# POI-HPBF - Java API To Access Microsoft Publisher Format Files

## Overview

**by Nick Burch**

### 1. Overview

HPBF is the POI Project's pure Java implementation of the Publisher file format.

Currently, HPBF is in an early stage, whilst we try to figure out the file format. So far, we have basic text extraction support, and are able to read some parts within the file. Writing is not yet supported, as we are unable to make sense of the Contents stream, which we think has lots of offsets to other parts of the file.

Our initial aim is to provude a text extractor for the format (now done), and be able to extract hyperlinks from within the document (partly supported). Additional low level code to process the file format may follow, if there is demand and developer interest warrant it.

Text Extraction is available via the *org.apache.poi.hpbf.extractor.PublisherTextExtractor* class.

At this time, there is no *usermodel* api or similar. There is only low level support for certain parts of the file, but by no means all of it.

Our current understanding of the file format is documented here.

**Note:**

This code currently lives the scratchpad area of the POI SVN repository. Ensure that you have the scratchpad jar or the scratchpad build area in your classpath before experimenting with this code.