Redmine - Feature #2371

character encoding for attachment file

2008-12-22 07:21 - youngseok yi

Status:	Closed	**Start date:**	2008-12-22
Priority:	Low	**Due date:**	
Assignee:	Toshi MARUYAMA	**% Done:**	100%
Category:	Attachments	**Estimated time:**	0.00 hour
Target version:	1.3.0		
Resolution:	Fixed		

Description

As r814, default encoding for repository can be configured.

diff or patch attachment requires similar configuration.

- default encoding for diff or patch attachment (Admin -> Settings -> Attachment -> diff/patch encodings ?).
- follow encoding of repository. (source:/trunk/app/helpers/repositories_helper.rb@1900#L109)

I thinks 2nd option may be enough and useful.

Related issues:

Related to Redmine - Defect #9143: Partial diff comparison should be done on ...	**Closed**	**2011-08-29**
Related to Redmine - Defect #4608: Mail attachment name encoding is incorectl...	**Closed**	**2010-01-19**
Has duplicate Redmine - Feature #4577: convert text file attached an issue to...	**Closed**	**2010-01-14**
Has duplicate Redmine - Defect #3652: Unicode Support for TXT-Files	**Closed**	**2009-07-22**

Associated revisions

Revision 7823 - 2011-11-17 08:00 - Toshi MARUYAMA

attachment: add a functional test to show UTF-8 text file (#2371)

Revision 7824 - 2011-11-17 08:01 - Toshi MARUYAMA

attachment: add a functional test to show invalid UTF-8 text file (#2371)

Stripping invalid UTF-8 is Redmine 1.2 behaviour.

Revision 7825 - 2011-11-17 12:53 - Toshi MARUYAMA

move repositories helper to_utf8 logic to lib/redmine/codeset_util.rb for common use (#2371)

Revision 7828 - 2011-11-18 15:57 - Toshi MARUYAMA

move repositories helper to_utf8 tests to new CodesetUtilTest (#2371)

Revision 7866 - 2011-11-20 12:46 - Toshi MARUYAMA

attachment: use repositories setting to convert contents character encoding (#2371)

This commit results replacing invalid encoding instead to stripping.

Revision 7867 - 2011-11-20 12:47 - Toshi MARUYAMA

attachment: add a functional test to show an ISO-8859-1 patch (#2371)

Revision 7868 - 2011-11-20 12:47 - Toshi MARUYAMA

attachment: add a functional test to show an ISO-8859-1 content file (#2371)

Revision 7869 - 2011-11-20 12:48 - Toshi MARUYAMA

attachment: move repositories encodings setting to the general tab and update the label (#2371)

Revision 7870 - 2011-11-20 12:48 - Toshi MARUYAMA

update Japanese translation of attachments and repositories encodings setting label (#2371)

Revision 7871 - 2011-11-20 13:04 - Toshi MARUYAMA

scm: attachment: remove "to_utf8" methods from helpers (#2371)

It is confusing that same name methods are in several helpers.

Revision 7873 - 2011-11-20 13:17 - Toshi MARUYAMA

attachment: update a functional test to switch "side by side" and "inline" for ISO-8859-1 patches (#2371, #9612)

Revision 7885 - 2011-11-22 01:08 - Toshi MARUYAMA

attachment: add missing diff type at functional tests (#2371, #9612)

History

#1 - 2010-06-27 12:14 - Yuya Nishihara

- File attachment-encoding.patch added

youngseok yi wrote:

> - follow encoding of repository.

Attached patch implements it with minimal changes. attachment-encoding.patch

Proper solution will be something like:

1. move to_utf8 to separate module, e.g. RepoFilesHelper
2. make AttachmentsHelper and RepositoriesHelper include RepoFilesHelper

#2 - 2011-05-30 15:37 - Toshi MARUYAMA

- Assignee set to Toshi MARUYAMA

#3 - 2011-05-30 15:48 - Toshi MARUYAMA

- Target version set to Candidate for next major release

#4 - 2011-06-05 09:23 - Toshi MARUYAMA

- Target version changed from Candidate for next major release to 1.3.0

#5 - 2011-11-17 07:49 - Toshi MARUYAMA

- Subject changed from encoding for diff or patch attachment file to encoding for attachment file

#6 - 2011-11-17 07:50 - Toshi MARUYAMA

- Subject changed from encoding for attachment file to character encoding for attachment file

#7 - 2011-11-17 10:36 - Etienne Massip

Toshi, won't your last commit prevent me from attaching an iso8859-1 encoded patch to this issue and seeing it fine?

