## Redmine - Defect #12641

**Diff outputs become ??? in some non ASCII words.**

2012-12-19 09:30 - Toshi MARUYAMA

<table>
<thead>
<tr>
<th>Status:</th>
<th>Closed</th>
<th>Start date:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority:</td>
<td>Normal</td>
<td>Due date:</td>
<td></td>
</tr>
<tr>
<td>Assignee:</td>
<td>Toshi MARUYAMA</td>
<td>% Done:</td>
<td>0%</td>
</tr>
<tr>
<td>Category:</td>
<td>I18n</td>
<td>Estimated time:</td>
<td>0.00 hour</td>
</tr>
<tr>
<td>Target version:</td>
<td>2.3.0</td>
<td>Affected version:</td>
<td>2.1.4</td>
</tr>
<tr>
<td>Resolution:</td>
<td>Fixed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Description

An example is r11052 in #12640#note-2.

![diff-r11052.png](attachment:diff-r11052.png)

### Related issues:

- Related to Redmine - Patch # 12640: Russian "about_x_hours" translation change - Closed

### Associated revisions

**Revision 11544 - 2013-03-07 08:24 - Toshi MARUYAMA**

remove unnecessary h() from diff filename (#12641)

On Rails3, escaping is default.

**Revision 11545 - 2013-03-07 09:31 - Toshi MARUYAMA**

move utf8 encoding from view to UnifiedDiff (#12641)

**Revision 11546 - 2013-03-07 09:47 - Toshi MARUYAMA**

code cleanup (#12641)

**Revision 11547 - 2013-03-07 11:17 - Toshi MARUYAMA**

set html encoding utf8 at Diff class (#12641)

**Revision 11549 - 2013-03-07 11:36 - Toshi MARUYAMA**

fix that diff outputs become ??? in some non ASCII words (#12641)

Contributed by Filou Centrinov.

**Revision 11550 - 2013-03-07 13:53 - Toshi MARUYAMA**

svn propset svn:eol-style native to fixtures (#12641)
Revision 11551 - 2013-03-07 22:03 - Toshi MARUYAMA

Merged r11544, r11545, r11546, r11547, r11549 from trunk to 2.3-stable (#12641)

fix that diff outputs become ??? in some non ASCII words.

Contributed by Filou Centrinov.

Revision 11552 - 2013-03-07 22:07 - Toshi MARUYAMA

2.3-stable: svn propset svn:eol-style native to fixtures (#12641)

History

#1 - 2013-03-05 00:17 - Filou Centrinov
- File unified_diff.rb.diff added

The Problem is, that for example the following diff-lines

- часа
+ часов

are parsed in Redmine as UTF-8 like this:

\xD1\x87\xD0\xB0\xD1\xB0\xD0\xB0\xD1\x81\xD0\x80&lt;span&gt;\xD0\xB0&lt;/span&gt;&amp;quot;
\xD1\x87\xD0\xB0\xD1\x81\xD0\x80&lt;span&gt;\xD0\xBE\xD0\xB2&lt;/span&gt;&amp;quot;

This is wrong, because the leading byte D0 is part of the cyrillic 2-Byte character "а" in the <span>-tag, but it's actually outside of the <span>-tag. Therefore characters will be misinterpreted and will be displayed with "?".

Correct UTF-8 would be:

\xD1\x87\xD0\xB0\xD1\x81\xD0\xB0\xD1\x81\xD0\x80&lt;span&gt;\xD0\xB0&amp;quot;
\xD1\x87\xD0\xB0\xD1\x81\xD0\x80&lt;span&gt;\xD0\xBE\xD0\xB2&lt;/span&gt;&amp;quot;

So we have for the first line "...<span>\xD0\x80</span>..." instead of "...\xD0&lt;span&gt;\xB0&lt;/span&gt;...". The attached patch searches for the last leading byte, if the unmatching byte is a continuation byte (and not a leading byte or a single character byte).

A continuation byte has the binary format 10xxxxxx, so we can determine it with myContinuationByte.ord.between?(128, 191)

This problem occurs always, when the first determined difference between two bytes are continuation bytes. An other example in japanese you find in #13350.

#2 - 2013-03-05 12:56 - Filou Centrinov
- File unified_diff.rb.2.diff added

A much better way to fix this problem is to set an UTF-8 encoding. :-)

2020-04-11
The affected version is also 2.3 (devel)

- Category set to I18n
- Assignee set to Toshi MARUYAMA
- Target version set to 2.4.0

- Target version changed from 2.4.0 to 2.3.0

- Status changed from New to Closed
- Resolution set to Fixed

Committed in, thanks.

Files

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Date</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>diff-r11052.png</td>
<td>20 KB</td>
<td>2012-12-19</td>
<td>Toshi MARUYAMA</td>
</tr>
<tr>
<td>unified_diff.rb.diff</td>
<td>787 Bytes</td>
<td>2013-03-04</td>
<td>Filou Centrinov</td>
</tr>
<tr>
<td>unified_diff.rb.2.diff</td>
<td>621 Bytes</td>
<td>2013-03-05</td>
<td>Filou Centrinov</td>
</tr>
</tbody>
</table>