TYPO3 Core - Bug #20612

scandinavian letters are translittered wrong

2009-06-12 22:24 - Katja Lampela

Status:	Needs Feedback	**Start date:**	2009-06-12
Priority:	Should have	**Due date:**	
Assignee:		**% Done:**	0%
Category:	Localization	**Estimated time:**	0.00 hour
Target version:	Candidate for Major Version		
TYPO3 Version:	4.2	**Complexity:**	
PHP Version:		**Is Regression:**	No
Tags:		**Sprint Focus:**	

Description

The scandinavian letters ä, ö and å are rendered, with for example realurl, in a wrong way. Ä is ae, ö is oe.. Ä should be a, ö o and å a.

This should be added to the file typo3_src-X/t3lib/unidata/Translit.txt:

1. scandinavian
   00e4; 0061; LATIN SMALL LETTER A WITH UMLAUTS => a (finnish)
   00c4; 0041; LATIN CAPITAL LETTER A WITH UMLAUTS => A (finnish)
   00f6; 006f; LATIN SMALL LETTER O WITH UMLAUTS => o (finnish)
   00d6; 004f; LATIN CAPITAL LETTER O WITH UMLAUTS => O (finnish)
   00e5; 0061; LATIN SMALL LETTER SWEDISH A (Å) => a (finnish)
   00c5; 0041; LATIN CAPITAL LETTER SWEDISH A (Å) => a (finnish)
   (issue imported from #M11322)

Related issues:

Related to TYPO3 Core - Bug #67187: recursiveFileListSortingHelper natural so...	**New**	**2015-05-29**
Related to TYPO3 Core - Task #83546: Unit test CharsetConverter::specCharsToA...	**Closed**	**2018-01-12**
Has duplicate TYPO3 Core - Bug #83438: Respect suomi in specCharsToASCII conv...	**Closed**	**2017-12-28**

Associated revisions

Revision ec5d31ee - 2018-01-12 15:38 - Reiner Teubner

[TASK] Test cases for function specCharsToASCII()

Add a new test for the function specCharsToASCII().

Resolves: #83546
Related: #20612
Releases: master
Change-Id: Id255ab953ef7c1865a7db1892b9b5d5fac87c547
Reviewed-on: https://review.typo3.org/55333
Reviewed-by: Reiner Teubner <rteubner@me.com>
Tested-by: Reiner Teubner <rteubner@me.com>
Tested-by: TYPO3com <no-reply@typo3.com>
Reviewed-by: Oliver Klee <typo3-coding@oliverklee.de>
Reviewed-by: Anja Leichsenring <aleichsenring@ab-softlab.de>
Tested-by: Anja Leichsenring <aleichsenring@ab-softlab.de>
Reviewed-by: Christian Kuhn <lolli@schwarzbu.ch>
Tested-by: Christian Kuhn <lolli@schwarzbu.ch>

History

#1 - 2009-06-18 14:05 - Martin Kutschker

Unfortunately the transliteration is currently language independent. The vowels with umlauts are transliterated according to German custom. Which is fine for the huge German user base.

As a workaround you may change Translit.txt and and delete all files in typo3temp/cs.

#2 - 2013-05-08 21:14 - Alexander Opitz

- Status changed from New to Needs Feedback

- Target version deleted (0)

- TYPO3 Version set to 4.2

The issue is very old, does this issue exists in newer versions of TYPO3 CMS (4.5 or 6.1)?

#3 - 2013-05-09 06:12 - Katja Lampela

Yes it still exists in 4.5-4.7. Haven't tried 6, but I suspect nothing has been done there either.

#4 - 2013-05-09 08:34 - Alexander Opitz

- Status changed from Needs Feedback to New

#5 - 2015-01-16 17:20 - Mathias Schreiber

- Target version set to 7.2 (Frontend)

- Is Regression set to No

#6 - 2015-04-08 12:03 - Riccardo De Contardi

Still present in 6.2.11 I guess, the file /typo3/sysext/core/Resources/Private/Charsets/unidata/Translit.txt does not contain a "Scandinavian" section.

What about adding an option (in Install tool?) where to specify a custom file instead of modifying the original in the core?

#7 - 2015-06-15 18:12 - Benni Mack

- Target version changed from 7.2 (Frontend) to 7.4 (Backend)

#8 - 2015-08-05 10:31 - Susanne Moog

- Target version changed from 7.4 (Backend) to 7.5

#9 - 2015-09-08 05:21 - Joonas Kauhanen

Riccardo De Contardi wrote:

> Still present in 6.2.11 I guess, the file /typo3/sysext/core/Resources/Private/Charsets/unidata/Translit.txt does not contain a "Scandinavian" section.
>
> What about adding an option (in Install tool?) where to specify a custom file instead of modifying the original in the core?

This would be very handy. Now we have to manually update the Translit.txt file every time after updating TYPO3 source. As many of the developers are from central Europe, they don't understand the issue. And I think it's not an issue strictly to scandinavians, as the transliterations should really be language specific, not system wide.

Hoping this to be part of 7 LTS

#10 - 2015-09-24 20:42 - Benni Mack

- Target version changed from 7.5 to 7 LTS

#11 - 2015-10-16 18:24 - Jan Helke

- Assignee set to Jan Helke

#12 - 2015-10-16 18:32 - Gerrit Code Review

- Status changed from New to Under Review

Patch set 1 for branch **master** of project **Packages/TYPO3.CMS** has been pushed to the review server.
It is available at http://review.typo3.org/44115

#13 - 2015-10-29 09:52 - Christian Kuhn

- Status changed from Under Review to New

#14 - 2016-03-18 15:24 - Riccardo De Contardi

- Target version changed from 7 LTS to Candidate for Major Version

#15 - 2016-10-26 10:12 - Bart Lammers

This still seems to be the case. For a client this is a big issue in regards to speaking URLs that have (in their eyes) wrong letters in them.
The solution as provided earlier by Jan Helke seems like a nice solution, using the Install Tool to override the default transliteration files.

