The listingsutf8 package

Heiko Oberdiek*
<heiko.oberdiek at googlemail.com>

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 . . . 3
2.4 Add support for UTF-8 . . . 4
2.4.1 Conversion . . . 4
2.4.2 Convert CR/LF pairs to LF . . . 4
2.4.3 Patch \lst@InputListing . . . 5

3 Test **5**
3.1 Catcode checks for loading . . . 5
3.2 Test example for latin1 . . . 6

4 Installation **6**
4.1 Download . . . 6
4.2 Bundle installation . . . 7
4.3 Package installation . . . 7
4.4 Refresh file name databases . . . 7
4.5 Some details for the interested . . . 7

5 Catalogue **8**

6 References **8**

7 History **9**
[2007/10/22 v1.0] . . . 9
[2007/11/11 v1.1] . . . 9
[2011/11/10 v1.2] . . . 9
[2016/05/16 v1.3] . . . 9

8 Index **9**

*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 %
```

```
\catcode35=6 % #
\catcode64=11 % @
\catcode123=1 % {
\catcode125=2 % }
\def\TMP@EnsureCode#1#2{%
  \edef\lstU@AtEnd{%
    \lstU@AtEnd
    \catcode#1=\the\catcode#1\relax
  }%
  \catcode#1=#2\relax
}
\TMP@EnsureCode{10}{12}% ^^J
\TMP@EnsureCode{33}{12}% !
\TMP@EnsureCode{36}{3}% $
\TMP@EnsureCode{38}{4}% &
\TMP@EnsureCode{39}{12}% '
\TMP@EnsureCode{40}{12}% (
\TMP@EnsureCode{41}{12}% )
\TMP@EnsureCode{42}{12}% *
\TMP@EnsureCode{43}{12}% +
\TMP@EnsureCode{44}{12}% ,
\TMP@EnsureCode{45}{12}% -
\TMP@EnsureCode{46}{12}% .
\TMP@EnsureCode{47}{12}% /
\TMP@EnsureCode{58}{12}% :
\TMP@EnsureCode{60}{12}% <
\TMP@EnsureCode{62}{12}% >
\TMP@EnsureCode{91}{12}% [
\TMP@EnsureCode{93}{12}% ]
\TMP@EnsureCode{94}{7}% ^ (superscript)
\TMP@EnsureCode{95}{8}% _ (subscript)
\TMP@EnsureCode{96}{12}% `
\TMP@EnsureCode{124}{12}% |
\TMP@EnsureCode{126}{13}% ~ (active)
\edef\lstU@AtEnd{\lstU@AtEnd\noexpand\endinput}
```

Package identification.

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

2.2 Package options

Just pass options to package listings.

```
\DeclareOption*{%
  \PassOptionsToPackage\CurrentOption{listings}%
}
\ProcessOptions*
```

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

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

Ensure that \inputencoding is provided.

```
\AtBeginDocument{%
  \@ifundefined{inputencoding}{%
    \RequirePackage{inputenc}%
  }{}%
}
```

2.3 Check prerequisites

```
\RequirePackage{pdftexcmds}[2011/04/22]
```

```
\def\lstU@temp#1#2{%
  \begingroup\expandafter\expandafter\expandafter\endgroup
```

```
  \expandafter\ifx\csname #1\endcsname\relax
    \PackageWarningNoLine{listingsutf8}{%
      Package loading is aborted because of missing %
      \@backslashchar#1.\MessageBreak
      #2%
    }%
    \expandafter\lstU@AtEnd
  \fi
}
\lstU@temp{scantokens}{It is provided by e-TeX}%
\lstU@temp{pdf@unescapehex}{It is provided by pdfTeX >= 1.30}%
\lstU@temp{pdf@filedump}{It is provided by pdfTeX >= 1.30}%
\lstU@temp{pdf@filesize}{It is provided by pdfTeX >= 1.30}%
\RequirePackage{stringenc}[2010/03/01]
```

2.4 Add support for UTF-8

\iflstU@utfviii

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

\lstU@inputenc

```
\def\lstU@inputenc#1{%
  \expandafter\lstU@@inputenc#1utf8/utf8/\@nil
}
```

\lstU@@inputenc

```
\lst@Key{inputencoding}\relax{%
  \def\lst@inputenc{#1}%
  \lstU@inputenc{#1}%
}
```

2.4.1 Conversion

\lstU@input

```
\def\lstU@input#1{%
  \iflstU@utfviii
    \edef\lstU@text{%
      \pdf@unescapehex{%
        \pdf@filedump{0}{\pdf@filesize{#1}}{#1}%
      }%
    }%
    \lstU@CRLFtoLF\lstU@text
    \StringEncodingConvert\lstU@text\lstU@text{utf8}\lst@inputenc
    \def\lstU@temp{%
      \scantokens\expandafter{\lstU@text}%
    }%
  \else
    \def\lstU@temp{%
      \input{#1}%
    }%
  \fi
  \lstU@temp
}
```

2.4.2 Convert CR/LF pairs to LF

\lstU@CRLFtoLF

