Reconstruction of Soft Hyphens

Author: Roger Dunham

Date: 30th March 2020

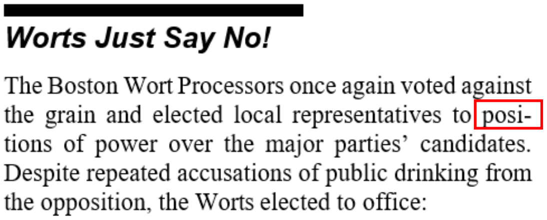
Version 1.2

What is a Soft Hyphen?

“a hyphen inserted into a word not otherwise hyphenated, to be displayed or typeset only if it falls at the end of a line of text.”

What this means is that if the text flow or layout changes so that a word that has a soft hyphen is no longer positioned at the end of the line, it will adjust to remove the hyphen that is no longer necessary.

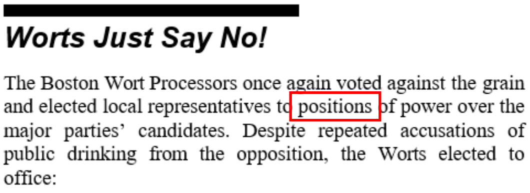
This is illustrated in the following figures.



Worts Just Say No!

The Boston Wort Processors once again voted against
the grain and elected local representatives to posi-
tions of power over the major parties’ candidates.
Despite repeated accusations of public drinking from
the opposition, the Worts elected to office:

Figure 1 - Text showing a soft hyphen.

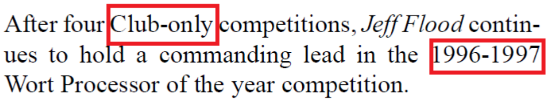


Worts Just Say No!

The Boston Wort Processors once again voted against the grain
and elected local representatives to positions of power over the
major parties’ candidates. Despite repeated accusations of
public drinking from the opposition, the Worts elected to
office:

Figure 2 - The same text, but at a smaller font size which has caused the word "positions" to no longer split over two lines. The soft hyphen is no longer visible.

This compares with “Hard Hyphens” which should always be present



After four Club-only competitions, *Jeff Flood* contin-
ues to hold a commanding lead in the 1996-1997
Wort Processor of the year competition.

Figure 3 – Text showing two hard hyphens.
Note that there is also a soft hyphen in the word “continues”.

Being able to tell the two types of hyphen apart helps to reconstruct Word documents that look like the PDF, and are easy to edit and search.

Solid Framework, Soft Hyphens and Reconstructing Documents

When Solid Framework finds a hyphen at the end of a line of text it identifies whether it should be treated as a hard or soft hyphen.

If converting to a Word document then different codes are used for each type of hyphen (which allows Word to correctly show or hide it).

If converting to a text document then soft hyphens are removed entirely.

In either case, one of the benefits of soft hyphen detection is the ability to search for words that wrap across a line break. If soft hyphens are not detected as such, then the parts of the word before and after the hyphen are considered to be separate. As such the word "continues" shown in Figure 3 (see above) would not be found when searching the text.

Using the New Functionality

The ability to detect soft hyphens when converting to Word[1] was implemented in Solid Framework 10.0.10054.

By default, soft hyphen detection is turned off. This means that a PDF reconstructed with this version of Solid Framework will, by default, treat hyphens in the same way as previous versions of Solid Framework. To use the new functionality it must be explicitly enabled.

To enable soft hyphen detection, use:

converter.DetectSoftHyphens = true;

where converter is an instance of a PdfToWordConverter.

Explicitly Encoded Soft Hyphens in PDFs

Unicode character U+00AD represents a soft hyphen. If the PDF contains this character in a Unicode Map then it is possible that the hyphen may render in the PDF, but the instructions when exporting the document indicate that it should not be visible. This is a conflict within the PDF itself.

Solid Framework aims to recreate a Word document that looks like the PDF. To resolve this conflict Solid Framework substitutes the soft hyphen with a "real" hard hyphen during reconstruction, provided that DetectSoftHyphens is enabled[2].

[1] It has been available when converting to text since 9.2.8528.

[2] It appears that Word has its own logic for "optional" hyphens that does not include handling `U+00AD` characters. As such they are always shown as hyphens, regardless of where they appear in a line.