Back then the reasoning to not accept this was to rework the CharsetConverter completely, but now all trough version 4.x, 6.2LTS, 7.6LTS and now 8.x this has not been addressed.

Is it possible to reconsider the given solution?

#16 - 2017-05-10 21:41 - Benni Mack

Hey Bart,

maybe we can use a PHP library to do the work now that we cleaned up CharsetConverter big time, do you know any good places to look for?

#17 - 2017-10-29 16:45 - Susanne Moog

- Category set to Localization

- Assignee deleted (Jan Helke)

#18 - 2018-01-11 13:59 - Markus Klein

- Has duplicate Bug #83438: Respect suomi in specCharsToASCII conversion method added

#19 - 2018-01-12 13:25 - Christian Kuhn

- Related to Task #83546: Unit test CharsetConverter::specCharsToASCII() added

#20 - 2018-09-30 00:46 - Benni Mack

- Status changed from New to Needs Feedback

Hey,

this issue should be fixed with 9 LTS and site handling. Please let us know if the new version will solve your issue, otherwise we'll close this ticket in the next weeks.

Benni.

#21 - 2018-10-29 16:19 - Joonas Kauhanen

We have now tested the new 9 LTS with Site Handling and Url Routing. Unfortunately the issue is still not completely resolved, but we are getting near!

Let me go through the issue by example:

- Editor creates a page in Finnish language, titled "Ääni" (that is "Sound" in Finnish)
- In the Page properties, TYPO3 shows generated URL Segment "/aeaeni"
- This is incorrect for Finnish (and Swedish, Danish, etc) language, so the editor must manually override the URL Segment as "/aani"

For pages, manually correcting all URL segments is manageable, but somehow irritating. However, if the URL contains content from other records such as News or an integrated Product database, there may not even be an opportunity to input correct URL segments and all the generated slugs are translittered wrong.

I have tracked the URL generation down to sysext/core/Classes/DataHandling/SlugHelper.php where the function sanitize is used to make slugs URL compatible. This function uses CharsetConverter->specCharsToASCII to convert extended letters to ASCII characters. No locale or language parameters are provided, and the CharsetConverter uses a hardcoded path to Resources/Private/Charsets/unidata/Translit.txt where the character replacements are defined. This file is only compatible with German way of spelling umlaut characters.

Do you have ideas how this could be improved so that we could somehow provide conversion tables for languages other than German? Essentially we still have to edit Resources/Private/Charsets/unidata/Translit.txt manually after updating TYPO3 and remember to clear typo3temp cache after that.

#22 - 2019-03-04 22:48 - Susanne Moog

Some research notes:

- PHP has a Transliterator in the `intl` extension, but we'd need to provide custom rules there, too.
- Our Symfony `intl` polyfill does not polyfill the Transliterator meaning we'd have to introduce `intl` as hard dependency
- There seems to be no decent PHP transliteration solution covering our use cases out of the box

May be possible:
- Provide the possibility to register own translit.txt file in localconf to allow integrators to use custom transliterations

#23 - 2019-12-22 16:41 - Alexander Schnitzler

Susanne Moog wrote:

> Some research notes:
>
> - PHP has a Transliterator in the `intl` extension, but we'd need to provide custom rules there, too.
> - Our Symfony `intl` polyfill does not polyfill the Transliterator meaning we'd have to introduce `intl` as hard dependency
> - There seems to be no decent PHP transliteration solution covering our use cases out of the box
>
> May be possible:
> - Provide the possibility to register own translit.txt file in localconf to allow integrators to use custom transliterations

Just a short notice that **symfony/string**, which has been released just a couple of days ago, does a really great job, normalizing and converting string, respecting locales properly:

```
(new AsciiSlugger('de'))->slug('Näe ja koe')->toString(); // Naee-ja-koe
(new AsciiSlugger('dk'))->slug('Näe ja koe')->toString(); // Nae-ja-koe
```

We should evaluate how much own logic we actually need if using this library.
Could replace our current slugger and parts of the CharsetConverter (maybe even all of it).

#24 - 2020-09-10 14:50 - Mathias Bolt Lesniak

- File transliterate-norwegian.png added

- File transliterate-english.png added

- File transliterate-german.png added

- File transliterate-swedish.png added

Alexander Schnitzler wrote:

> We should evaluate how much own logic we actually need if using this library.
> Could replace our current slugger and parts of the CharsetConverter (maybe even all of it).

I made a first implementation the AsciiSlug class in symfony/string. It's working OK, but I notice TYPO3 doesn't transliterate e.g. Chinese and Hindi characters. AsciiSlug does.

This implementation uses the page language to determine how to transliterate the string, so & can be transliterated as "og" in Norwegian and "und" in German.

Implementation: https://github.com/pixelant/transliterator

Files

transliterate-norwegian.png	63.7 KB	2020-09-10	Mathias Bolt Lesniak
transliterate-english.png	55.3 KB	2020-09-10	Mathias Bolt Lesniak
transliterate-german.png	63.9 KB	2020-09-10	Mathias Bolt Lesniak
transliterate-swedish.png	63.8 KB	2020-09-10	Mathias Bolt Lesniak