

The listingsutf8 package

Heiko Oberdiek
<oberdiek@uni-freiburg.de>

2007/10/22 v1.0

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	1
1.1	User interface	1
1.2	Future	2
2	Implementation	2
2.1	Catcodes and identification	2
2.2	Package options	3
2.3	Check prerequisites	3
2.4	Add support for UTF-8	3
2.4.1	Conversion	4
2.4.2	Patch <code>\lst@InputListing</code>	4
3	Test	4
3.1	Catcode checks for loading	4
3.2	Test example for latin1	5
4	Installation	6
4.1	Download	6
4.2	Bundle installation	6
4.3	Package installation	6
4.4	Refresh file name databases	6
4.5	Some details for the interested	7
5	References	7
6	History	7
	[2007/10/22 v1.0]	7
7	Index	7

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 \expandafter\edef\csname lstU@AtEnd\endcsname{%
3   \catcode64 \the\catcode64\relax
4 }
5 \catcode64 11 % @
6 \def\TMP@EnsureCode#1#2{%
7   \edef\lstU@AtEnd{%
8     \lstU@AtEnd
9     \catcode#1 \the\catcode#1\relax
10  }%
11  \catcode#1 #2\relax
12 }
13 \TMP@EnsureCode{10}{12}% ^^J
14 \TMP@EnsureCode{33}{12}% !
15 \TMP@EnsureCode{36}{3}% $
16 \TMP@EnsureCode{38}{4}% &
17 \TMP@EnsureCode{39}{12}% '
18 \TMP@EnsureCode{40}{12}% (
19 \TMP@EnsureCode{41}{12}% )
20 \TMP@EnsureCode{42}{12}% *
21 \TMP@EnsureCode{43}{12}% +
22 \TMP@EnsureCode{44}{12}% ,
23 \TMP@EnsureCode{45}{12}% -
24 \TMP@EnsureCode{46}{12}% .
25 \TMP@EnsureCode{47}{12}% /
26 \TMP@EnsureCode{58}{12}% :
27 \TMP@EnsureCode{60}{12}% <
28 \TMP@EnsureCode{61}{12}% =
29 \TMP@EnsureCode{62}{12}% >
30 \TMP@EnsureCode{94}{7}% ^ (superscript)
```

```

31 \TMP@EnsureCode{95}{8}% _ (subscript)
32 \TMP@EnsureCode{96}{12}% ‘
33 \TMP@EnsureCode{124}{12}% |
34 \TMP@EnsureCode{126}{13}% ~ (active)
35 \g@addto@macro\lstU@AtEnd{\endinput}

Package identification.
36 \NeedsTeXFormat{LaTeX2e}
37 \ProvidesPackage{listingsutf8}%
38 [2007/10/22 v1.0 Adding support for UTF-8 to listings (HO)]

```

2.2 Package options

Just pass options to package listings.

```

39 \DeclareOption*{%
40   \PassOptionsToPackage\CurrentOption{listings}%
41 }
42 \ProcessOptions*

```

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

```

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

```

Ensure that \inputencoding is provided.

```

44 \AtBeginDocument{%
45   \ifundefined{inputencoding}{%
46     \RequirePackage{inputenc}%
47   }{}%
48 }

```

2.3 Check prerequisites

```

49 \def\lstU@temp#1#2{%
50   \begingroup\expandafter\expandafter\expandafter\endgroup
51   \expandafter\ifx\csname #1\endcsname\relax
52     \PackageWarningNoLine{listingsutf8}{%
53       Package loading is aborted because of missing %
54       \@backslashchar#1.\MessageBreak
55       #2%
56     }%
57   \expandafter\lstU@AtEnd
58   \fi
59 }
60 \lstU@temp{scantokens}{It is provided by e-TeX}
61 \lstU@temp{pdfunescapehex}{It is provided by pdfTeX >= 1.30}
62 \lstU@temp{pdffiledump}{It is provided by pdfTeX >= 1.30}
63 \lstU@temp{pdffilesize}{It is provided by pdfTeX >= 1.30}
64 \RequirePackage{stringenc}[2007/10/22]

```

2.4 Add support for UTF-8

```
\iflstU@utfviii
```

```
65 \newif\iflstU@utfviii
```

```
\lstU@inputenc
```

```

66 \def\lstU@inputenc#1{%
67   \expandafter\lstU@@inputenc#1utf8/utf8/\@nil
68 }

```

```
\lstU@@inputenc
```

```

69 \lst@Key{inputencoding}\relax{%
70   \def\lst@inputenc#1{%
71     \lstU@inputenc#1}%
72 }

```

2.4.1 Conversion

