opentsv prevents the corruption of scientific data by Excel

Peter De Rijk[1], Svenn D'Hert[1], Mojca Strazisar[1]

[1] Neuromics Support Facility, Center for Molecular Neurology, VIB - University of Antwerp, Antwerp, Belgium

Abstract

Summary

Researchers commonly use spreadsheet software Microsoft Excel to view, edit and format tab- or comma-separated data files, for example to prepare supplementary data for a paper. When loading such a file Excel converts values that it interprets as dates or numbers into an internal format, sometimes corrupting the data in this process by e.g. changing some gene names to dates. Unfortunately these overzealous "smart" features cannot be turned off and have been causing errors in a substantial number of published supplementary data files.
Opentsv was made to effectively circumvent this problem at the root. It provides an easy and transparent way to open delimited data files in Excel without these conversions.

Availability and Implementation

Opentsv is freely available as a tclkit based single file executable for Windows at https://github.com/derijkp/opentsv. The source code is available under the MIT license.

Introduction

Bioinformatics analysis software often produces results in tab-separated value (tsv) or comma-separated value (csv) format. For viewing, annotating, editing or formatting these data files, biologists commonly use Microsoft Excel (Microsoft Corp., Redmond, WA, USA) because it provides an easy to use interface to tabular data they are already familiar with. Also, for many journals it is the only accepted file type that can contain tabular data with formatting.

However, already more than a decade ago researchers found a serious flaw in Excel for this type of use (Zeeberg et al. 2004); When developing their software for data mining published gene expression data sets (Zeeberg et al. 2003) they kept coming across unknown gene names: They found that these were caused by the erroneous conversions by Excel: Gene names that look like dates were transformed to an internal date format when opening tsv or csv

data files. Other identifiers such as the RIKEN "2310009E13" were also found to be converted irreversibly to the floating-point number "2.31E+13".

Similarly, in non-biological fields people are reporting conversion errors. Leading zeros (e.g. in zip-codes and telephone numbers) or trailing zeros (e.g. in version numbers such as 1.10) can be lost in a conversion to a number. Any identifier consisting of numbers and hyphens or dots (e.g. size ranges, product codes, hyphenated room numbers, storage location codes) risk conversion to a date. Ironically, not even dates are completely safe: An unambiguous iso 8601 conformant date such as 2018-02-04 is converted to an ambiguous, locale dependent format (2-4-2018 or 4-2-2018) upon export (http://ecologybits.com/index.php/2016/07/06/beware-this-scary-thing-excel-can-do-to-your-data/). A date where only the month and year are indicated (e.g. april 2001) will be changed to apr-01, which can, after export, be easily confused with the first of April instead.

Zeeberg et al (2004) already expressed the hope that these problems would be solved by making the automatic conversions a non-default option in Excel. However, after more than a decade, there still is no such option.
A recent screening of supplementary files accompanying articles and Excel files deposited in the NCBI Gene Expression Omnibus showed that gene name errors continue to be a problem in supplementary files (Ziemann et al. 2016): In different journals tested, from 5 to over 30% of papers with supplementary gene lists in Excel files have converted gene symbols; 39.7% of Excel files deposited to NCBI GEO were found to contain gene name errors.

Recently some tools were developed to mitigate the problem: One converts the delimited file into an escaped format before loading in Excel (Welsh et al. 2017), while the other detects and fixes misconverted gene symbols (Mallona et al. 2017) to reduce the damage after the fact. We wanted a transparent fix for the problem at the root avoiding (mis)conversions all together.

Opentsv

Opentsv was created to provide a simple solution to the problem by giving the desired option to users. It is a small program that supports, by simply double clicking, opening tsv and csv files in Excel without automatic conversions. Data files loaded via opentsv are by default (after confirmation) saved as tsv files. This is useful in situations where tsv files can be downloaded, edited and uploaded to a server for bulk changes of tabular data. With opentsv, Excel can be safely used for editing these types of files.

Opentsv is available as a portable executable that can be used with or without installation. Running the opentsv executable directly displays a dialog with some help on the software and several options, including installation (Figure 1). The Install button will copy the executable into the program files folder and link it to the appropriate file types. Alternatively, the "Register opentsv as default program for opening" button can be used to link file types to the program at its current location, allowing its easy use without needing administrator access to the program

files folder. Opentsv can be linked to other file types as well using the default Windows methods.
When linked to the file types, double clicking will open the file using opentsv, which subsequently opens Excel and uses the Excel programmatic interface to open the file without the erroneous conversions.

The "Open file" button will allow the user to select and open a file in Excel directly from opentsv using the current settings, which are shown below it. The following options for handling values are available:

- The "**All as text**" option will cause all fields to be loaded as text fields. In this case no conversion happens at all.
- Using "**Only convert numerical columns**", opentsv parses the first 10 data lines (ignoring potential comments and header line). Columns that have a numeric value in all 10 lines, will be imported as generic, allowing conversion to numeric values. As this may be useful for calculations in Excel, and minimizes the chance for incorrect conversions, this method is the default setting.
- "**Convert everything**" uses the generic conversion for all fields.
- Finally "**Excel default**" can be used to return (temporarily) to Excel opening data files without intervention of opentsv.

