QGIS Application - Bug report #6500
Language Encoding very broken in 1.8 Lisboa

2012-10-11 02:34 PM - Lex Berman

Status:	Closed		
Priority:	Normal		
Assignee:			
Category:			
Affected QGIS version:	master	**Regression?:**	No
Operating System:	all	**Easy fix?:**	No
Pull Request or Patch supplied:	No	**Resolution:**	
Crashes QGIS or corrupts data:	No	**Copied to github as #:**	15720

Description

Shapefiles that have known charset encodings cannot be viewed or managed properly since 1.7.4 and the problem persists in 1.8.

System: Windows 7, Mac Snow Leopard, Mac Lion
Versions: QGIS 1.7.4 - 1.8
Data format: SHP

Last known working version QGIS 1.7.3

I know that there are many issues opened and closed on the topic, but my PRIMARY use of QGIS is to manage datasets in various encodings, because ESRI products simply cannot accomplish this task. I am a huge QGIS fan, and teach QGIS at the GIS center where I work. http://maps.cga.harvard.edu/qgis/

The main argument I used to convince the center to offer QGIS instruction, (instead of ArcMap), is the ability of QGIS to handle various encodings so well.

Attached images show how great QGIS 1.7.3 is, able to read and edit Chinese in GBK or UTF-8, as well as Russian and Uighur scripts in UTF-8 all on the same screen, and works flawlessly!

Unfortunately the identical layers in QGIS 1.8 always fail to view the non-ascii scripts, every time.

Some test cases:

1) .csv can be opened in UTF8 without a problem. QGIS UI handles everything perfectly. .csv can be joined to SHP and the attribute table is still working fine in QGIS. however, as soon as the SHP is saved out, the non-ascii contents are completely corrupted.

2) .kml can be opened in UTF8 and works fine. as in previous case, saving out to SHP ruins the contents.

Whether this is an LDID problem, or a GDAL problem, or something else, the issues I have read here are all closed and no solution works in my case.

Ideally, whatever worked fine up to QGIS 1.7.3 can be restored? We just want the declared encoding at the time of the layer being opened in QGIS to be actually respected. This must hold true at the time the layer is opened, as well as when it is saved. Something has broken in the way the encoding is interpreted by QGIS when using SHP. Is there some way that the encoding declaration from the QGIS UI can over-ride the system default AND the LDID, and then over-write the LDID if the file is saved in that encoding? Otherwise I can no longer make sense out of QGIS behaviour in version 1.8, which seems to disrespect the encoding as declared and also corrupt it whenever it is saved.

Can somebody please revisit this and finally fix it?

Related issues:		
Duplicates QGIS Application - Feature request # 6216: User should be able to ...	**Closed**	**2012-08-17**
Duplicates QGIS Application - Bug report # 5911: Language Driver ID in dbf fi...	**Closed**	**2012-06-30**

History

#1 - 2012-10-11 02:36 PM - Jürgen Fischer

- Subject changed from Langauge Encoding very broken in 1.8 Lisboa to Language Encoding very broken in 1.8 Lisboa

#2 - 2012-10-11 11:46 PM - Alexander Bruy

- Status changed from Open to Feedback

- Operating System changed from Windows to all

- OS version deleted (7)

This already fixed in master, see commit:75dc85b4d652116814873bb7674cab15ce6cde66. So use master or, if you want 1.8 — try NextGIS build

#3 - 2012-10-12 01:21 AM - Borys Jurgiel

What about adding "autodetect" to the encoding combobox rather than the present solution with the "ignore" checkbox?

#4 - 2012-10-12 02:58 AM - Minoru Akagi

Disabling the encoding conversion of Shapefile layer is the easiest way to avoid garbling text in shapefiles outputted from QGIS. It will recover previous behavior about encoding.

If you have installed QGIS 1.8 with OSGeo4W, you can disable the encoding conversion of Shapefile layer by adding the following line to qgis.bat.

SET SHAPE_ENCODING=DISABLED

"DISABLED" is charset name that does not exist. This solution needs newer gdal19.dll, which is installed by OSGeo4W. That of Standalone Installer (version 1.8) is old to do it.

#5 - 2012-10-12 03:43 AM - Alexander Bruy

Minoru Akagi wrote:

> *This solution needs newer gdal19.dll, which is installed by OSGeo4W. That of Standalone Installer (version 1.8) is old to do it.*

This is not correct. Standalone installer also used GDAL 1.9.0

#6 - 2012-10-12 04:33 AM - Minoru Akagi

Alexander Bruy wrote:

> *This is not correct. Standalone installer also used GDAL 1.9.0*

It's good to set empty string to SHAPE_ENCODING to avoid any encoding conversion of Shapefile layer, but it seems to be impossible to set empty string

to an environment variable in Windows. Though the solution I have suggested is made possible by GDAL changeset 24554, the creation date of gdal19.dll in the current Standalone Installer we can get from https://issues.qgis.org/projects/quantum-gis/wiki/Download is 2012/05/17.