```
\begingroup
  \endlinechar=-1 %
  \@makeother\^^J %
  \@makeother\^^M %
```

```
  \gdef\lstU@CRLFtoLF#1{%
    \edef#1{%
      \expandafter\lstU@CRLFtoLF@aux#1^^M^^J\@nil
    }%
  }%
  \gdef\lstU@CRLFtoLF@aux#1^^M^^J#2\@nil{%
    #1%
    \ifx\relax#2\relax
      \@car
    \fi
    ^^J%
    \lstU@CRLFtoLF@aux#2\@nil
  }%
\endgroup %
```

2.4.3 Patch \lst@InputListing

```
\def\lstU@temp#1\def\lst@next#2#3\@nil{%
  \def\lst@InputListing##1{%
    #1%
    \def\lst@next{\lstU@input{##1}}%
    #3%
  }%
}
\expandafter\lstU@temp\lst@InputListing{#1}\@nil
\lstU@AtEnd%
⟨/package⟩
```

3 Test

3.1 Catcode checks for loading

```
⟨*test1⟩
\NeedsTeXFormat{LaTeX2e}
\documentclass{minimal}
\makeatletter
\def\RestoreCatcodes{}
\count@=0 %
\loop
  \edef\RestoreCatcodes{%
    \RestoreCatcodes
    \catcode\the\count@=\the\catcode\count@\relax
  }%
\ifnum\count@<255 %
  \advance\count@\@ne
\repeat

\def\RangeCatcodeInvalid#1#2{%
  \count@=#1\relax
  \loop
    \catcode\count@=15 %
  \ifnum\count@<#2\relax
    \advance\count@\@ne
  \repeat
}
\def\Test{%
  \RangeCatcodeInvalid{0}{47}%
  \RangeCatcodeInvalid{58}{64}%
  \RangeCatcodeInvalid{91}{96}%
  \RangeCatcodeInvalid{123}{127}%
  \catcode`\@=12 %
  \catcode`\\=0 %
```

```
  \catcode`\{=1 %
  \catcode`\}=2 %
  \catcode`\#=6 %
  \catcode`\[=12 %
  \catcode`\]=12 %
  \catcode`\%=14 %
  \catcode`\ =10 %
  \catcode13=5 %
  \RequirePackage{listingsutf8}[2016/05/16]\relax
  \RestoreCatcodes
}
\Test
\csname @@end\endcsname
\end
```

⟨/test1⟩

3.2 Test example for latin1

⟨*test2⟩

```
\NeedsTeXFormat{LaTeX2e}
\documentclass{minimal}
\usepackage{filecontents}
\def\do#1{%
  \ifx#1\^%
  \else
    \noexpand\do\noexpand#1%
  \fi
}
\expandafter\let\expandafter\dospecials\expandafter\empty
\expandafter\edef\expandafter\dospecials\expandafter{\dospecials}
\begin{filecontents*}{ExampleUTF8.java}
public class ExampleUTF8 {
    public static String testString =
        "Umlauts: " +
        "^^c3^^84^^c3^^96^^c3^^9c^^c3^^a4^^c3^^b6^^c3^^bc^^c3^^9f";
    public static void main(String[] args) {
        System.out.println(testString);
    }
}
\end{filecontents*}
\usepackage{listingsutf8}[2016/05/16]
\def\Text{%
  Umlauts: %
  ^^c3^^84^^c3^^96^^c3^^9c^^c3^^a4^^c3^^b6^^c3^^bc^^c3^^9f%
}
\begin{document}
\lstinputlisting[%
  language=Java,%
  inputencoding=utf8/latin1,%
]{ExampleUTF8.java}
\end{document}
```

⟨/test2⟩

4 Installation

4.1 Download

Package. This package is available on CTAN[1]:

CTAN:macros/latex/contrib/oberdiek/listingsutf8.dtx The source file.

CTAN:macros/latex/contrib/oberdiek/listingsutf8.pdf Documentation.

[1] http://ctan.org/pkg/listingsutf8

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 TEX 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 TEX:

```
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
test/listingsutf8-test1.tex → doc/latex/oberdiek/test/listingsutf8-test1.tex
test/listingsutf8-test2.tex → doc/latex/oberdiek/test/listingsutf8-test2.tex
test/listingsutf8-test3.tex → doc/latex/oberdiek/test/listingsutf8-test3.tex
test/listingsutf8-test4.tex → doc/latex/oberdiek/test/listingsutf8-test4.tex
test/listingsutf8-test5.tex → doc/latex/oberdiek/test/listingsutf8-test5.tex
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.

