POI-HSLF - A Quick Guide

Overview

by Nick Burch

1. Basic Text Extraction

For basic text extraction, make use of `org.apache.poi.extractor.PowerPointExtractor`. It accepts a file or an input stream. The `getText()` method can be used to get the text from the slides, and the `getNotes()` method can be used to get the text from the notes. Finally, `getText(true,true)` will get the text from both.

2. Specific Text Extraction

To get specific bits of text, first create a `org.apache.poi.usermodel.SlideShow` (from a `org.apache.poi.HSLFSlideShow`, which accepts a file or an input stream). Use `getSlides()` and `getNotes()` to get the slides and notes. These can be queried to get their page ID (though they should be returned in the right order). You can also call `getTextRuns()` on these, to get their blocks of text. From the `TextRun`, you can extract the text, and check what type of text it is (eg Body, Title)

3. Poor Quality Text Extraction

If speed is the most important thing for you, you don't care about getting duplicate blocks of text, you don't care about getting text from master sheets, and you don't care about getting old text, then `org.apache.poi.extractor.QuickButCruddyTextExtractor` might be of use.

QuickButCruddyTextExtractor doesn't use the normal record parsing code, instead it uses a tree structure blind search method to get all text holding records. You will get all the text, including lots of text you normally wouldn't ever want. However, you will get it back very very fast!

There are two ways of getting the text back. `getTextAsString()` will return a single string with all the text in it. `getTextAsVector()` will return a vector of strings, one for each text record found in the file.

4. Changing Text

It is possible to change the text via `TextRun.setText(String)`. However, if the length of the text is changed, things will break because PowerPoint has internal file references in byte offsets. We currently update all of these byte references that we know about when writing out, but there are a few more still to be found.

5. Guide to key classes

- `org.apache.poi.hslf.HSLFSlideShow` Handles reading in and writing out files. Calls `org.apache.poi.hslf.record.record` to build a tree of all the records in the file, which it allows access to.
- `org.apache.poi.hslf.record.record` Base class of all records. Also provides the main record generation code, which will build up a tree of records for a file.
- `org.apache.poi.hslf.usermode.SlideShow` Builds up model entries from the records, and presents a user facing view of the file
- `org.apache.poi.hslf.extractor.PowerPointExtractor` Uses the model code to allow extraction of text from files