Apache POI - HWPF - Java API to Handle Microsoft Word Files

Overview

by Nicola Ken Barozzi, Andrew C. Oliver, Ryan Ackley, Rainer Klute

1. Overview

HWPF is the name of our port of the Microsoft Word 97(-2007) file format to pure Java. It *does not* support the new Word 2007 .docx file format, which is not OLE2 based.

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

Source in the *org.apache.poi.hwpf.model* tree is the old legacy code refactored into an object model. Source code in the *org.apache.poi.hwpf.extractor* tree is a wrapper of this to facilitate easy extraction of interesting things (eg the Text). Source code in the *org.apache.poi.hdf* tree is the old legacy code.

1.1. HWPF Pointman Needed!

At the moment we unfortunately do not have someone taking care for HWPF and fostering its development. What we need is someone to stand up, take this thing under his hood as his baby and push it forward. Ryan Ackley, who put a lot of effort into HWPF, is no longer on board, so HWPF is an orphan child waiting to be adopted.

If **you** are interested in becoming the new HWPF pointman, you should look into the Microsoft Word internals. A good starting point seems to be Ryan Ackley's overview. This document contains a link to a detailled Word format description you can find somewhere at http://www.wotsit.org/. Please do not contact Ryan Ackley directly, because he is working for a company now that signed a NDA with Microsoft and thus he will be no longer able to answer questions.

As a first step you should familiarize yourself with the source code, examples, test cases, and the HWPF patches available at Bugzilla (if any). Then you should compile an overview of

- the current HWPF status,
- the patches in Bugzilla to be checked in (and those that should better be ditched),
- the available test cases and the test cases still to be written,
- the available documentation and the docs to be written,
- anything else that seems reasonable

When you start start coding, you will not yet have write access to the CVS repository. Please submit your patches to Bugzilla and nag Rainer Klute until he commits them. Besides the actual checking in of HWPF patches Rainer will also do some minor reviews now and then of your source code patches, test cases and documentation to help ensure software quality. But most of the time you will be on your own.

Please do not forget to write JUnit test cases and documentation! We won't accept code that doesn't come with test cases. And please consider that other contributors should be able to understand your source code easily. If you need any help getting started with JUnit test cases for HWPF, please ask on the developers' mailing list! If you show that you are prepared to stick at it you will most likely be given CVS commit access.

Important: It is legally vital for POI that you have never seen any documentation or specification from Microsoft that required you or your employer to sign an NDA to get it. Please do read the "Contribution to POI" page for details! This page also contains further information for you to start POI development.

Of course we will help you as best as we can. However, presently there is no committer who is really familiar with the Word format, so you'll be mostly on your own. We are looking forward for you and your contributions! Honor and glory of becoming a POI committer are waiting!