Redmine - Defect #41464

CSV file encoding auto-detection may fail with multibyte characters

2024-10-10 09:30 - Go MAEDA

Status:	Closed	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	Go MAEDA	**% Done:**	0%
Category:	Importers	**Estimated time:**	0.00 hour
Target version:	6.0.0		
Resolution:	Fixed	**Affected version:**	

Description

When importing a CSV file, Redmine attempts to auto-detect the file's encoding (see #34718). However, this auto-detection may fail when the file contains multibyte characters.

This is because the Redmine::CodesetUtil.guess_encoding method checks the first 256 bytes of the file without considering character boundaries. As a result, an incomplete multibyte character will be present at the end of the data. For example, the Japanese Hiragana character "" in UTF-8 encoding consists of three bytes, "\xE3\x81\x82", but it may appear at the end as "\xE3\x81". When this occurs, the guess_encoding method fails to identify the encoding, as String#valid_encoding? in the method returns false due to the presence of a partial character.

The attached patch fixes this issue by truncating the data at the last line break to discard the last line that may contain such an incomplete multibyte character. This approach ensures that characters causing String#valid_encoding? to return false are excluded.

Associated revisions

Revision 23150 - 2024-10-20 08:47 - Go MAEDA

Fix CSV import file encoding auto-detection failure with multibyte characters (#41464).

Patch by Go MAEDA (user:maeda).

History

#1 - 2024-10-19 09:10 - Go MAEDA

- File 0001-Fix-CSV-file-encoding-auto-detection-failure-with-mu.patch added

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

I have updated the patch.

This version moved the logic to discard the last line from lib/redmine/codeset_util.rb to app/models/import.rb because I think data cleanup should be done within the method that reads the data from a file.

#2 - 2024-10-20 08:48 - Go MAEDA

- Status changed from New to Closed

- Assignee set to Go MAEDA

- Target version changed from Candidate for next major release to 6.0.0

- Resolution set to Fixed

Committed the fix in r23150.

Files

fix-guess_encoding.patch	2.38 KB	2024-10-10	Go MAEDA
0001-Fix-CSV-file-encoding-auto-detection-failure-with-mu.patch	7.34 KB	2024-10-19	Go MAEDA