

The listingsutf8 package

Heiko Oberdiek
<heiko.oberdiek at gmail.com>

2007/11/11 v1.1

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	7
4.5	Some details for the interested	7
5	References	7
6	History	7
	[2007/10/22 v1.0]	7
	[2007/11/11 v1.1]	8
7	Index	8

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
3   \catcode123 1 % {
4   \catcode125 2 % }
5   \def\x{\endgroup
6     \expandafter\edef\csname lstU@AtEnd\endcsname{%
7       \catcode35 \the\catcode35\relax
8       \catcode64 \the\catcode64\relax
9       \catcode123 \the\catcode123\relax
10      \catcode125 \the\catcode125\relax
11    }%
12  }%
13 \x
14 \catcode35 6 % #
15 \catcode64 11 % @
16 \catcode123 1 % {
17 \catcode125 2 % }
18 \def\TMP@EnsureCode#1#2{%
19   \edef\lstU@AtEnd{%
20     \lstU@AtEnd
21     \catcode#1 \the\catcode#1\relax
22   }%
23   \catcode#1 #2\relax
24 }
25 \TMP@EnsureCode{10}{12}% ^^J
26 \TMP@EnsureCode{33}{12}% !
27 \TMP@EnsureCode{36}{3}% $
28 \TMP@EnsureCode{38}{4}% &
29 \TMP@EnsureCode{39}{12}% '
30 \TMP@EnsureCode{40}{12}% (
```

```

31 \TMP@EnsureCode{41}{12}% )
32 \TMP@EnsureCode{42}{12}% *
33 \TMP@EnsureCode{43}{12}% +
34 \TMP@EnsureCode{44}{12}% ,
35 \TMP@EnsureCode{45}{12}% -
36 \TMP@EnsureCode{46}{12}% .
37 \TMP@EnsureCode{47}{12}% /
38 \TMP@EnsureCode{58}{12}% :
39 \TMP@EnsureCode{60}{12}% <
40 \TMP@EnsureCode{61}{12}% =
41 \TMP@EnsureCode{62}{12}% >
42 \TMP@EnsureCode{94}{7}% ^ (superscript)
43 \TMP@EnsureCode{95}{8}% _ (subscript)
44 \TMP@EnsureCode{96}{12}% '
45 \TMP@EnsureCode{124}{12}% |
46 \TMP@EnsureCode{126}{13}% ~ (active)
47 \g@addto@macro\lstU@AtEnd{\endinput}

Package identification.
48 \NeedsTeXFormat{LaTeX2e}
49 \ProvidesPackage{listingsutf8}%
50 [2007/11/11 v1.1 Adding support for UTF-8 to listings (HO)]

```

2.2 Package options

Just pass options to package listings.

```

51 \DeclareOption*{%
52   \PassOptionsToPackage\CurrentOption{listings}%
53 }
54 \ProcessOptions*

```

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

```

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

```

Ensure that \inputencoding is provided.

```

56 \AtBeginDocument{%
57   \ifundefined{inputencoding}{%
58     \RequirePackage{inputenc}%
59   }{}%
60 }

```

2.3 Check prerequisites

```

61 \RequirePackage{pdftexcmds}[2007/11/11]
62 \def\lstU@temp#1#2{%
63   \begingroup\expandafter\expandafter\expandafter\endgroup
64   \expandafter\ifx\csname #1\endcsname\relax
65     \PackageWarningNoLine{listingsutf8}{%
66       Package loading is aborted because of missing %
67       \@backslashchar#1.\MessageBreak
68       #2%
69     }%
70     \expandafter\lstU@AtEnd
71   \fi
72 }
73 \lstU@temp{scantokens}{It is provided by e-TeX}
74 \lstU@temp{pdf@unescapehex}{It is provided by pdfTeX >= 1.30}
75 \lstU@temp{pdf@filedump}{It is provided by pdfTeX >= 1.30}
76 \lstU@temp{pdf@filesize}{It is provided by pdfTeX >= 1.30}
77 \RequirePackage{stringenc}[2007/10/22]

```

2.4 Add support for UTF-8

```
\iflstU@utfviii
```

```

78 \newif\iflstU@utfviii

\lstU@inputenc

79 \def\lstU@inputenc#1{%
80   \expandafter\lstU@@inputenc#1utf8/utf8/\@nil
81 }

```

\lstU@@inputenc

```

82 \lst@Key{inputencoding}\relax{%
83   \def\lst@inputenc{#1}%
84   \lstU@inputenc{#1}%
85 }

```

2.4.1 Conversion

\lstU@input

```

86 \def\lstU@input#1{%
87   \iflstU@utfviii
88     \edef\lstU@text{%
89       \pdf@unescapehex{%
90         \pdf@filedump{0}{\pdf@filesize{#1}}{#1}%
91       }%
92     }%
93     \StringEncodingConvert\lstU@text\lstU@text{utf8}\lst@inputenc
94     \def\lstU@temp{%
95       \scantokens\expandafter{\lstU@text}%
96     }%
97   \else
98     \def\lstU@temp{%
99       \input{#1}%
100     }%
101   \fi
102   \lstU@temp
103 }

```

2.4.2 Patch \lst@InputListing

```

104 \def\lstU@temp#1\def\lst@next#2#3\@nil{%
105   \def\lst@InputListing##1{%
106     #1%
107     \def\lst@next{\lstU@input{##1}}%
108     #3%
109   }%
110 }
111 \expandafter\lstU@temp\lst@InputListing{#1}\@nil
112 \lstU@AtEnd
113 \</package>

```

3 Test

3.1 Catcode checks for loading

