all lines from conversion table with the same first byte in ⟨*byte_sequence*⟩. It is sufficient to do it if you write the next line to conversion table with the one byte length of ⟨*byte_sequence*⟩ and with the same ⟨*first_token*⟩ as this ⟨*byte_sequence*⟩. For example:

```
\mubyte A A\endmubyte
```

removes all lines from conversion table with the ⟨*byte_sequence*⟩ beginning by the letter A. If ⟨*first_token*⟩ is equal to ⟨*byte_sequence*⟩ then `\mubyte` primitive really clears the lines with ⟨*byte_sequence*⟩ beginning by ⟨*first_token*⟩ from the conversion table and frees the main memory of TEX where these data are stored. In all other cases, `\mubyte` primitive only adds the new line into conversion table.

The following code clears the whole conversion table:

```
{\catcode`\^^@=12
\gdef\clearmubytes{\bgroup \count255=1
   \loop \uccode`X=\count255
       \uppercase{\mubyte XX\endmubyte}%
       \advance\count255 by1
       \ifnum\count255<256 \repeat
   \mubyte ^^@^^@\endmubyte
   \egroup}
}
\clearmubytes
```

3.4. Input and output sides of the conversion table

The conversion table consists from two independent parts: input side used by input processor and output side used during `\write` or printing to the log and terminal. You can save the record only to one of this parts by using the nonempty ⟨*optional_prefix*⟩. If the ⟨*optional_prefix*⟩ is empty then the same record is stored twice: into input and output sides. If ⟨*optional_prefix*⟩ is a token of catcode 8 (usually the _ character) then the record is stored only into input side . If ⟨*optional_prefix*⟩ is a pair of tokens catcode 8 (usually __) then the record is stored only into output side.

If the optional prefix has a form of __ then the following ⟨*byte_sequence*⟩ can be empty. EncTEX removes the record corresponding to the ⟨*first_token*⟩ from output side in such situation.

The macro code for clearing the conversion table from previous section clears all records from input side but only the records concerned to the ⟨*first_token*⟩ in "one byte form" from output side. You can remove the record concerned to control sequence from output side only by `\mubyte \foo __\endmubyte`.

3.5. Inserted control sequences

If the ⟨*first_token*⟩ is the control sequence and the ⟨*optional_prefix*⟩ is one token of catcode 6 (usually the # character) then the ⟨*byte_sequence*⟩ is kept by input processor and only the declared control sequence is inserted before ⟨*byte_sequence*⟩. The example of usage:

```
\mubyte \warntwobytes  #^^c3^^80\endmubyte
\mubyte \warntwobytes  #^^c3^^82\endmubyte
\mubyte \warntwobytes  #^^c3^^83\endmubyte
% etc...
\def\warntwobytes #1#2{\bgroup\mubyteout=0
   \message{WARNING: the UTF-8 code: #1#2 is not defined i my macros.}
   \egroup}
```

If `\mubytein=1` and if the control sequence is inserted before ⟨*byte_sequence*⟩ then the ⟨*byte_sequence*⟩ is kept absolutely. It means no part of ⟨*byte_sequence*⟩ is converted again. On the other hand, if `\mubytein>1` then the parts of ⟨*byte_sequence*⟩ can be converted by other rules given in conversion table:

```
\mubyte \foo #ABC\endmubyte  \mubyte X BC\endmubyte
\mubytein=1  ABC is converted to \foo ABC
\mubytein=2  ABC is converted to \foo AX
```

3.6. The virtual start line mark

If `\mubytein>0` and if the first byte in ⟨*byte_sequence*⟩ is equal to `\endlinechar` (it means ⟨*byte_sequence*⟩ has a format ⟨*endlinechar*⟩⟨*rest*⟩) then input processor checks the matching of the ⟨*rest*⟩ with the begin of every line. If it matches then the given conversion is done. The example:

```
\bgroup \uccode‘X=\endlinechar \uppercase{\gdef\echar{X}}\egroup
\mubyte \fooB \echar ABC\endmubyte % ABC matches at begin of line
\mubyte \fooE ABC\echar \endmubyte % ABC matches at end of line
\mubyte \fooW \spce\space ABC\space \endmubyte
     % ABC matches as a word with spaces before and after
\mubyte \foo #\echar XYZ\endmubyte %
     % if XYZ is at begin of line the \foo is inserted before them
```

