The sanitize-umlaut package
Manual for version 1.3.0 (2023/05/15)

Thomas F. Sturm

https://www.ctan.org/pkg/sanitize-umlaut
https://github.com/T-F-S/sanitize-umlaut

Abstract
The packages sanitizes umlauts to be used directly in index entries for MakeIndex and friends with \pdflatex. This means, that inside \index an umlaut can be used as "U or Ü. In both cases, the letter is written as "U into the raw index file for correct processing with MakeIndex and \pdflatex. \lualatex and \xelatex are also supported with a different approach.

Contents

1 Purpose of the Package 2
2 Important Compatibility Informations 3
  2.1 Past 3
  2.2 Present 3
  2.3 Future 3
3 Package Usage 4
  3.1 Prerequisites 4
  3.2 Package Application 4
  3.3 Sanitized Characters 4
  3.4 Technical Information 5
4 Application Examples 6
1 Purpose of the Package

The package sanitizes umlauts to be used directly in index entries for \texttt{makeindex} and friends with \texttt{pdflatex}. This means, that inside \texttt{\index} an umlaut can be used as "Ü" or Ü. In both cases, the letter is written as "U" into the raw index file for correct processing with \texttt{makeindex} and \texttt{pdflatex}.

The package is intended

- for documents in German language using the babel package with a setting identical or similar to \texttt{\usepackage[ngerman]{babel}}.
- for documents which are processed by \texttt{latex} or \texttt{pdflatex} (also for \texttt{lualatex} or \texttt{xelatex}, but with more compilation overhead).
- for documents with an index which is processed using the MakeIndex program.
- for authors who like to use \texttt{\index{Übermaß}} instead of \texttt{\index{"Uberma"s}}.

All these conditions are satisfiable by simply including the \texttt{sanitize-umlaut} package.

An alternative would be to filter the resulting raw .\texttt{idx} index before \texttt{makeindex} is applied to create the final .\texttt{ind} index. Another alternative is to replace MakeIndex by Xindy or another index processor.
2 Important Compatibility Informations

2.1 Past

Until 2018, the default encoding for LATEX files was 7-bit ASCII. For other encodings, packages like inputenc had to be loaded. Also, inputenc used to expand characters like umlauts during \index output. The package sanitize-umlaut version 1.00 replaced this expansion code for \index output to get "\={U}" instead of "Ü", etc.

2.2 Present

Since April 2018, the default encoding for LATEX files has been changed to UTF-8. This is done by preloading the UTF-8 settings of the package inputenc by default LATEX, i.e. if you want to use UTF-8 (recommended!), you do not longer need to load inputenc inside your preamble. But, also the implementation of inputenc changed for UTF-8 (October 2019?). Nowadays, characters like umlauts are not longer expanded during \index output, but are preserved as is. Therefore, sanitize-umlaut version 1.00 is not compatible to inputenc with UTF-8 dating from 2019 or newer.

sanitize-umlaut version 1.10 (or newer) patches some UTF-8 code of LATEX/inputenc to return and replace character expansion during \index output. This patch is not compatible to older versions of LATEX/inputenc (before October 2019). Therefore, if your LATEX distribution is not reasonable up to date, you should stay at version 1.00 of sanitize-umlaut.

With the 2022 June release of LATEX, characters defined via utf8.def are now defined as \protected macros. Therefore, sanitize-umlaut version 1.2.0 (or newer) patches some relevant parts of two-octets characters during \index back to pre 2022 June behaviour. Obviously, you loose \protected here, if you load sanitize-umlaut.

sanitize-umlaut version 1.3.0 (or newer) also supports lualatex and xelatex with a different approach. Here, \index is patched such that its argument is processed to replace umlauts.

2.3 Future

As always, the future is dark and difficult to see. Further changes of inputenc implementation may force further changes of sanitize-umlaut. Hopefully, this will not happen too soon or too often. Also, if some miracle happens, MakeIndex may be updated one day to recognize UTF-8 properly to make sanitize-umlaut superfluous.
3 Package Usage

3.1 Prerequisites

The source document may need some encoding by inputenc since pdflatex is assumed as engine. For example:

\usepackage[latin1]{inputenc}

For utf8 (UTF-8), modern \LaTeX{} does not need this package inclusion any more! Also, for lualatex and xelatex this has to be omitted.

Just some few encodings are supported by sanitize-umlaut. These are the most important for German language texts:

<table>
<thead>
<tr>
<th>encoding</th>
<th>recognized as</th>
</tr>
</thead>
<tbody>
<tr>
<td>utf8</td>
<td>utf8</td>
</tr>
<tr>
<td>utf8-2018</td>
<td>utf8-2018</td>
</tr>
<tr>
<td>latin1, ansinew, cp1252</td>
<td>latin1</td>
</tr>
<tr>
<td>applemac</td>
<td>applemac</td>
</tr>
</tbody>
</table>

Further, the babel package with German settings is needed:

\usepackage[ngerman]{babel}

3.2 Package Application

Now, the package application is simple. You just put

\usepackage{sanitize-umlaut}

into your document preamble after inputenc and, maybe, after babel. That is all.

3.3 Sanitized Characters

The umlauts and the sharp s are replaced by their babel shorthand codes which are written to the .idx file.

