

The listingsutf8 package

Heiko Oberdiek*

2016/05/16 v1.3

Abstract

Package `listings` does not support files with multi-byte encodings such as UTF-8. In case of `\lstinputlisting` a simple workaround is possible if an one-byte encoding exists that the file can be converted to. Also ε -TeX and pdfTeX regardless of its mode are required.

Contents

1	Documentation	2
1.1	User interface	2
1.2	Future	2
2	Implementation	2
2.1	Catcodes and identification	2
2.2	Package options	3
2.3	Check prerequisites	4
2.4	Add support for UTF-8	4
2.4.1	Conversion	4
2.4.2	Convert CR/LF pairs to LF	5
2.4.3	Patch <code>\lst@InputListing</code>	5
3	Installation	5
3.1	Download	5
3.2	Bundle installation	6
3.3	Package installation	6
3.4	Refresh file name databases	6
3.5	Some details for the interested	6
4	References	7
5	History	7
	[2007/10/22 v1.0]	7
	[2007/11/11 v1.1]	7
	[2011/11/10 v1.2]	7
	[2016/05/16 v1.3]	7
6	Index	7

*Please report any issues at <https://github.com/ho-tex/oberdiek/issues>

1 Documentation

1.1 User interface

Load this package after or instead of package `listings` [2]. The package does not define own options and passes given options to package `listings`.

The syntax of package `listings`' key `inputencoding` is extended:

```
inputencoding=utf8/⟨one-byte-encoding⟩  
Example: inputencoding=utf8/latin1
```

That means the file is encoded in UTF-8 and can be converted to the given *⟨one-byte-encoding⟩*. The available encodings for *⟨one-byte-encoding⟩* are listed in section “1.2 Supported encodings” of package `stringenc`'s documentation [3]. Of course, the encoding must encode its characters with one byte exactly. This excludes the unicode encodings (`utf8`, `utf16`, ...).

Only `\lstinputlisting` is supported by the syntax extension of key `inputencoding`.

Internally package `listingsutf8` reads the file as binary file via primitives of pdfTeX (`\pdffiledump`). Then the file contents is converted as string using package `stringenc` and finally the string is read as virtual file by ε -TeX's `\scantokens`.

1.2 Future

Workarounds are not provided for

- `\lstinline`
- Environment `lstlisting`.
- Environments defined by `\lstnewenvironment`.

Perhaps someone will find time to extend package `listings` with full native support for UTF-8. Then this package would become obsolete.

2 Implementation

```
1 ⟨*package⟩
```

2.1 Catcodes and identification

```
2 \begingroup\catcode61\catcode48\catcode32=10\relax%  
3 \catcode13=5 % ^~M  
4 \endlinechar=13 %  
5 \catcode123=1 % {  
6 \catcode125=2 % }  
7 \catcode64=11 % @  
8 \def\x{\endgroup  
9 \expandafter\edef\csname lstU@AtEnd\endcsname{%  
10 \endlinechar=\the\endlinechar\relax  
11 \catcode13=\the\catcode13\relax  
12 \catcode32=\the\catcode32\relax  
13 \catcode35=\the\catcode35\relax  
14 \catcode61=\the\catcode61\relax  
15 \catcode64=\the\catcode64\relax  
16 \catcode123=\the\catcode123\relax  
17 \catcode125=\the\catcode125\relax  
18 }%
```

```

19 }%
20 \x\catcode61\catcode48\catcode32=10\relax%
21 \catcode13=5 % ^^M
22 \endlinechar=13 %
23 \catcode35=6 % #
24 \catcode64=11 % @
25 \catcode123=1 % {
26 \catcode125=2 % }
27 \def\TMP@EnsureCode#1#2{%
28   \edef\lstU@AtEnd{%
29     \lstU@AtEnd
30     \catcode#1=\the\catcode#1\relax
31   }%
32   \catcode#1=#2\relax
33 }
34 \TMP@EnsureCode{10}{12}% ^^J
35 \TMP@EnsureCode{33}{12}% !
36 \TMP@EnsureCode{36}{3}% $
37 \TMP@EnsureCode{38}{4}% &
38 \TMP@EnsureCode{39}{12}% '
39 \TMP@EnsureCode{40}{12}% (
40 \TMP@EnsureCode{41}{12}% )
41 \TMP@EnsureCode{42}{12}% *
42 \TMP@EnsureCode{43}{12}% +
43 \TMP@EnsureCode{44}{12}% ,
44 \TMP@EnsureCode{45}{12}% -
45 \TMP@EnsureCode{46}{12}% .
46 \TMP@EnsureCode{47}{12}% /
47 \TMP@EnsureCode{58}{12}% :
48 \TMP@EnsureCode{60}{12}% <
49 \TMP@EnsureCode{62}{12}% >
50 \TMP@EnsureCode{91}{12}% [
51 \TMP@EnsureCode{93}{12}% ]
52 \TMP@EnsureCode{94}{7}% ^ (superscript)
53 \TMP@EnsureCode{95}{8}% _ (subscript)
54 \TMP@EnsureCode{96}{12}% '
55 \TMP@EnsureCode{124}{12}% |
56 \TMP@EnsureCode{126}{13}% ~ (active)
57 \edef\lstU@AtEnd{\lstU@AtEnd\noexpand\endinput}