4.4 Refresh file name databases

If your TEX distribution (teTEX, mikTEX, ...) relies on file name databases, you must refresh these. For example, teTEX users run texhash or mktexlsr.

4.5 Some details for the interested

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 pdfLaTeX:

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

5 Catalogue

The following XML file can be used as source for the TeX Catalogue. The elements caption and description are imported from the original XML file from the Catalogue. The name of the XML file in the Catalogue is listingsutf8.xml.

```
⟨*catalogue⟩
<?xml version='1.0' encoding='us-ascii'?>
<!DOCTYPE entry SYSTEM 'catalogue.dtd'>
<entry datestamp='$Date$' modifier='$Author$' id='listingsutf8'>
  <name>listingsutf8</name>
  <caption>Allow UTF-8 in listings input.</caption>
  <authorref id='auth:oberdiek'/>
  <copyright owner='Heiko Oberdiek' year='2007,2011'/>
  <license type='lppl1.3'/>
  <version number='1.3'/>
  <description>
    Package <xref refid='listings'>listings</xref> does not support files
    with multi-byte encodings such as UTF-8.  In the case of
    <tt>\lstinputlisting</tt>, a simple workaround is possible if a
    one-byte encoding exists that the file can be converted to.  The
    package requires the e-TeX extensions under pdfTeX (in either PDF
    or DVI output mode).
    <p/>
    The package is part of the <xref refid='oberdiek'>oberdiek</xref> bundle.
  </description>
  <documentation details='Package documentation'
      href='ctan:/macros/latex/contrib/oberdiek/listingsutf8.pdf'/>
  <ctan file='true' path='/macros/latex/contrib/oberdiek/listingsutf8.dtx'/>
  <miktex location='oberdiek'/>
  <texlive location='oberdiek'/>
  <install path='/macros/latex/contrib/oberdiek/oberdiek.tds.zip'/>
</entry>
⟨/catalogue⟩
```

6 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.

7 History

[2007/10/22 v1.0]

- First version.

[2007/11/11 v1.1]

- Use of package pdftexcmds.

[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.

8 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

\# 175
\% 178
\@ 171
\@backslashchar 77
\@car 127
\@ifundefined 67
\@makeother 117, 118
\@ne 155, 163
\@nil 90, 121, 124, 130, 133, 140
\[. 176
\\ 172
\{ 173
\} 174
\] 177
\^ 117, 118, 193

\␣ 179

A

\advance 155, 163
\AtBeginDocument 66

B

\begin 200, 215

C

\catcode 2, 3, 5, 6, 7, 11, 12, 13, 14, 15, 16, 17, 20, 21, 23, 24, 25, 26, 30, 32, 152, 161, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180
\count@ 148, 152, 154, 155, 159, 161, 162, 163
\csname 9, 74, 185
\CurrentOption 62

D

\DeclareOption 61
\do 192, 195
\documentclass 145, 190
\dospecials 198, 199

E

\empty 198
\end 186, 209, 220
\endcsname 9, 74, 185
\endinput 57
\endlinechar 4, 10, 22, 116

G

\gdef 119, 124

I

\iflstU@utfviii 88, 97
\ifnum 154, 162
\ifx 74, 126, 193
\input 110

L

\loop 149, 160
\lst@inputenc 93, 104
\lst@InputListing 134, 140
\lst@Key 92
\lst@next 133, 136
\lstinputlisting 216, 235
\lstU@@inputenc 90, 92
\lstU@AtEnd 28, 29, 57, 80, 141
\lstU@CRLFtoLF 103, 115
\lstU@CRLFtoLF@aux . . 121, 124, 130
\lstU@input 96, 136
\lstU@inputenc 89, 94

\lstU@temp 72, 83, 84, 85, 86, 105, 109, 113, 133, 140
\lstU@text 98, 103, 104, 106

M

\makeatletter 146
\MessageBreak 77

N

\NeedsTeXFormat 58, 144, 189
\newif 88

P

\PackageWarningNoLine 75
\PassOptionsToPackage 62
\pdf@filedump 100
\pdf@filesize 100
\pdf@unescapehex 99
\ProcessOptions 64
\ProvidesPackage 59

R

\RangeCatcodeInvalid 158, 167, 168, 169, 170
\repeat 156, 164
\RequirePackage . . . 65, 68, 71, 87, 181
\RestoreCatcodes . . 147, 150, 151, 182

S

\scantokens 106
\StringEncodingConvert 104

T

\Test 166, 184
\Text 211
\the 10, 11, 12, 13, 14, 15, 16, 17, 30, 152
\TMP@EnsureCode 27, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56

U

\usepackage 191, 210

X

\x 8, 20