<table>
<thead>
<tr>
<th>character</th>
<th>replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>ä</td>
<td>&quot;a&quot;</td>
</tr>
<tr>
<td>ö</td>
<td>&quot;o&quot;</td>
</tr>
<tr>
<td>ü</td>
<td>&quot;u&quot;</td>
</tr>
<tr>
<td>Ä</td>
<td>&quot;A&quot;</td>
</tr>
<tr>
<td>Ù</td>
<td>&quot;O&quot;</td>
</tr>
<tr>
<td>Ù</td>
<td>&quot;U&quot;</td>
</tr>
<tr>
<td>ß</td>
<td>&quot;s&quot;</td>
</tr>
</tbody>
</table>
3.4 Technical Information

The package uses `\inputencodingname` (set by \LaTeX{} and the \texttt{inputencoding} package) to determine the current encoding.

The package redefines the `\sanitize` macro at the begin of the document. It adds some encoding redefinitions to this macro. `\sanitize` is used inside `\index` in a local group. If another package (besides \texttt{babel}) also changes this macro or uses it outside `\index`, strange things may happen.

If `\inputencodingname` is not present, the package checks, if the current engine is \texttt{luatex} or \texttt{xetex} and patches the `\index` macro itself. All umlauts inside the argument of `\index` are replaced by their \texttt{babel} shorthand codes using \LaTeX{}3 token replacement. This increases compilation time considerably compared to the `\sanitize` hack for \texttt{pdflatex}. A very rough figure is approximately a plus of 0.8 seconds per 10000 `\index` calls (will differ on other machines and other example codes).
4 Application Examples

file "german.ist" for the examples

actual '=' % instead of @
quote '!.' % instead of "
level '>' % instead of !

% !TeX encoding=UTF-8
% arara: pdflatex
% arara: makeindex: { style: german.ist, german: true }
% arara: pdflatex
\documentclass\[a4paper,12pt\]{article}
\usepackage\[T1\]{fontenc}
%\usepackage[utf8]{inputenc} % utf8 is default now
\usepackage\[ngerman\]{babel}
\usepackage{makeidx}
\usepackage{sanitize-umlaut}
\makeindex
\begin{document}
\section{Basic Example}
Test äöüÄÖÜß.
\index{Aber}
\index{Arg}
\index{Ärger}
\index{Ofen}
\index{Ö - wie schön}
\index{oberhalb}
\index{Ufer}
\index{Übermaß}
\index{LATEX=La\TeX\}}
\index{Ärger>Index}
Test äöüÄÖÜß.
\printindex
\end{document}
Example with hyperref

Test äöüÄÖÜß.

\index{Aber} \index{Arg} \index{Ärger}
\index{Ofen} \index{Ö - wie schön} \index{oberhalb}
\index{Ufer} \index{Übermaß}
\index{Latex=\LaTeX} \index{Ärger>Index}

Test äöüÄÖÜß.

\printindex

Index
Aber, 1
Arg, 1
Ärger, 1
Ofen, 1
Ö - wie schön, 1
oberhalb, 1
Ufer, 1
Übermaß, 1
\LaTeX, 1
1 Example with imakeidx

Test äöüÄÖÜß.

\index{Aber} \index{Arg} \index{Ärger}
\index{Ofen} \index{Ö - wie schön} \index{oberhalb}
\index{Ufer} \index{Übermaß}
\index{Latex=\LaTeX} \index{Ärger>Index}

Test äöüÄÖÜß.

\printindex

\end{document}
Example with imakeidx and hyperref

Test äöüÄÖÜß.

\begin{document}
\section{Example with imakeidx and hyperref}
Test äöüÄÖÜß.
\printindex
\end{document}
# Example with multiple indexes

Test äöüÄÖÜß.

## Personenregister

<table>
<thead>
<tr>
<th>Name</th>
<th>Seite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huber, Hans</td>
<td>1</td>
</tr>
<tr>
<td>Hübner, Jörg</td>
<td>1</td>
</tr>
</tbody>
</table>

## Allgemeines Register

<table>
<thead>
<tr>
<th>Wörter</th>
<th>Seite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aber</td>
<td>1</td>
</tr>
<tr>
<td>Ärger</td>
<td>1</td>
</tr>
<tr>
<td>Arg</td>
<td>1</td>
</tr>
<tr>
<td>LATEX</td>
<td>1</td>
</tr>
<tr>
<td>Ö - wie schön</td>
<td>1</td>
</tr>
<tr>
<td>Ofen</td>
<td>1</td>
</tr>
<tr>
<td>Übermaß</td>
<td>1</td>
</tr>
<tr>
<td>Ufer</td>
<td>1</td>
</tr>
<tr>
<td>LATEX=\LaTeX</td>
<td>1</td>
</tr>
<tr>
<td>Ärger&gt;Index</td>
<td>1</td>
</tr>
</tbody>
</table>

Test äöüÄÖÜß.

\clearpage
\printindex[allgemein]
\printindex[personen]
Example with multiple indexes for lualatex

Test äöüÄÖÜß.

\index[personen]{Huber, Hans} \index[personen]{Hübner, Jörg}
\index[allgemein]{Aber} \index[allgemein]{Arg}
\index[allgemein]{Ärger} \index[allgemein]{Ofen}
\index[allgemein]{O - wie schon} \index[allgemein]{oberhalb}
\index[allgemein]{Ufer} \index[allgemein]{Übermaß}
\index[allgemein]{Latex=\LaTeX} \index[allgemein]{Ärger>Index}

Test äöüÄÖÜß.
\printindex[allgemein]
\printindex[personen]