#8 - 2011-11-17 11:17 - Toshi MARUYAMA

- File general-settings.png added

Etienne Massip wrote:

> Toshi, won't your last commit prevent me from attaching an iso8859-1 encoded patch to this issue and seeing it fine?

This feature issue goal is that attachment **file** and **patch** encoding are converted by repositories setting.

general-settings.png

#9 - 2011-11-17 13:33 - Etienne Massip

I'm not sure this is a good idea; repositories may return data using a specific encoding, but attachments are usually stored on FS without transformation, so assuming that they're "very likely to be encoded the same way data in SCM is" is not necessarily true.

For example, my encoding list starts with UTF-8 and my locale (Fr) would assume that files uploaded by users are probably encoded in ISO-8859-15/CP1252; so assuming that the text files uploaded are in UTF-8 mean that they will be rendered stripped and that I will probably often loose some chars, which is the actual situation.

I would prefer to be able to specify a distinct default encoding for text attachments which would be ISO-8859-15/CP1252 (could be defaulted to default server encoding) and render with something like bom_present?(str) ? str : Iconv.conv('UTF-8', Setting.default_encoding).

#10 - 2011-11-17 14:05 - Toshi MARUYAMA

UTF-8 is very strict.
It is very rare case that miss understanding ISO-8859-1 characters as UTF-8.
http://groups.google.com/group/thg-dev/browse_thread/thread/6c258628e3fce8/09e9dbe4a030e51d

#11 - 2011-11-17 14:14 - Toshi MARUYAMA

Redmine 1.2.2 repository converting encoding is this line.
source:tags/1.2.2/app/helpers/repositories_helper.rb#L140

In case of "UTF-8,ISO-8859-1",
if converting error in "UTF-8", Redmine converts from ISO-8859-1.

Japanese use three encoding, UTF-8, EUC-JP and Shift-JIS (CP932).
This Redmine feature is big advantage in Japan.

#12 - 2011-11-17 14:20 - Etienne Massip

So if I understand well, according to encoding list order, it will try and fail to convert the ISO-8859-1 file from UTF-8 to UTF-8 and then will try and success to convert it from ISO-8859-1 to UTF-8?

Guess it will work...

#13 - 2011-11-17 14:23 - Etienne Massip

What if the administrator does not set UTF-8 at the start of the list?
Can't you str.is_utf8? ? str : try Iconv.conv('UTF-8', Setting.encodings)?

#14 - 2011-11-17 14:25 - Toshi MARUYAMA

Etienne Massip wrote:

> repositories may return data using a specific encoding,

It is not true.
SCMs does not have encoding information (meta data) of **file contents**.
http://mercurial.selenic.com/wiki/EncodingStrategy?action=recall&rev=21#Unknown_byte_strings

#15 - 2011-11-17 14:28 - Etienne Massip

Toshi MARUYAMA wrote:

> It is not true.
> SCMs does not have encoding information (meta data) of **file contents**.

Well, that's why I said *may* :-)

#16 - 2011-11-17 15:39 - Toshi MARUYAMA

Etienne Massip wrote:

> What if the administrator does not set UTF-8 at the start of the list?

This is very rare case in Japan.
It is popular "UTF-8,EUC-JP,Shift_JIS in Japan.
This order is strict order.

If Single Byte Character Set (e.g. ISO-8859-1) is the start of the list, all characters are converted to UTF-8.
But, I think this is very rare case in the whole world.

> Can't you str.is_utf8? ? str : try Iconv.conv('UTF-8', Setting.encodings)?

Default repository encoding setting is **empty**.

This is equivalent that default is UTF-8.
And I think it is better that administrator set UTF-8 in the start of the list explicitly.

#17 - 2011-11-20 13:19 - Toshi MARUYAMA

- % Done changed from 0 to 100

#18 - 2011-11-24 12:20 - Anton Statutov

Is this feature fixes #4608?

#19 - 2011-11-24 21:36 - Mischa The Evil

Anton Statutov wrote:

> Is this feature fixes #4608?

I don't think so.

#20 - 2011-11-30 00:09 - Toshi MARUYAMA

- Status changed from New to Closed

- Resolution set to Fixed

Committed in r7885.

Files

attachment-encoding.patch	1017 Bytes	2010-06-27	Yuya Nishihara
general-settings.png	28.2 KB	2011-11-17	Toshi MARUYAMA