POI-HWPF - A Quick Guide

Overview

by Nick Burch

HWPF is still in early development. It is in the <u>scratchpad section of the SVN</u>. You will need to ensure you either have a recent SVN checkout, or a recent SVN nightly build (including the scratchpad jar!)

1. Basic Text Extraction

For basic text extraction, make use of org.apache.poi.hwpf.extractor.WordExtractor. It accepts an input stream or a HWPFDocument. The getText() method can be used to get the text from all the paragraphs, or getParagraphText() can be used to fetch the text from each paragraph in turn. The other option is getTextFromPieces(), which is very fast, but tends to return things that aren't text from the page. YMMV.

2. Specific Text Extraction

To get specific bits of text, first create a org.apache.poi.hwpf.HWPFDocument. Fetch the range with getRange(), then get paragraphs from that. You can then get text and other properties.

3. Changing Text

It is possible to change the text via insertBefore() and insertAfter() on a Range object (either a Range, Paragraph or CharacterRun). It is also possible to delete a Range, but this code is know to have bugs in it.

4. Further Examples

For now, the best source of additional examples is in the unit tests. <u>Browse the HWPF unit tests.</u>