3.7. The suppression of the expansion in write parameters

If you need to convert the control sequences back to its ⟨*byte_sequences*⟩ then the expansion of such control sequences is not welcome. You can suppress the expansion by `\let\macro=\relax` before `\write` starts the expansion of its parameter. But `\write` works asynchronously in most situations and you can manipulate with hundreds or thousands control sequences declared as UTF-8 codes. The encTEX serves a simple tool to solve this problem: If `\mubyteout>=3` then encTEX gives the `\relax` meaning to each control sequence declared in output side of the conversion table before the `\write` starts its expansion and it returns back these control sequences to their original meaning immediately after `\write` finish its work. Example:

```
\mubyte \foo ABC\endmubyte  \def\foo{macro body}
\mubyteout=2
\immediate\write16{testwrite: \foo}  % prints "testwrite: macro body"
\immediate\write16{testwrite: \noexpand\foo} % prints "testwrite: ABC"
\mubyteout=3
\immediate\write16{testwrite: \foo}  % prints "testwrite: ABC"
\message{testmessage: \foo}          % prints "testmessage: macro body"
\message{testmessage: \noexpand\foo} % prints "testmessage: \foo"
\edef\a{testedef: \foo}              % expands to macro body
\foo                                 % expands to macro body
\immediate\write16{\meaning\foo}     % prints "\relax"
\message{\meaning\foo}               % prints "macro:->macro body"
```

Note the difference between `\message` and `\immediate\write16`. The control sequences in `\message` parameter are always expanded and never converted to the ⟨*byte_sequence*⟩.

You can set the “noexpand” flag (for `\write` parameters only) to any ⟨*control_sequence*⟩ and you need not declare the ⟨*byte_sequence*⟩ for it. Write `\mubyte` ⟨*control˙sequence*⟩ `\relax \endmubyte` for this purposes. This has the same effect as `\mubyte` ⟨*control_sequence*⟩ `__\string` ⟨*control_sequence*⟩`\space\endmubyte`, but this second solution is more memory consuming because TEX have to store the ⟨*byte_sequence*⟩ as a string to the pool.

3.8. The asynchronous write command and the mubyteout value

If you don’t use `\immediate` then the `\write` command first gets its parameter but it expands and prints this parameter at another time. The `\write` command stores the actual value of the `\mubyteout` register when it gets its parameter. This value is used late when parameter is expanded and written to the file.

This feature gives the possibility to write to more files, first (for table of contents, for example) is written with conversion to UTF-8 and another files are written without this conversion, because (for example) this file is an input for a program which cannot read the UTF-8 encoding. You can try:

```
\newwrite\tocfile \newwrite\indexfile
\immediate\openout\tocfile=\jobname.toc
\immediate\openout\indexfile=\jobname.idx
\mubyteout=3
\write\tocfile{this parameter will be converted to UTF-8}
{\mubyteout=0 \write\indexfile{this parameter stay unchanged}}
\write\tocfile{this parameter will be converted to UTF-8}
\end % now, all three writes are actually done
```

3.9. Summary of the mubyteout values

Apart from the values 0, 1, 2 and 3, you can set the `\mubyteout` register to the value −1 or −2. The summary table of meanings of these values follows:

`\mubyteout`	⟨*byte*⟩`->`⟨*byte_sequence*⟩	⟨*cs_name*⟩`->`⟨*byte_sequence*⟩	`noexpanding`
0	off	off	off
1	on	off	off
2	on	on	off
3	on	on	on
-1	on	off	on
-2	off	off	on

If the ⟨*byte*⟩`->`⟨*byte_sequence*⟩ conversion is on then all texts written to the `\rite` files, log file and to the terminal are converted. On the other hand, the ⟨*cs_name*⟩`->`⟨*byte_sequence*⟩ conversion and the noexpanding are related only to the `\write` arguments (and the `\special` arguments, see the following chapter).

4. The arguments of the special primitive

The plain texts of non-english languages can occur in the `\special` arguments. The PDF-outlines are the good example of this situation. May be, you need to save these arguments in UTF-8 encoding. The encTEX gives the possibility to do it.

The argument of `\special` is processed by the value of the integer primitive register `\specialout`. This register is introduced by encTEX and its default value is zero.

- `\speialout=0` – no conversion.
- `\speialout=1` – only the xchr conversion.
- `\speialout=2` – only the `\mubyteout` conversion.
- `\speialout=3` – the `\mubyteout` conversion followed by the xchr conversion.

The `\special` primitive expands its argument immediately. If `\specialout` is 2 or 3 then the expansion is done by `\mubyteout` value in the same manner as during the `\write` expansion. Moreower, `\special` saves the current values of `\specialout` and `\mubyteout` registers to its memory and use them at the time of the output to the `dvi` file.

5. The macro files

The encTEX package includes some encoding tables inputted by `\input` during format generation. These tables support encodings widely used in Czech texts. The more information about these macro files are in comments of these files and in the Czech version of the documentation.