CSV Driver (auto) for OpenJUMP

Since version 1.6, OpenJUMP PLUS is bundled with a new CSV Driver which replaces the old txt-driver. CSV driver 0.9 is the 3rd major revision of this new driver and makes it possible to read csv files included in a compressed archive (read-only). It is available on Jump-Pilot-Project and on http://geo.michaelm.free.fr/
In this documentation, we'll describe the CSV Driver (auto) which is a 0-option implementation which tries to guess what the structure of the csv file is. A CSV file without geometry or with a geometry which cannot be recognized will be imported with empty GeometryCollection's in the Geometry attribute.
Version 2.0.0 of the driver, released on 2021, april, is the first version compatible with OpenJUMP 2 and JTS 1.18.

Goal of CSV Driver (auto)

The goal of this CSV Driver implementation is to make things easy for the user. Just choose the file, the driver will guess how to parse it.

Limitations

Drawbacks of a 100% automatic driver are of two kinds :

- some options are disabled (ex. the user cannot choose encoding)
- the parser will import empty GeometryCollections as geometries if geometry column cannot be recognized automatically. To recognize geometry columns, they must have current names such as X, Y, Z, latitude, longitude, altitude, WKT or GEOMETRY (most often, it means that a header line with column names must be there).

How parameters are guessed

Encoding

~~The driver try to guess the file encoding~~. This capability has been removed after it has been proved more harmful than useful. CSV_auto now uses the default platform encoding.

Comment lines

The parser check first line of the file. If the line starts with #, all lines starting with # are considered as comment lines. It does the same for //, - -, and $.
Note 1 : # is not considered as a comment line if it is immediately followed by FID or X to keep compatibility with previous parser.
Note 2 : if three consecutive line start with $, file is considered to follow the pirol format

Field Separator

After comment lines, the parser takes a maximum of 5 lines and tries each of the following separators : **tab**, **comma**, **semi-column**, **pipe** and **whitespace**. It will set the separator which always split lines into the same number of tokens, and if several separators fulfill this requirement, it will keep the one producing the maximum number of tokens.
Column names
The file may have a header line with column name or not. The parser will test the first line after comment lines and check if it contains pure numbers (a record containing geometric data should contain at least one number).
If the first line contains pure numbers, it is considered as a data line, and the csv file columns will be called col1, col2, col3...
If the first line contains no pure number, it is considered as a header line containing column names.

Data types

It is not common in csv files to have a line containing data types, but pirol csv format uses that, as well as the former version of this driver.
The parser read the first line after the header line. If all fields in this line match one of "string", "double", "integer", "date", "geometry", "wkt", it is considered as a line describing data types. Otherwise, all column are considered as strings.

Geometry columns

A very important part of the work is to localize columns containing the geometry. By default, the parser will try to parse first column as a WKT geometry. If columns have names, it will recognize X, Y[, Z] or LONGITUDE, LATITUDE[, ALTITUDE], or EASTING, NORTHING labels for puntal geometries, and GEOMETRY or WKT labels for any geometry.
Axis order can be any of X/Y or Y/X, matches are case insensitive and some abreviations will also be recognized.
Note that a CSV file without recognized geometry column will be imported in a layer with a GEOMETRY attribute filled with empty GeometryCollections.