

QGIS Application - Bug report #19639

CSV: "Detect field types" doesn't update the sample view

2018-08-17 08:17 AM - Tobias Wendorff

Status:	Closed	
Priority:	Normal	
Assignee:		
Category:	Data Provider/Delimited Text	
Affected QGIS version:	3.2.1	Regression?: No
Operating System:	Microsoft Windows 7, 64-bit	Easy fix?: No
Pull Request or Patch supplied:	No	Resolution: no timely feedback
Crashes QGIS or corrupts data:	No	Copied to github as #: 27466
Description		
<p>When selecting "Detect field types", the sample view doesn't change. It's expected to have a preview, of how QGIS will modify the content (since it's still very buggy, a preview is important).</p> <p>All newer versions are affected (at least >= 3.1).</p>		

History

#1 - 2018-08-19 11:16 AM - Giovanni Manghi

- Status changed from Open to Feedback

What is "very buggy"? Do you mean in general or compared to 2.18/LTR?

#2 - 2018-08-19 11:31 AM - Tobias Wendorff

Giovanni Manghi wrote:

What is "very buggy"? Do you mean in general or compared to 2.18/LTR?

In general. I've filed a bug some months ago. It partially go fixed, but "Detect field types" is still broken. Numbers like "04595" still get parsed into "4595", which creates corrupted data (please check, how OGR does it... it's working perfect) - but that's not part of this ticket.

Since the preview of "Detect field types" doesn't work, you can only see the corrupted data in the attribute table. Some guys have very big CSV files, so it's hard for them to find the corruption at all.

#3 - 2018-08-20 11:00 AM - Giovanni Manghi

Tobias Wendorff wrote:

(please check, how OGR does it... it's working perfect)

for example when translating a CSV to a shapefile with ogr2ogr?

#4 - 2018-08-20 08:13 PM - Tobias Wendorff

Giovanni Manghi wrote:

Tobias Wendorff wrote:

(please check, how OGR does it... it's working perfect)

for example when translating a CSV to a shapefile with ogr2ogr?

Yes, like this:

```
ogr2ogr -overwrite --config PG_USE_COPY YES PG:"host=127.0.0.1 port=xxxx dbname=xxxx user=xxxx" "xxxx.csv" -oo HEADERS=YES -oo AUTODETECT_SIZE_LIMIT=0 -oo AUTODETECT_TYPE=YES -oo AUTODETECT_WIDTH=YES -oo X_POSSIBLE_NAMES=lon* -oo Y_POSSIBLE_NAMES=lat* -oo KEEP_GEOM_COLUMNS=NO -a_srs EPSG:4326 -nlt point -nln xxxx -lco GEOMETRY_NAME=geom
```

"AUTODETECT_SIZE_LIMIT=0" means: scan the whole file (data gets loaded into a buffer instead reading from STDIN), normally it's 100,000 rows (which is too low on some of my datasets). Importing data into "R" works similar, it's another workaround.

CSVT works inside of QGIS, BUT you can't make QGIS use the CSV's header... I think, when loading a CSV with CSVT, OGR gets used. But you can't tell it to use the first line as a header :-(

#5 - 2018-08-21 08:30 PM - Giovanni Manghi

- File *TM_WORLD_BORDERS-0.csv* added

CSVT works inside of QGIS, BUT you can't make QGIS use the CSV's header... I think, when loading a CSV with CSVT, OGR gets used. But you can't tell it to use the first line as a header :-(

I just loaded the attached CSV in QGIS (using the 'add vector layer' dialog) and the first line **was** indeed used as header.

#6 - 2018-09-10 07:45 PM - Tobias Wendorff

- Status changed from Feedback to Open

Giovanni Manghi wrote:

*I just loaded the attached CSV in QGIS (using the 'add vector layer' dialog) and the first line **was** indeed used as header.*

Nah, I was talking about CSVT. When opening a CSV, which has a CSVT, the header line of the CSV is loaded as a data line. It can't be disabled.

After all, the reported bug is still open. Please have a look, how OGR did it. It's a pretty simple, but effective logic. **Right now, the function is broken and should be disabled.**

#7 - 2018-09-11 02:11 PM - Giovanni Manghi

- Status changed from Open to Feedback

Tobias Wendorff wrote:

Giovanni Manghi wrote:

I just loaded the attached CSV in QGIS (using the 'add vector layer' dialog) and the first line **was** indeed used as header.

Nah, I was talking about CSVT.

the title nor the description talks about CSVT files, can you please help clarify? If there are different issues here they must be filed in separate tickets.

#8 - 2018-09-11 02:31 PM - Giovanni Manghi

Tobias Wendorff wrote:

Giovanni Manghi wrote:

I just loaded the attached CSV in QGIS (using the 'add vector layer' dialog) and the first line **was** indeed used as header.

Nah, I was talking about CSVT. When opening a CSV, which has a CSVT, the header line of the CSV is loaded as a data line. It can't be disabled.

just tried, both loading the CSV as a table or as a point layer (using the delimited text provider). In the latter case the CSVT is not used, I think this is expected.

After all, the reported bug is still open. Please have a look, how OGR did it. It's a pretty simple, but effective logic. **Right now, the function is broken and should be disabled.**

In my case the fields types were detected correctly (using master), could you please attach sample data, thanks.

#9 - 2018-09-15 12:59 AM - Tobias Wendorff

Yay, it really works for CSVT now, but normal CSV files still get bad results.

first.csvt String(255),Real,String(255),Real

first.csv zipcode;number_science;number_comma;number_point
01234578;3.333333333333333E-01;1,234567890;1.23456789

second.csv zipcode;number_science;number_comma;number_point
01234578;3.333333333333333E-01;1,234567890;1.23456789

Good work on **first.csv** - works as expected now. Good work! **second.csv** reads all fields as text when field detection is disabled; this is fine. But when it's enabled, field **zipcode** gets integer again. The leading zero shouldn't be dropped. Like stated above, OGR has a nice way to figure out the value's real type (R works similar): it scans the fields and stops when the transformed value is different from the original one. **012345678** fits into string/text only, so the field can't be INT.

#10 - 2018-09-15 01:00 AM - Tobias Wendorff

- Status changed from Feedback to Open

Whoops, forgot to open it again.

#11 - 2018-09-15 03:46 PM - Giovanni Manghi

- Status changed from Open to Feedback

second.csv reads all fields as text when field detection is disabled; this is fine. But when it's enabled, field **zipcode** gets integer again. The leading zero shouldn't be dropped. Like stated above, OGR has a nice way to figure out the value's real type (R works similar): it scans the fields and stops when the transformed value is different from the original one. **012345678** fits into string/text only, so the field can't be INT.

this is a **different** issue from the one in the description/subject of this title and should be reported in a separated ticket(?).

#12 - 2019-02-23 08:37 PM - Jürgen Fischer

- Resolution set to no timely feedback

- Status changed from Feedback to Closed

Bulk closing 82 tickets in feedback state for more than 90 days affecting an old version. Feel free to reopen if it still applies to a current version and you have more information that clarify the issue.

Files

TM_WORLD_BORDERS-0.csv	14 KB	2018-08-21	Giovanni Manghi
------------------------	-------	------------	-----------------