```

114 \<test1>
115 \NeedsTeXFormat{LaTeX2e}
116 \documentclass{minimal}
117 \makeatletter
118 \def\RestoreCatcodes{}
119 \count@=0 %
120 \loop
121   \edef\RestoreCatcodes{%

```

```

122     \RestoreCatcodes
123     \catcode\the\count@=\the\catcode\count@\relax
124 }%
125 \ifnum\count@<255 %
126     \advance\count@\@ne
127 \repeat
128
129 \def\RangeCatcodeInvalid#1#2{%
130     \count@=#1\relax
131     \loop
132         \catcode\count@=15 %
133     \ifnum\count@<#2\relax
134         \advance\count@\@ne
135     \repeat
136 }
137 \def\Test{%
138     \RangeCatcodeInvalid{0}{47}%
139     \RangeCatcodeInvalid{58}{64}%
140     \RangeCatcodeInvalid{91}{96}%
141     \RangeCatcodeInvalid{123}{127}%
142     \catcode'\@=12 %
143     \catcode'\=0 %
144     \catcode'\{=1 %
145     \catcode'\}=2 %
146     \catcode'\#=6 %
147     \catcode'\[=12 %
148     \catcode'\]=12 %
149     \catcode'\%=14 %
150     \catcode'\ =10 %
151     \catcode13=5 %
152     \RequirePackage{listingsutf8}[2007/11/11]\relax
153     \RestoreCatcodes
154 }
155 \Test
156 \csname @@end\endcsname
157 \end
158 </test1>

```

3.2 Test example for latin1

```

159 (*test2)
160 \NeedsTeXFormat{LaTeX2e}
161 \documentclass{minimal}
162 \usepackage{filecontents}
163 \def\do#1{%
164     \ifx#1\^%
165     \else
166         \noexpand\do\noexpand#1%
167     \fi
168 }
169 \expandafter\let\expandafter\dospecials\expandafter\empty
170 \expandafter\edef\expandafter\dospecials\expandafter{\dospecials}
171 \begin{filecontents*}{ExampleUTF8.java}
172 public class ExampleUTF8 {
173     public static String testString =
174         "Umlauts: " +
175         "^^c3^^84^^c3^^96^^c3^^9c^^c3^^a4^^c3^^b6^^c3^^bc^^c3^^9f";
176     public static void main(String[] args) {
177         System.out.println(testString);
178     }
179 }
180 \end{filecontents*}
181 \usepackage{listingsutf8}[2007/11/11]

```

```

182 \def\Text{%
183   Umlauts: %
184   ^^c3^^84^^c3^^96^^c3^^9c^^c3^^a4^^c3^^b6^^c3^^bc^^c3^^9f%
185 }
186 \begin{document}
187 \lstinputlisting[%
188   language=Java,%
189   inputencoding=utf8/latin1,%
190 ]{ExampleUTF8.java}
191 \end{document}
192 </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:install/macros/latex/contrib/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>

¹<http://ftp.ctan.org/tex-archive/>

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 \TeX distribution (`te\TeX`, `mik\TeX`, ...) relies on file name databases, you must refresh these. For example, `te\TeX` users run `texhash` or `mktextlsr`.

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.

- Use of package pdfdoccmds.

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		L	
\#	146	\loop	120, 131
\%	149	\lst@inputenc	83, 93
\@	142	\lst@InputListing	105, 111
\@backslashchar	67	\lst@Key	82
\@ifundefined	57	\lst@next	104, 107
\@ne	126, 134	\lstinputlisting	187
\@nil	80, 104, 111	\lstU@inputenc	80, <u>82</u>
\[147	\lstU@AtEnd	19, 20, 47, 70, 112
\]	143	\lstU@input	<u>86</u> , 107
\{	144	\lstU@inputenc	<u>79</u> , 84
\}	145	\lstU@temp	62,
\]	148		73, 74, 75, 76, 94, 98, 102, 104, 111
\^	164	\lstU@text	88, 93, 95
A		M	
_	150	\makeatletter	117
B		\MessageBreak	67
\advance	126, 134	N	
\AtBeginDocument	56	\NeedsTeXFormat	48, 115, 160
C		\newif	78
\begin	171, 186	P	
\catcode	3,	\PackageWarningNoLine	65
	4, 7, 8, 9, 10, 14, 15, 16, 17,	\PassOptionsToPackage	52
	21, 23, 123, 132, 142, 143, 144,	\pdf@filedump	90
	145, 146, 147, 148, 149, 150, 151	\pdf@filesize	90
\count@	119,	\pdf@unescapehex	89
	123, 125, 126, 130, 132, 133, 134	\ProcessOptions	54
\csname	6, 64, 156	\ProvidesPackage	49
\CurrentOption	52	R	
D		\RangeCatcodeInvalid	129, 138, 139, 140, 141
\DeclareOption	51	\repeat	127, 135
\do	163, 166	\RequirePackage	55, 58, 61, 77, 152
\documentclass	116, 161	\RestoreCatcodes	118, 121, 122, 153
\dospecials	169, 170	S	
E		\scantokens	95
\empty	169	\StringEncodingConvert	93
\end	157, 180, 191	T	
\endcsname	6, 64, 156	\Test	137, 155
\endinput	47	\Text	182
G		\the	7, 8, 9, 10, 21, 123
\g@addto@macro	47	\TMP@EnsureCode	18, 25, 26, 27,
I			28, 29, 30, 31, 32, 33, 34, 35, 36,
\iflstU@utfviii	<u>78</u> , 87		37, 38, 39, 40, 41, 42, 43, 44, 45, 46
\ifnum	125, 133	U	
\ifx	64, 164	\usepackage	162, 181
\input	99	X	
		\x	5, 13