```

Package identification.

```

58 \NeedsTeXFormat{LaTeX2e}
59 \ProvidesPackage{listingsutf8}%
60 [2016/05/16 v1.3 Allow UTF-8 in listings input (HO)]

```

2.2 Package options

Just pass options to package listings.

```

61 \DeclareOption*{%
62   \PassOptionsToPackage\CurrentOption{listings}%
63 }
64 \ProcessOptions*

```

Key inputencoding was introduced in version 2002/04/01 v1.0 of package listings.

```

65 \RequirePackage{listings}[2002/04/01]

```

Ensure that \inputencoding is provided.

```

66 \AtBeginDocument{%
67   \@ifundefined{inputencoding}{%
68     \RequirePackage{inputenc}%

```

```

69 }{}%
70 }

```

2.3 Check prerequisites

```

71 \RequirePackage{pdftexcmds}[2011/04/22]
72 \def\lstU@temp#1#2{%
73   \begingroup\expandafter\expandafter\expandafter\endgroup
74   \expandafter\ifx\csname #1\endcsname\relax
75     \PackageWarningNoLine{listingsutf8}{%
76       Package loading is aborted because of missing %
77       \@backslashchar#1.\MessageBreak
78       #2%
79     }%
80   \expandafter\lstU@AtEnd
81   \fi
82 }
83 \lstU@temp{scantokens}{It is provided by e-TeX}%
84 \lstU@temp{pdf@unescapehex}{It is provided by pdfTeX >= 1.30}%
85 \lstU@temp{pdf@filedump}{It is provided by pdfTeX >= 1.30}%
86 \lstU@temp{pdf@filesize}{It is provided by pdfTeX >= 1.30}%
87 \RequirePackage{stringenc}[2010/03/01]

```

2.4 Add support for UTF-8

```

\iflstU@utfviii
28 \newif\iflstU@utfviii

\lstU@inputenc
89 \def\lstU@inputenc#1{%
90   \expandafter\lstU@@inputenc#1utf8/utf8/\@nil
91 }

```

```

\lstU@@inputenc

92 \lst@Key{inputencoding}\relax{%
93   \def\lst@inputenc{#1}%
94   \lstU@inputenc{#1}%
95 }

```

2.4.1 Conversion

```

\lstU@input
96 \def\lstU@input#1{%
97   \iflstU@utfviii
98     \edef\lstU@text{%
99       \pdf@unescapehex{%
100         \pdf@filedump{0}{\pdf@filesize{#1}}{#1}%
101       }%
102     }%
103     \lstU@CRLFtoLF\lstU@text
104     \StringEncodingConvert\lstU@text\lstU@text{utf8}\lst@inputenc
105     \def\lstU@temp{%
106       \scantokens\expandafter{\lstU@text}%
107     }%
108   \else
109     \def\lstU@temp{%

```

```

110     \input{#1}%
111 }%
112 \fi
113 \lstU@temp
114 }

```

2.4.2 Convert CR/LF pairs to LF

\lstU@CRLFtoLF

```

115 \begingroup
116 \endlinechar=-1 %
117 \@makeother\^^J %
118 \@makeother\^^M %
119 \gdef\lstU@CRLFtoLF#1{%
120   \edef#1{%
121     \expandafter\lstU@CRLFtoLF@aux#1^^M^^J\@nil
122   }%
123 }%
124 \gdef\lstU@CRLFtoLF@aux#1^^M^^J#2\@nil{%
125   #1%
126   \ifx\relax#2\relax
127     \@car
128   \fi
129   ^^J%
130   \lstU@CRLFtoLF@aux#2\@nil
131 }%
132 \endgroup %

```

2.4.3 Patch \lst@InputListing

```

133 \def\lstU@temp#1\def\lst@next#2#3\@nil{%
134   \def\lst@InputListing##1{%
135     #1%
136     \def\lst@next{\lstU@input{##1}}%
137     #3%
138   }%
139 }
140 \expandafter\lstU@temp\lst@InputListing{#1}\@nil
141 \lstU@AtEnd%
142 \</package>

```

3 Installation

3.1 Download

Package. This package is available on CTAN¹:

[CTAN:macros/latex/contrib/oberdiek/listingsutf8.dtx](#) The source file.

