Apache POI - POIFS - Documents embeded in other documents

Overview

by Nick Burch, Yegor Kozlov

1. Overview

It is possible for one OLE 2 based document to have other OLE 2 documents embeded in it. For example, and Excel file may have a word document and a powerpoint slideshow embeded as part of it.

Normally, these other documents are stored in subdirectories of the OLE 2 (POIFS) filesystem. The exact location of the embeded documents will vary depending on the type of the master document, and the exact directory names will differ each time. To figure out exactly which directory to look in, you will either need to process the appropriate OLE 2 linking entry in the master document, or simple iterate over all the directories in the filesystem.

As a general rule, you will find the same OLE 2 entries in the subdirectories, as you would've found at the root of the filesystem were a document to not be embeded.

1.1. Files embeded in Excel

Excel normally stores embeded files in subdirectories of the filesystem root. Typically these subdirectories are named starting with MBD, with 8 hex characters following.

1.2. Files embeded in Word

Word normally stores embeded files in subdirectories of the ObjectPool directory, itself a subdirectory of the filesystem root. Typically these subdirectories and named starting with an underscore, followed by 10 numbers.

1.3. Files embeded in PowerPoint

PowerPoint does not normally store embeded files in the OLE2 layer. Instead, they are held within records of the main PowerPoint file.
See the HSLF Tutorial for how to retrieve embedded OLE objects from a presentation

2. Listing POIFS contents

POIFS provides a simple tool for listing the contents of OLE2 files. This can allow you to see what your POIFS file contents, and hence if it has any embeded documents in it, and where.

The tool to use is *org.apache.poi.poifs.dev.POIFSLister*. This tool may be run from the command line, and takes a filename as its parameter. It will print out all the directories and files contained within the POIFS file.

3. Opening embeded files

All of the POIDocument classes (HSSFWorkbook, HSLFSlideShow, HWPFDocument and HDGFDiagram) can either be opened from a POIFSFileSystem, or from a specific directory within a POIFSFileSystem. So, to open embeded files, simply locate the appropriate DirectoryNode that represents the subdirectory of interest, and pass this + the overall POIFSFileSystem to the constructor.

I you want to extract the textual contents of the embeded file, then open the appropriate POIDocument, and then pass this to the extractor class, instead of simply passing the POIFSFilesystem to the extractor.