QGIS Application - Bug report #4547
Garbled Japanese characters in GRASS plugin

2011-11-17 05:30 AM - Masaru Narazaki Narazaki

Status:	Closed		
Priority:	Normal		
Assignee:	Giuseppe Sucameli		
Category:	GRASS		
Affected QGIS version:	master	**Regression?:**	No
Operating System:		**Easy fix?:**	No
Pull Request or Patch supplied:	Yes	**Resolution:**	
Crashes QGIS or corrupts data:	No	**Copied to github as #:**	14461

Description

In Japan if we try to use GRASS plugin with Japanese, we can not find correct japanese leter because of the Garbring as adding files.
They say this phenomina begun at version 1.0 of QGIS.
Please collect this phenomina.

Associated revisions

Revision c53c8581 - 2012-11-04 07:00 PM - Giuseppe Sucameli

grass plugin: avoid garbled japanese/cyrillic chars in the tools' GUI (fix #4547, #3164)

Thanks Minoru Akagi for patches!

History

#1 - 2011-11-17 05:35 AM - Giovanni Manghi

- Target version changed from Version 1.6.0 to Version 1.8.0

- Subject changed from Garbring Japanes character in GRASS plugin to Garbled Japanese characters in GRASS plugin

#2 - 2012-05-12 04:39 AM - Alexander Bruy

- Affected QGIS version set to master

- Crashes QGIS or corrupts data set to No

Looks like duplicate of #3164 (same issue for cyrillic)

#3 - 2012-09-04 11:55 AM - Paolo Cavallini

- Target version changed from Version 1.8.0 to Version 2.0.0

#4 - 2012-10-25 05:52 PM - Paolo Cavallini

So, this turned out a practically unsolvable problem in GRASS. Quoting Glynn Clements:

===
There are two issues for which there is no viable solution:

1. OEM encoding.
2. Shift-JIS.

Regarding #1: GRASS neither knows nor cares whether a string is in ANSI or OEM encoding. Much of it doesn't care about encodings at all, and just treats strings as sequences of bytes. Anything which needs to care about the encoding (e.g. the GUI) will just use "the locale's encoding", which on Windows means "the ANSI codepage". If you use the OEM codepage for anything, you lose.

Suggestions as to how to determine whether a string uses the ANSI or OEM page are welcome, if unlikely.

Regarding #2: On Windows, any byte within the range 0-127 is assumed to represent the corresponding ASCII character. For encodings which assign other characters to any byte within that range (either individually or as part of a multi-byte sequence), that is likely to cause problems.

The most obvious example is that any occurrence of the byte 0x5C within a filename is assumed to be a directory separator. Unfortunately, Shift-JIS uses 0x5C as the second byte of a multi-byte sequence, meaning that Japanese filenames may be parsed incorrectly.

Neither EUC-JP nor UTF-8 have this problem (as these only re-purpose codes above 128), but unfortunately Windows doesn't provide locales which uses either of these encodings.

And I can't think of any solution which doesn't involve re-writing all code which handles pathnames.

Similar issues may exist with the other punctuation characters which are "mingled" with the alphabetic characters, i.e. "[\\]^_{|}~" (e.g. | is commonly used as a field separator, so tabular data which includes Japanese text may be parsed incorrectly).

While such cases are probably less common than the pathname issue, a fix is even less viable (i.e. fixing all string-handling code).

-- Glynn Clements <glynn@gclements.plus.com>===

So the solution seems just to switch to EN, just for Windows.
Seems an easy fix.

#5 - 2012-10-25 11:46 PM - Minoru Akagi

- File grassplugin1.patch added

In Japanese Windows environment, GRASS commands output xml text of interface description that begins with the following line.

<?xml version="1.0" encoding="CP932"?>

QDomDocument has ability to detect encoding, but it doesn't recognize most of codepage name "CPxxx". See
http://qt-project.org/doc/qt-4.8/QTextCodec.html

I think it's not better to rely the current encoding conversion ability of QDomDocument. Since GRASS commands usually output text in system default encoding, we maybe should treat encoding name that Qt doesn't recognize as system encoding.

#6 - 2012-10-25 11:47 PM - Paolo Cavallini

- Pull Request or Patch supplied changed from No to Yes

#7 - 2012-10-26 12:53 AM - Marco Hugentobler

- Assignee set to Radim Blazek

#8 - 2012-11-01 05:55 PM - Paolo Cavallini

May be a duplicate of #3164. Please close it if this is the case.

#9 - 2012-11-01 06:42 PM - Minoru Akagi

- File grassplugin2.patch added

Okay, I attach a patch including patch for #3164 anew.

#10 - 2012-11-03 06:38 PM - Giuseppe Sucameli

Hi Minoru,
the patch looks good to me.

I'm adding a check so if we are not able to get the encoding from the XML declaration (using utf8 and the regular expression) then we'll let Qt detects the encoding of the XML (current behaviour).

This will make it working even whether the encoding name is not found, e.g. the encoding attribute is missing (though we are quite sure GRASS won't remove it) or the XML content is a UTF-16 or UTF-32 encoded string (the regexp doesn't match the text).

Since I cannot test it with Japanese lang, please, could you try the branch grass_jp_enc from my repo and report if it works?

#11 - 2012-11-03 08:52 PM - Minoru Akagi

Giuseppe Sucameli wrote:

> *Since I cannot test it with Japanese lang, please, could you try the branch grass_jp_enc from my repo and report if it works?*

I've just tested your branch and got good result. Thanks!

#12 - 2012-11-04 10:05 AM - Giuseppe Sucameli

- Status changed from Open to Closed

Fixed in changeset commit:"c53c85813f4723b75d4e9326d2565fb51eaa8355".

#13 - 2012-11-04 10:15 AM - Giuseppe Sucameli

- Assignee changed from Radim Blazek to Giuseppe Sucameli

Thanks Minoru Akagi!

I hope we haven't broken other languages :)

Now that the change is in master, please could other people test it and report here?

Files

Garbring_character.JPG	46.1 KB	2011-11-17	Masaru Narazaki Narazaki
grassplugin1.patch	1.33 KB	2012-10-25	Minoru Akagi
grassplugin2.patch	2.56 KB	2012-11-01	Minoru Akagi