[CTAN:macros/latex/contrib/oberdiek/listingsutf8.pdf](#) Documentation.

Bundle. All the packages of the bundle ‘oberdiek’ are also available in a TDS compliant ZIP archive. There the packages are already unpacked and the documentation files are generated. The files and directories obey the TDS standard.

[CTAN:install/macros/latex/contrib/oberdiek.tds.zip](#)

¹[CTAN:pkg/listingsutf8](#)

TDS refers to the standard “A Directory Structure for T_EX Files” (CTAN:pkg/t_{ds}). Directories with `texmf` in their name are usually organized this way.

3.2 Bundle installation

Unpacking. Unpack the `oberdiek.tds.zip` in the TDS tree (also known as `texmf` tree) of your choice. Example (linux):

```
unzip oberdiek.tds.zip -d ~/texmf
```

3.3 Package installation

Unpacking. The `.dtx` file is a self-extracting `docstrip` archive. The files are extracted by running the `.dtx` through plain T_EX:

```
tex listingsutf8.dtx
```

TDS. Now the different files must be moved into the different directories in your installation TDS tree (also known as `texmf` tree):

```
listingsutf8.sty → tex/latex/oberdiek/listingsutf8.sty
listingsutf8.pdf → doc/latex/oberdiek/listingsutf8.pdf
listingsutf8.dtx → source/latex/oberdiek/listingsutf8.dtx
```

If you have a `docstrip.cfg` that configures and enables `docstrip`’s TDS installing feature, then some files can already be in the right place, see the documentation of `docstrip`.

3.4 Refresh file name databases

If your T_EX distribution (T_EX Live, mikT_EX, ...) relies on file name databases, you must refresh these. For example, T_EX Live users run `texhash` or `mktexlsr`.

3.5 Some details for the interested

Unpacking with L^AT_EX. The `.dtx` chooses its action depending on the format:

plain T_EX: Run `docstrip` and extract the files.

L^AT_EX: Generate the documentation.

If you insist on using L^AT_EX for `docstrip` (really, `docstrip` does not need L^AT_EX), then inform the autodetect routine about your intention:

```
latex \let\install=y\input{listingsutf8.dtx}
```

Do not forget to quote the argument according to the demands of your shell.

Generating the documentation. You can use both the `.dtx` or the `.drv` to generate the documentation. The process can be configured by the configuration file `ltxdoc.cfg`. For instance, put this line into this file, if you want to have A4 as paper format:

```
\PassOptionsToClass{a4paper}{article}
```

An example follows how to generate the documentation with pdfL^AT_EX:

```
pdflatex listingsutf8.dtx
makeindex -s gind.ist listingsutf8.idx
pdflatex listingsutf8.dtx
makeindex -s gind.ist listingsutf8.idx
pdflatex listingsutf8.dtx
```

4 References

- [1] Alan Jeffrey, Frank Mittelbach, *inputenc.sty*, 2006/05/05 v1.1b. [CTAN:pkg/inputenc](#)
- [2] Carsten Heinz, Brooks Moses: *The listings package*; 2007/02/22; [CTAN:pkg/listings](#).
- [3] Heiko Oberdiek: *The stringenc package*; 2007/10/22; [CTAN:pkg/stringenc](#).

5 History

[2007/10/22 v1.0]

- First version.

[2007/11/11 v1.1]

- Use of package pdfdoccmds.

[2011/11/10 v1.2]

- DOS line ends CR/LF normalized to LF to avoid empty lines (Bug report of Thomas Benkert in de.comp.text.tex).

[2016/05/16 v1.3]

- Documentation updates.

6 Index

Numbers written in *italic* refer to the page where the corresponding entry is described; numbers underlined refer to the code line of the definition; plain numbers refer to the code lines where the entry is used.

Symbols		E	
\@backslashchar	77	\endcsname	9, 74
\@car	127	\endinput	57
\@ifundefined	67	\endlinechar	4, 10, 22, 116
\@makeother	117, 118		
\@nil	90, 121, 124, 130, 133, 140		
\^	117, 118		
A		G	
\AtBeginDocument	66	\gdef	119, 124
C		I	
\catcode	2, 3, 5, 6, 7, 11, 12, 13, 14, 15, 16, 17, 20, 21, 23, 24, 25, 26, 30, 32	\iflstU@utfviii	88, 97
\csname	9, 74	\ifx	74, 126
\CurrentOption	62	\input	110
D		L	
\DeclareOption	61	\lst@inputenc	93, 104
		\lst@InputListing	134, 140
		\lst@Key	92
		\lst@next	133, 136
		\lstU@@inputenc	90, <u>92</u>
		\lstU@AtEnd	28, 29, 57, 80, 141

