POI-HDGF - Java API To Access Microsoft Visio Format Files

Overview

by Nick Burch

1. Overview

HDGF is the POI Project's pure Java implementation of the Visio file format.

Currently, HDGF provides a low-level, read-only api for accessing Visio documents. It also provides a way to extract the textual content from a file.

At this time, there is no *usermodel* api or similar, only low level access to the streams, chunks and chunk commands. Users are advised to check the unit tests to see how everything works. They are also well advised to read the documentation supplied with vsdump to get a feel for how Visio files are structured.

To get a feel for the contents of a file, and to track down where data of interest is stored, HDGF comes with VSDDumper to print out the contents of the file. Users should also make use of vsdump to probe the structure of files.

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.

1.1. Steps required for write support

Currently, HDGF is only able to read visio files, it is not able to write them back out again. We believe the following are the steps that would need to be taken to implement it.

1. Re-write the decompression support in LZW4HDGF as HDGFLZW, which will be much better documented, and also under the ASL. **Completed October 2007**
2. Add compression support to HDGFLZW. **In progress - works for small streams but encoding goes wrong on larger ones**
3. Have HDGF just write back the raw bytes it read in, and have a test to ensure the file is

un-changed.

4. Have HDGF generate the bytes to write out from the Stream stores, using the compressed data as appropriate, without re-compressing. Plus test to ensure file is un-changed.
5. Have HDGF generate the bytes to write out from the Stream stores, re-compressing any streams that were decompressed. Plus test to ensure file is un-changed.
6. Have HDGF re-generate the offsets in pointers for the locations of the streams. Plus test to ensure file is un-changed.
7. Have HDGF re-generate the bytes for all the chunks, from the chunk commands. Tests to ensure the chunks are serialized properly, and then that the file is un-changed
8. Alter the data of one command, but keep it the same length, and check visio can open the file when written out.
9. Alter the data of one command, to a new length, and check that visio can open the file when written out.