#7 - 2012-10-12 08:20 AM - Lex Berman

- File QGIS_bat_fix_not_working.jpg added

Thanks so much for you replies!

I have installed: QGIS-OSGeo4W-1.8.0-1-Setup.exe

Now I edited gqis.bat like this:
@echo off

SET OSGEO4W_ROOT=C:\\PROGRA~2\\QUANTU~2
SET SHAPE_ENCODING=DISABLED
call "%OSGEO4W_ROOT%"\\bin\\o4w_env.bat
....

however, the encoding functions are still failing completely...

Apparently this is the standalone, running with GDAL 1.9, found here:
C:\\Program Files (x86)\\Quantum GIS Lisboa\\bin\\gdalplugins\\1.9

I also tried setting the path explicitly in the GDAL tools settings to
C:\\Program Files (x86)\\Quantum GIS Lisboa\\bin
But they had no effect...

Still hoping there is some way to view and edit multibyte encodings!

thanks gain...

#8 - 2012-10-22 01:42 AM - Mathieu Pellerin - nIRV

I've just ran into this problem with Lao users complaining that "while 1.7.4 supported input Lao scripts strings, qgis 1.8 broke it, doesn't support Lao". That is, user creating a new shapefile via QGIS will loose unicode string data entered upon saving work. Since the user isn't given a choice for shapefile character encoding, the assumption is that it'll be utf-8. (as it worked with previous versions).

While there might be nothing much that can be done for 1.8, can the upcoming version 2.0 fix this issue? Assuming it's a GDAL issue (even for the case scenarios when a shapefile is created through the QGIS interface), is there an GDAL issue filed against this problem?

Alexander, you pointed out to a recently added [ ] ignore shapefile encoding option. That's a good workaround but I'm not entirely sure this will be read by the user as the way to fix newly created shapefiles not saving unicode data.

#9 - 2012-10-22 03:48 AM - Minoru Akagi

Garbled characters in shapefiles outputted by QGIS are caused by encoding conversion of Shapefile layer, which converts encoding of passed characters from UTF-8 to ISO-8859-1 by default. Thinking that Shapefile layer should not convert character encoding if the user doesn't specify the ENCODING layer creation option, I submitted a ticket. See https://trac.osgeo.org/gdal/ticket/4808

#10 - 2012-10-22 07:30 PM - Mathieu Pellerin - nIRV

In the region I'm based, I confirm that, under QGIS >= 1.8, Thai, Lao, Burmese, and Khmer script strings are lost when editing newly created or already existing shapefiles. Googling around also shows Japanese language being affected (http://osgeo-org.1560.n6.nabble.com/QGIS-version-1-8-td5006318.html).

Basically, any non-latin language is broken when editing shapefiles (newly created or otherwise) under QGIS >= 1.8 with default settings. That's a pretty big regression that affects a large part of the world.

Proper Unicode support has always been a good argument when getting people to migrate from ArcGIS to QGIS. This regression should also be viewed considering this :)

IMO, until the GDAL shapefile encoding detection feature works properly, the "[ ] Ignore shapefile encoding" option should be on *by default*. Having a default setting that leads to non-latin data lost should not be acceptable behavior.

#11 - 2012-10-23 09:41 PM - Mathieu Pellerin - nIRV

Russian also affected: http://gis.stackexchange.com/questions/37342/shape-file-encoding-problem-in-qgis-1-9-0-built-with-gdal-1-9-2

What would be the downside of setting QGIS to ignore GDAL's shapefile encoding by default?

#12 - 2012-10-23 11:53 PM - Minoru Akagi

nirvn - wrote:

> *What would be the downside of setting QGIS to ignore GDAL's shapefile encoding by default?*

Probably nothing. I also think it is preferable to disable encoding conversion of Shapefile layer in the present situation, for example, the handling of LDID/87 and OGR interface about encoding. All the non-ASCII characters using users will need the option, so it should be checked by default.

#13 - 2012-10-24 02:00 AM - Mathieu Pellerin - nIRV

Minoru Akagi wrote:

> *All the non-ASCII characters using users will need the option, so it should be checked by default.*

All users using non-ASCII characters currently need to ignore GDAL shapefile encoding feature, and all users using ASCII characters will not be affected by turning this option on by default.

That's a win win solution :)

#14 - 2013-01-08 02:12 AM - marisn -

Welcome to QGIS/OGR broken Shapefile encoding club! The first rule of club is to not talk about club!

Man ones: #5255 #5911
Others: #6327 #5982 #5927 #5900 #5508 #5340 #4343

Correct me, if I missed some.

#15 - 2013-02-16 02:28 AM - Jürgen Fischer

- Status changed from Feedback to Closed

Fixed in changeset commit:"7fb46498c9fb3c14a2d0b0fcc8e634dba2f1cade".

Files

qgis_1-7-3_wroclaw_utf_works.jpg	382 KB	2012-10-11	Lex Berman
qgis_1-8_lisboa_utf_broken.jpg	378 KB	2012-10-11	Lex Berman
QGIS_bat_fix_not_working.jpg	252 KB	2012-10-12	Lex Berman