

Cleaning up “dirty” text

By John Allen

There are two things that can be annoying about using Copy/Paste when copying formatted text from a web page, PDF file, rich text email etc:

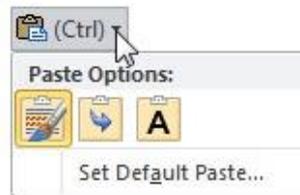
1. The format of the pasted text is different to existing text in the document/email etc.
2. Text in a paragraph appears in shortened lines and does not wrap at the end of each line.

Pasted text can contain a large amount of invisible ‘baggage’, such as HTML tags, Word formatting structure, etc. This baggage accompanies the text when you copy and paste it. To get around this problem, there are some options available for pasting into Word, or you can ‘launder’ the text through a simple text processor and paste it into any word, email, etc. document.

Pasting in Word (using Word 2010, but earlier versions are similar)

If you are pasting into a *Microsoft Word* document, there is a paste option that appears at the end of the pasted text, where you can choose whether to paste *as copied*, *match destination formatting* or *paste as unformatted text*. For example, if you copy some text from another Word document that has a different font and size, bold, italics etc, you can choose how to paste it. The following image shows what happens when I paste some text from another document into my current document –:

The quick brown fox jumps over the lazy dog



If I click on the down arrow and choose the  option, then the italics, font and size are changed to my default Word text format, and all formatting is lost. If I choose the  option, the pasted text is changed to the same format as surrounding text in the document.

And finally, the  option pastes the text as it was copied.

That fixes the first problem, but what about making a single paragraph from a number of lines separated by a carriage return?

Say that I copy the following text from a PDF file of the *Eurobodalla Voice* to a Word document.

ity Calendar

.30am)	South East Arts Every month South East Arts sends out an e-news alert to over 1000 subscribers across the region and beyond. The alert directs people to www.southeastarts.org.au . If you have an arts event, exhibition, festival or latest news to promote, you can include it on the News and What's On sections of the website. News entries feature unlimited text and space for one image. News entries can also include attached documents such as entry forms, posters and links to website. What's On entries can include text information describing your event as well as details on date, time, cost.
30m (1pm)	
5am)	
4 Hall	
4 Hall	If you would like to feature a News entry the website has the following categories: Latest News,

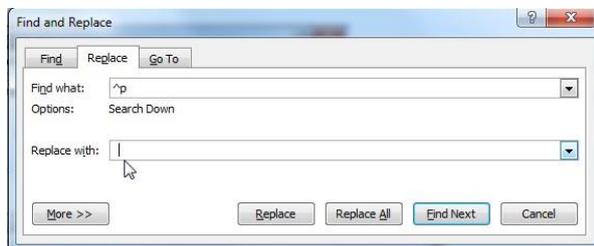
Pasting as unformatted text in Word gives me this:

South East Arts
Every month South East Arts sends out an e-news alert to over 1000 subscribers across the region and beyond. The alert directs people to www.southeastarts.org.au. If you have an arts event, exhibition, festival or latest news to promote, you can include it on the News and What's On sections of the website. News entries feature unlimited text and space for one image. News entries can also include attached documents such as entry forms, posters and links to website. What's On entries can include text information describing your event as well as details on date, time, cost.

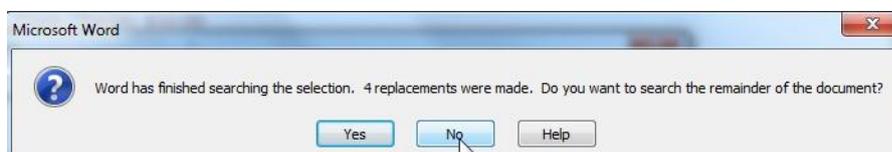
I now want to replace the carriage returns with blanks to join the lines into one paragraph. Select all the lines except the first one (which I still want on a separate line) then choose *Replace*.



A carriage return is specified in the *Find what* section as *carat p (^p)*. (the carat is upper case 6.) Then in the *Replace with* section - put a single space. Then click *Replace All*. (Word remembers these *Replace* settings for the current session, by the way)



Click *No* after the replacement has finished



Now the text looks like this – the lines have been joined:

South East Arts

Every month South East Arts sends out an e-news alert to over 1000 subscribers across the region and beyond. The alert directs people to www.southeastarts.org.au. If you have an arts event, exhibition, festival or latest news to promote, you can include it on the News and What's On sections of the website. News entries feature unlimited text and space for one image. News entries can also include attached documents such as entry forms, posters and links to website. What's On entries can include text information describing your event as well as details on date, time, cost.

All this works fine in *Word*, but no such paste options are available in email or some other types of documents. There is simple alternative.

'Laundering' the text through *NoteTab Lite*

The idea here is to paste formatted or rich text into a text editor, which automatically strips it of all the 'baggage' mentioned above. Then you copy the text from there and paste it where you want.

Notetab Lite is a very useful free text editor that is better than *Notepad*. You can use it to resolve the 'dirty' text and unjoined lines problems, then copy the laundered text and paste it into Email, Word, or any other document.

Removing text baggage

If you paste any formatted text into *NoteTab Lite*, all text formatting baggage, images etc. will be removed. Then go *Edit / Copy All*, then paste into your Word or email document.

Joining lines

If you need to join the lines together (as we just looked at in Word), this is very easy in *NoteTab Lite* – select all the lines (or *Ctrl A* to select all) that you want to join, then go *Modify / Lines / Join Lines* (or *Ctrl J*) to Join the lines. Then go *Edit / Copy All*, then paste into your Word or email document.

NoteTab Lite is available from www.notetab.com (downloads page)

Hint: when you initially install *NoteTab Lite*, change the settings as follows:

- *Tools / Clipbook Properties / Hide Clipbook*
- *View / Options / General – uncheck Reload Open Documents and Load Favorites on Startup*
- *View / Options / Files – uncheck Make backups, Save Files on close, and Automatic Save ...*
- If you want *Notetab* as your default text editor, go *View / Options / Associations – enter "txt" (without quotes) and click Add*