```
\lstU@input
73 \def\lstU@input#1{%
74   \iflstU@utfviii
75     \edef\lstU@text{%
76       \pdfunescapehex{%
77         \pdffiledump length\pdffilesize{#1}{#1}%
78       }%
79     }%
80     \StringEncodingConvert\lstU@text\lstU@text{utf8}\lst@inputenc
81   \def\lstU@temp{%
82     \scantokens\expandafter{\lstU@text}%
83   }%
84   \else
85     \def\lstU@temp{%
86       \input{#1}%
87     }%
88   \fi
89   \lstU@temp
90 }
```

2.4.2 Patch \lst@InputListing

```
91 \def\lstU@temp#1\def\lst@next#2#3\@nil{%
92   \def\lst@InputListing##1{%
93     #1%
94     \def\lst@next{\lstU@input{##1}}%
95     #3%
96   }%
97 }
98 \expandafter\lstU@temp\lst@InputListing{#1}\@nil
99 \lstU@AtEnd
100 </package>
```

3 Test

3.1 Catcode checks for loading

```
101 <*test1>
102 \NeedsTeXFormat{LaTeX2e}
103 \documentclass{minimal}
104 \makeatletter
105 \def\RestoreCatcodes{}
106 \count@=0 %
107 \loop
108   \edef\RestoreCatcodes{%
109     \RestoreCatcodes
110     \catcode\the\count@=\the\catcode\count@\relax
111   }%
112   \ifnum\count@<255 %
113     \advance\count@\@ne
114   \repeat
115
116 \def\RangeCatcodeInvalid#1#2{%
117   \count@=#1\relax
118   \loop
119     \catcode\count@=15 %
120     \ifnum\count@<#2\relax
121       \advance\count@\@ne
122     \repeat
123 }
```

```

124 \def\Test{%
125   \RangeCatcodeInvalid{0}{47}%
126   \RangeCatcodeInvalid{58}{64}%
127   \RangeCatcodeInvalid{91}{96}%
128   \RangeCatcodeInvalid{123}{127}%
129   \catcode'\@=12 %
130   \catcode'\=0 %
131   \catcode'\{=1 %
132   \catcode'\}=2 %
133   \catcode'\#=6 %
134   \catcode'\[=12 %
135   \catcode'\]=12 %
136   \catcode'\%=14 %
137   \catcode'\ =10 %
138   \catcode13=5 %
139   \RequirePackage{listingsutf8}[2007/10/22]\relax
140   \RestoreCatcodes
141 }
142 \Test
143 \csname @@end\endcsname
144 \end
145 </test1>

```

3.2 Test example for latin1

```

146 <*test2>
147 \NeedsTeXFormat{LaTeX2e}
148 \documentclass{minimal}
149 \usepackage{filecontents}
150 \def\do#1{%
151   \ifx#1\^%
152     \else
153       \noexpand\do\noexpand#1%
154     \fi
155 }
156 \expandafter\let\expandafter\dospecials\expandafter\empty
157 \expandafter\edef\expandafter\dospecials\expandafter{\dospecials}
158 \begin{filecontents*}{ExampleUTF8.java}
159 public class ExampleUTF8 {
160     public static String testString =
161         "Umlauts: " +
162         "^^c3^^84^^c3^^96^^c3^^9c^^c3^^a4^^c3^^b6^^c3^^bc^^c3^^9f";
163     public static void main(String[] args) {
164         System.out.println(testString);
165     }
166 }
167 \end{filecontents*}
168 \usepackage{listingsutf8}[2007/10/22]
169 \def\Text{%
170   Umlauts: %
171   ^^c3^^84^^c3^^96^^c3^^9c^^c3^^a4^^c3^^b6^^c3^^bc^^c3^^9f%
172 }
173 \begin{document}
174 \lstinputlisting[%
175   language=Java,%
176   inputencoding=utf8/latin1,%
177 ]{ExampleUTF8.java}
178 \end{document}
179 </test2>

```

4 Installation

4.1 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:macros/latex/contrib/oberdiek/oberdiek-tds.zip](#)

TDS refers to the standard “A Directory Structure for T_EX Files” ([CTAN:tds/tds.pdf](#)). Directories with `texmf` in their name are usually organized this way.

4.2 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
```

Script installation. Check the directory `TDS:scripts/oberdiek/` for scripts that need further installation steps. Package `attachfile2` comes with the Perl script `pdfatfi.pl` that should be installed in such a way that it can be called as `pdfatfi`. Example (linux):

```
chmod +x scripts/oberdiek/pdfatfi.pl
cp scripts/oberdiek/pdfatfi.pl /usr/local/bin/
```

