

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

<b>Status:</b>	New	<b>Priority:</b>	Normal
<b>Author:</b>	Michael Schmitt	<b>Category:</b>	Engine
<b>Created:</b>	01/11/2010	<b>Assigned to:</b>	
<b>Updated:</b>	08/22/2010	<b>Due date:</b>	
<b>Resolution:</b>			
<b>Complexity:</b>			
<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.		

### 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/^(
  [\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 0.8.0 to 0.9.0

#### 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.