A data cleaner's cookbook - Windows carriage returns

On Linux and Mac machines, a newline is built with just one character, the UNIX linefeed '\n' ('LF'). On Windows computers, a newline is created using two characters, one after the other: '\r\n' ('CRLF'), where '\r' is called a 'carriage return' ('CR'). Carriage returns aren't necessary in a data table and can cause problems in data cleaning. (For examples of problems, see this *BASHing data* post.)

There are several ways to find CR characters. You can use **sed -n 'l'** to visualise any '\r' in a table, and **grep** to select out the lines with a CR and print their line numbers. Alternatively, a CR character will be shown as '^M' if you use **cat -v**, where the '-v' option shows non-printing characters other than tabs and linefeeds. In the example below, the file **winCR** has an invisible Windows carriage return at the end of the first line:

```
$ cat winCR
aaa   bbb
ccc   ddd
eee   fff

 $ sed -n 'l' winCR
aaa\tbbb\r$
ccc\tddd$
eee\tfff$

 $ sed -n 'l' winCR | grep -n "\\r"
1:aaa\tbbb\r$

 $ cat -v winCR | grep -n "\^M"
1:aaa   bbb^M$
```

It's wise to run these commands with **grep**'s '-c' option first rather than '-n'. The '-c' option returns only the *number* of lines with a CR, and if that number is big, you avoid having large number of lines printed at high speed in your terminal. If your **grep** supports Perl-type regular expressions, you can count '\r' characters directly.

```
$ sed -n 'l' winCR | grep -c "\\r"
1

 $ cat -v winCR | grep -c "\^M"
1

 $ grep -cP "\r" winCR
1
```

Another command to find carriage returns is **file**, which will report on line endings if they're different from a single linefeed, but won't count them:

```
$ file example.csv
example.csv:     ASCII text, with CRLF line terminators
```

The easiest way to remove all Windows carriage returns from **table** is with **tr**:

```
$ tr -d '\r' < table > table_without_CR
```

Deleting all the carriage returns could be a mistake, however, if any of them are *within* data items.

The screenshot below shows a real-world example. In the file **afd1**, I used **sed** to replace each of the 2 carriage returns in line 67893 with a single whitespace. Note that this was an 'in-place' edit with **sed**'s '-i' option.

```
$ cat -v afd1 | grep -n "\^M"
67893:          safrina Spilopyra safrina Reid & Beatson, 2010C
HRYSOMELIDAE    Spilopyra               safrina         Synonym
synonym Species         Reid & Beatson  2010    Y       2d32135
3-4268-467d-8335-f5802ece5a50   20110419T01:26:11.999+0000    9
99f0a0b-3651-4174-a87f-ca286902ea0a     20120323T01:38:19.026+0
000             4f91518a-ae5e-4912-9505-3cd5f08557ce    Reid, C
.A.M. & Beatson, M.     2010    Revision of the Australo-Papuan
 genus <i>Spilopyra</i> Baly^M(Coleoptera: Chrysomelidae: Spilo
pyrinae)        1-32            Zootaxa                         <
!--MARK-->Reid, C.A.M. & Beatson, M. 2010. Revision of the Aust
ralo-Papuan genus <i>Spilopyra</i> Baly^M(Coleoptera: Chrysomel
idae: Spilopyrinae). <!--MARK--><!--MARK--><em>Zootaxa</em> <st
rong>2692</strong>: 1-32<!--MARK-->              Article in Jour
nal     999deae1-59ee-41db-9159-3d43c22206af    20110419T01:03:
41.080+0000     96466227-221c-4922-ac97-1ac2dd946d0a
$
$ sed -i '67893s/\r/ /g' afd1
$
$ sed -n '67893p' afd1
                safrina Spilopyra safrina Reid & Beatson, 2010C
HRYSOMELIDAE    Spilopyra               safrina         Synonym
synonym Species         Reid & Beatson  2010    Y       2d32135
3-4268-467d-8335-f5802ece5a50   20110419T01:26:11.999+0000    9
99f0a0b-3651-4174-a87f-ca286902ea0a     20120323T01:38:19.026+0
000             4f91518a-ae5e-4912-9505-3cd5f08557ce    Reid, C
.A.M. & Beatson, M.     2010    Revision of the Australo-Papuan
 genus <i>Spilopyra</i> Baly (Coleoptera: Chrysomelidae: Spilop
yrinae) 1-32            Zootaxa                         <!--MAR
K-->Reid, C.A.M. & Beatson, M. 2010. Revision of the Australo-P
apuan genus <i>Spilopyra</i> Baly (Coleoptera: Chrysomelidae: S
pilopyrinae). <!--MARK--><!--MARK--><em>Zootaxa</em> <strong>26
92</strong>: 1-32<!--MARK-->              Article in Journal    9
99deae1-59ee-41db-9159-3d43c22206af     20110419T01:03:41.080+0
000     96466227-221c-4922-ac97-1ac2dd946d0a
$
```