

Redmine - Defect #12641

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

2012-12-19 09:30 - Toshi MARUYAMA

Status:	Closed	Start date:	
Priority:	Normal	Due date:	
Assignee:	Toshi MARUYAMA	% Done:	0%
Category:	l18n	Estimated time:	0.00 hour
Target version:	2.3.0	Affected version:	2.1.4
Resolution:	Fixed		

Description

An example is [r11052](#) in [#12640#note-2](#).

Revision 11052

View differences: ☒ inline ☐ side by side

trunk/config/locales/ru.yml

115115

one: "около %{count} часа"

116116

few: "около %{count} часов"

117117

many: "около %{count} часов"

118

other: "????? %{count} ?????"

118

other: "????? %{count} ?????"

119119

x\_hours:

120120

one: "1 час"

121121

other: "%{count} часов"

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

```
- ґаґа"
+ ґаґоґв"
```

are parsed in Redmine as UTF-8 like this:

```
\xD1\x87\xD0\xB0\xD1\x81\xD0<span>\xB0</span>&quot;;
\xD1\x87\xD0\xB0\xD1\x81\xD0<span>\xBE\xD0\xB2</span>&quot;;
```

This is wrong, because the *leading byte* \xD0 is part of the cyrillic 2-Byte character "а" in the <span>-tag, but it's actually outside of the <span>-tag. Therefore charaters will be misinterpreted and will be displayed with "?".

Correct UTF-8 would be:

```
\xD1\x87\xD0\xB0\xD1\x81<span>\xD0\xB0</span>&quot;;
\xD1\x87\xD0\xB0\xD1\x81<span>\xD0\xBE\xD0\xB2</span>&quot;;
```

So we have for the first line "...<span>\xD0\xB0</span>..." instead of "...\xD0<span>\xB0</span>...". 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. :-)

**#3 - 2013-03-05 20:23 - Filou Centrinov**

The affected version is also 2.3 (devel)

**#4 - 2013-03-07 02:36 - Toshi MARUYAMA**

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

**#5 - 2013-03-07 09:28 - Toshi MARUYAMA**

- Target version changed from 2.4.0 to 2.3.0

**#6 - 2013-03-08 00:13 - Toshi MARUYAMA**

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

Committed in, thanks.

**Files**

---

diff-r11052.png	20 KB	2012-12-19	Toshi MARUYAMA
-----------------	-------	------------	----------------

unified_diff.rb.diff	787 Bytes	2013-03-04	Filou Centrinov
unified_diff.rb.2.diff	621 Bytes	2013-03-05	Filou Centrinov