Another Excel problem tackled by opentsv is the use of separators: In some locales (where the comma is used as a decimal separator) Excel diverges from the general csv format specification (https://tools.ietf.org/html/rfc4180) by using semicolons as separator instead of commas. Because of this, a user in such locale cannot easily open a properly formatted csv file (using commas) produced by software or international colleagues: Opentsv supports several options for the use of separator:

- By default (**Autodetect**), it will autodetect the separator from the data except for files with the extension csv and tsv, where respectively the comma and the tab character are used.
- With the "**Allways autodetect**" option, the separator is also autodetected for csv and tsv files.
- The separator can be manually set to **comma**, **tab**, **semicolon** or **space**.

Unfortunately the way Excel opens files with the extension .csv seems to be hardcoded in the program and cannot be changed with the APIs opentsv uses, requiring a hack: A .csv file is first copied to a file with the extension .tcsv and subsequently opened.

Conclusions

For more than a decade users have, without success, been asking Microsoft for an option to not convert values when loading comma or tab delimited data files in Excel because it can lead to erroneous data. Opentsv finally provides this option in an easy to install program that intervenes mostly transparently in loading data into Excel without the conversion. Its widespread use may help to reduce the significant of these errors in published literature.

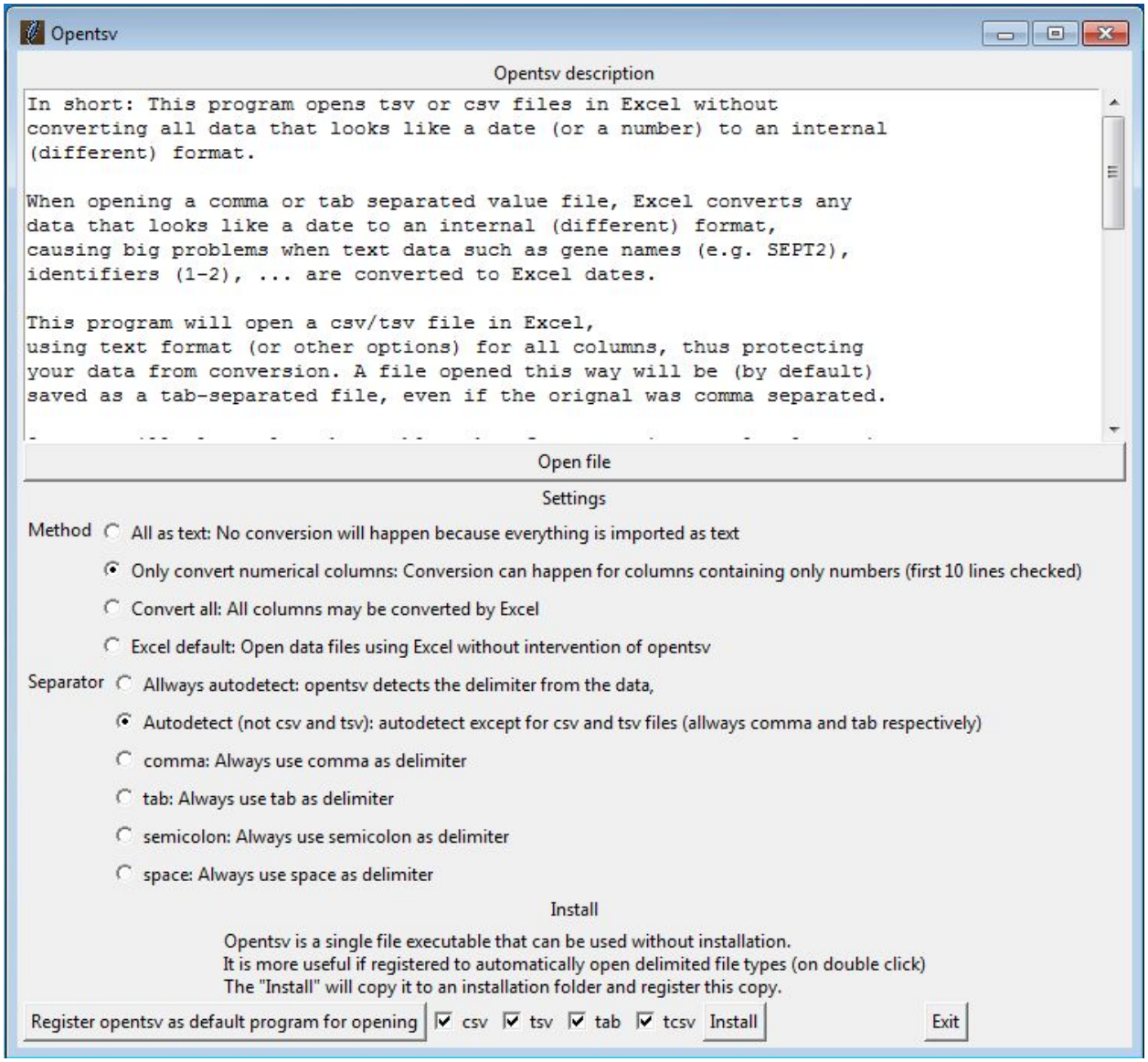

Figure 1

Settings and installation screen

References

Mallona,I. and Peinado,M.A. (2017) Truke, a web tool to check for and handle excel misidentified gene symbols. BMC Genomics, 18, 242.
Welsh,E.A. et al. (2017) Escape Excel: A tool for preventing gene symbol and accession conversion errors. PLoS ONE, 12, e0185207.
Zeeberg,B.R. et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol, 4, R28.

Zeeberg,B.R. et al. (2004) Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics, 5, 80.
Ziemann,M. et al. (2016) Gene name errors are widespread in the scientific literature. Genome Biology, 17, 177.