

## Smuxi - Feature # 288: automatic character recoding (e.g. latin1 <-> utf8)

<b>Status:</b>	Closed	<b>Priority:</b>	Normal
<b>Author:</b>	Michael Schmitt	<b>Category:</b>	Engine IRC
<b>Created:</b>	01/11/2010	<b>Assigned to:</b>	Ondrej Hoříšek
<b>Updated:</b>	05/27/2013	<b>Due date:</b>	
<b>Complexity:</b>	High		
<b>Subject:</b>	automatic character recoding (e.g. latin1 <-> utf8)		
<b>Description:</b>	I did a small survey, as most users are ignorant and do not want to change their encoding, smuxi should recode as necessary as all major IRC clients do it nowadays anyway.		

### Associated revisions

#### 05/26/2013 10:28 PM - Ondrej Hoříšek

[Engine-(IRC), Frontend-GNOME] Support try-UTF-8-first encoding (closes #288).

Add a new connection option, AutoConvertUTF8, which will make outgoing messages UTF-8 and incoming messages UTF-8 with falling back to the chosen encoding if invalid UTF-8.

### History

#### 01/11/2010 12:00 PM - Mirco Bauer

The perl regex on this page might help to detect UTF-8 characters:  
<http://www.w3.org/International/questions/qa-forms-utf-8.en.php>

#### 01/11/2010 12:08 PM - Mirco Bauer

- Assigned to deleted (Mirco Bauer)

#### 08/21/2010 10:34 AM - Mirco Bauer

In case the URL breaks or the content vanishes, here a snapshot of it:

```
<pre>
$field =~
  m\A(
    [\x09\x0A\x0D\x20-\x7E]      # ASCII
    | [\xC2-\xDF][\x80-\xBF]    # non-overlong 2-byte
    | \xE0[\xA0-\xBF][\x80-\xBF] # excluding overlongs
    | [\xE1-\xEC\xEE\xEF][\x80-\xBF]{2} # straight 3-byte
    | \xED[\x80-\x9F][\x80-\xBF] # excluding surrogates
    | \xF0[\x90-\xBF][\x80-\xBF]{2} # planes 1-3
    | [\xF1-\xF3][\x80-\xBF]{3}  # planes 4-15
    | \xF4[\x80-\x8F][\x80-\xBF]{2} # plane 16
  )*\z/x;
```

This expression can be adapted to other programming languages. It takes care of various issues, such as illegal overlong encodings and illegal use of surrogates. It will return true if \$field is UTF-8, and false otherwise.

```
</pre>
```

#### 08/22/2010 01:18 PM - Mirco Bauer

- Target version changed from 0.8 to TBD

#### 08/22/2010 04:13 PM - Mirco Bauer

The branch that tries to deal with this:

[http://git.qnetp.net/?p=smuxi.git;a=shortlog;h=refs/heads/feature/%23288\\_automatic\\_character\\_recoding](http://git.qnetp.net/?p=smuxi.git;a=shortlog;h=refs/heads/feature/%23288_automatic_character_recoding)

It can detect UTF8 but the recode part is not working.

**09/16/2010 09:24 PM - Raphaël Hertzog**

+1 from me, this is really needed, it's one of the regressions that annoy me the most.

I see lots of ? instead of the accented characters on #debian-devel-fr. Some are going through correctly (for those that send UTF-8).

Example:

20:35 <bubulle> et, indirectement, ? cause d'un bug de dak, ?a m'emp?che d'envoyer une mise ? jour de s?curit? dans t-p-u

22:38 <KiBi> oué hein :)

22:39 <KiBi> (faire des IO ? quel drôle d'idée pour une machine qui fait du SQL..)

**09/23/2010 03:54 PM - Mirco Bauer**

- Target version changed from TBD to 0.10

- Complexity set to High

Ok, I can't recode from ISO 8859-1 if it was already converted from raw bytes to a string as it removes UTF-8 values during that. The IRC lib has to either expose the raw bytes or handle the transformation.

Here the chat with alan about this issue:

<pre>

17:48:40 <meebey> I think my issue is that the encoders are stripping invalid values

17:48:51 <meebey> but I am not sure, maybe I am just too stupid

17:49:01 <meebey> the initial issue is that the input is not byte[]

17:49:10 <meebey> it is already parsed bytes in strings

17:49:17 <alan> It's too late then :)

17:49:21 <meebey> say iso8859-15

17:49:27 <meebey> but it preserves the utf8 values

17:49:33 <meebey> at least I can see them

17:49:42 <alan> byte[] -> string is a lossy conversion if you have an invalid byte sequence

17:49:53 <meebey> sure?

17:49:55 <alan> so if you have invalid utf8 you discard those chars

17:49:55 <alan> aye

17:49:55 <alan> 100%

17:49:58 <alan> i hit this before :)

17:50:04 <alan> you have to use the raw bytes

17:50:10 <meebey> ok thanks

</pre>

**05/26/2013 09:16 PM - Mirco Bauer**

- Category changed from Engine to Engine IRC

- Target version changed from 0.10 to 0.9

**05/26/2013 11:54 PM - Mirco Bauer**

- Assigned to set to Ondrej HoÅjek

**05/27/2013 01:03 AM - Ondrej HoÅjek**

- Status changed from New to Closed

- % Done changed from 0 to 100

Applied in changeset commit:"bc2f323a9006ced64e3808550bead652e792c9e1".