4.3 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):

<code>listingsutf8.sty</code>	→ <code>tex/latex/oberdiek/listingsutf8.sty</code>
<code>listingsutf8.pdf</code>	→ <code>doc/latex/oberdiek/listingsutf8.pdf</code>
<code>test/listingsutf8-test1.tex</code>	→ <code>doc/latex/oberdiek/test/listingsutf8-test1.tex</code>
<code>test/listingsutf8-test2.tex</code>	→ <code>doc/latex/oberdiek/test/listingsutf8-test2.tex</code>
<code>test/listingsutf8-test3.tex</code>	→ <code>doc/latex/oberdiek/test/listingsutf8-test3.tex</code>
<code>test/listingsutf8-test4.tex</code>	→ <code>doc/latex/oberdiek/test/listingsutf8-test4.tex</code>
<code>test/listingsutf8-test5.tex</code>	→ <code>doc/latex/oberdiek/test/listingsutf8-test5.tex</code>
<code>listingsutf8.dtx</code>	→ <code>source/latex/oberdiek/listingsutf8.dtx</code>

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`.

4.4 Refresh file name databases

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

¹<http://ftp.ctan.org/tex-archive/>

4.5 Some details for the interested

Attached source. The PDF documentation on CTAN also includes the `.dtx` source file. It can be extracted by AcrobatReader 6 or higher. Another option is `pdftk`, e.g. unpack the file into the current directory:

```
pdftk listingsutf8.pdf unpack_files output .
```

Unpacking with \LaTeX . The `.dtx` chooses its action depending on the format:

plain- \TeX : Run `docstrip` and extract the files.

\LaTeX : Generate the documentation.

If you insist on using \LaTeX for `docstrip` (really, `docstrip` does not need \LaTeX), 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 `pdf \LaTeX` :

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

5 References

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

6 History

[2007/10/22 v1.0]

- First version.

7 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; numbers in roman refer to the code lines where the entry is used.

Symbols	\%	136
\#		133

\@	129	\lst@inputenc	70, 80
\@backslashchar	54	\lst@InputListing	92, 98
\@ifundefined	45	\lst@Key	69
\@one	113, 121	\lst@next	91, 94
\@nil	67, 91, 98	\lstinputlisting	174
\[134	\lstU@inputenc	67, 69
\	130	\lstU@AtEnd	7, 8, 35, 57, 99
\{	131	\lstU@input	73, 94
\}	132	\lstU@inputenc	66, 71
\]	135	\lstU@temp	49,
\^	151		60, 61, 62, 63, 81, 85, 89, 91, 98
		\lstU@text	75, 80, 82
_	137		
A			
\advance	113, 121	\makeatletter	104
\AtBeginDocument	44	\MessageBreak	54
B			
\begin	158, 173		
C			
\catcode	3,		
	5, 9, 11, 110, 119, 129, 130, 131,		
	132, 133, 134, 135, 136, 137, 138		
\count@	106,		
	110, 112, 113, 117, 119, 120, 121		
\csname	2, 51, 143		
\CurrentOption	40		
D			
\DeclareOption	39		
\do	150, 153		
\documentclass	103, 148		
\dospecials	156, 157		
E			
\empty	156		
\end	144, 167, 178		
\endcsname	2, 51, 143		
\endinput	35		
G			
\g@addto@macro	35		
I			
\iflstU@utfviii	65, 74		
\ifnum	112, 120		
\ifx	51, 151		
\input	86		
L			
\loop	107, 118		
M			
N			
\NeedsTeXFormat	36, 102, 147		
\newif	65		
P			
\PackageWarningNoLine	52		
\PassOptionsToPackage	40		
\pdffiledump	77		
\pdffilesize	77		
\pdfunescapehex	76		
\ProcessOptions	42		
\ProvidesPackage	37		
R			
\RangeCatcodeInvalid			
	116, 125, 126, 127, 128		
\repeat	114, 122		
\RequirePackage	43, 46, 64, 139		
\RestoreCatcodes	105, 108, 109, 140		
S			
\scantokens	82		
\StringEncodingConvert	80		
T			
\Test	124, 142		
\Text	169		
\the	3, 9, 110		
\TMP@EnsureCode	6, 13, 14, 15,		
	16, 17, 18, 19, 20, 21, 22, 23, 24,		
	25, 26, 27, 28, 29, 30, 31, 32, 33, 34		
U			
\usepackage	149, 168		