The sanitize-umlaut package
Manual for version 1.10 (2020/01/01)

Thomas F. Sturm

http://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 \texttt{pdflatex}. This means, that inside \texttt{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 \texttt{pdflatex}.

1 Purpose of the Package

2 Important Compatibility Informations
  2.1 Past ................................................................. 2
  2.2 Present ............................................................ 2
  2.3 Future ............................................................... 2

3 Package Usage
  3.1 Prerequisites ...................................................... 3
  3.2 Package Application ............................................... 3
  3.3 Sanitized Characters ............................................... 3
  3.4 Technical Information ............................................ 3

4 Application Examples

---

1Prof. Dr. Dr. Thomas F. Sturm, Institut für Mathematik und Informatik, Universität der Bundeswehr München, D-85577 Neubiberg, Germany; email: thomas.sturm@unibw.de
1 Purpose of the Package

The package 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 "Ü" or "Ü". In both cases, the letter is written as "Ü" into the raw index file for correct processing with `makeindex` and `pdflatex`.

The package is intended

- for documents in German language using the babel package with a setting identical or similar to `\usepackage[ngerman]{babel}`.
- for documents which are processed by `latex` or `pdflatex` (not `lualatex` or `xelatex`).
- for documents with an index which is processed using the MakeIndex program.
- for authors who like to use `\index{Übermaß}` instead of `\index{"Ubermaß"}`.

All these conditions are satisfiable by simply including the `sanitize-umlaut` package.

An alternative would be to filter the resulting raw `.idx` index before `makeindex` is applied to create the final `.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 "Ü" 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`.

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 \texttt{inputenc} since \texttt{pdflatex} is assumed as engine. For example:

\begin{verbatim}
\usepackage[latin1]{inputenc}
\end{verbatim}

For utf8 (UTF-8), modern \LaTeX{} does not need this package inclusion any more!

Just some few encodings are supported by \texttt{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 \texttt{babel} package with German settings is needed:

\begin{verbatim}
\usepackage[ngerman]{babel}
\end{verbatim}

3.2 Package Application

Now, the package application is simple. You just put

\begin{verbatim}
\usepackage{sanitize-umlaut}
\end{verbatim}

into your document preamble \texttt{after inputenc} and, maybe, after \texttt{babel}. That is all.

3.3 Sanitized Characters

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

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

3.4 Technical Information

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

The package redefines the \texttt{\@sanitize} macro at the begin of the document. It adds some encoding redefinitions to this macro. \texttt{\@sanitize} is used inside \texttt{\index} in a local group. If another package (besides \texttt{babel}) also changes this macro or uses it outside \texttt{\index}, strange things may happen.
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=\LaTeX}
\index{Ärger>Index}
Test äöüÄÖÜß.
\printindex
\end{document}
1 Example with hyperref
Test äöüÄÖÜß. Test äöüÄÖÜß.

Index
Aber, 1
Ärger, 1
Arg, 1
LateX, 1
oberhalb, 1
Ö - wie schön, 1
Ofen, 1
Übermaß, 1
Ufer, 1

Index
Aber, 1
Ärger, 1
Arg, 1
LateX, 1
oberhalb, 1
Ö - wie schön, 1
Ofen, 1
Übermaß, 1
Ufer, 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 \texttt{imakeidx} and \texttt{hyperref}

Test äöüÄÖÜß.

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

Test äöüÄÖÜß.

\printindex
Example with multiple indexes

Test äöüÄÖÜß.

Index

<table>
<thead>
<tr>
<th>Personen</th>
<th>Huber, Hans</th>
<th>Hübner, Jörg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allgemein</td>
<td>Aber</td>
<td>Arg</td>
</tr>
<tr>
<td>Allgemein</td>
<td>Arger</td>
<td>Ö - wie schön</td>
</tr>
<tr>
<td>Allgemein</td>
<td>Ofen</td>
<td>Oberhalb</td>
</tr>
<tr>
<td>Allgemein</td>
<td>Ufer</td>
<td>Übermaß</td>
</tr>
<tr>
<td>Allgemein</td>
<td>LATEX</td>
<td>Ärger &gt; Index</td>
</tr>
</tbody>
</table>

Test äöüÄÖÜß.

Clearpage

Printindex [Allgemein]

Printindex [Personen]

Allgemeines Register

| Aber | 1 |
| Arg | 1 |
| LATEX | 1 |

Personenregister

| Huber, Hans | 1 |
| Hübner, Jörg | 1 |

1 Example with multiple indexes

Test äöüÄÖÜß. Test äöüÄÖÜß.