History of Changes

Table of contents

RSS

(changes.rss)

1. Version 0.7.1 (04/10/2005)

- [1119408] Support named target for Bookmark extraction.(BJL)
- Created Resources/PDFBox_External_Fonts.properties to create a mapping for non-embedded fonts(BJL)
- Added implementation for PDF page articles(BJL)
- Created TextToPDF command line application(BJL)
- Created ImageToPDF example(BJL)
- [1119420] Extract and Update the Meta-Information as XML(BJL)
- [1119410] Extract text in/between bookmarks(BJL)
- [1164476] XFDFImport should fail with non XFDF document(BJL)
- **API Change** Renamed PDField.getName() to PDField.getPartialName(), added method getFullyQualifiedName() (BJL)
- **API Change** Renamed PDWidget to PDAnnotationWidget for naming consistency(BJL)
- Text is now extracted from embedded form xobjects.(BJL)
- Deployed site to new hosting vendor.(BJL)
- committed code for PDFHighlighter to highlight words in a PDF document.(BJL)
- Added command line application org.pdfbox.PDFToImage(BJL)
- Implemented runlength decoding(BJL)
- Added patch from Jorge Hernández Sellés to append content streams to existing page.(BJL)
- **API Change**renamed package from pdmodel.graphics.image to pdmodel.graphics.xobject(BJL)
- **API Change**Removed PDRadioButton, should use PDCheckbox instead(BJL)
- **API Change**COSStream now extends COSDictionary instead of containing a dictionary(BJL)
- [1021241] Text extraction should follow PDF article divisions(BJL)
- [1170068] text field is not found(BJL)
- fixed NPE issue where an image did not have any applied filters(BJL)
- Fixed issue where extra spaces were being added during text extraction for type3 fonts(BJL)
- fixed parsing of header where a trailing % exists(BJL)
- [1110029] Character ">" not quoted in COSName::writePDF(BJL)

2. Version 0.7.0 (1/22/2005)

- Added implementation for PDF Bookmarks(BJL)

- Added implementation for PDF Destinations(BJL)
- committed [1097913] Enhance LucenePDFDocument streams(thanks to Olivier Parent)(BJL)
- Updated website for better format for documentation(BJL)
- Now ExportFDF and ExportXFDF will default output files to pdfname.fdf and pdfname.xfdf(BJL)
- [1046278] ClassCastException when doing FDF/XFDF(BJL)
- ExtractText now allows you to extract text if you decrypt with the owner password(BJL)
- Added PDF 1.5 Object Stream support(BJL)
- Added pdmodel.common.PDStream to represent COSStream(BJL)
- changed PDPage.getContents to use PDStream instead of COSStream(BJL)
- Updated LucenePDFDocument Javadoc to tell which Lucene fields it populates(BJL)
- moved HelloWorld example from persistence to pdmodel and updated to use new PD Model features(BJL)
- Refactored PDFStreamEngine based on contributions from Christophe Huault(BJL)
- This class no longer uses a gigantic if/else statement for all of the operators they are defined as properties when instantiating the class(BJL)
- Updated AFM resources to be ones released on Adobe's site, include AFM license as well(BJL)
- Added ability to embed TTF fonts, only WinAnsiEncoding is supported at this time(BJL)
- Added ability to extract images, thanks to contributions by Brigitte Mathiak(BJL)
- COSWriter now generates the document id if it does not already exist(BJL)
- improved performance for text extraction(BJL)
- [1058693] TextPosition does not take account of tz operator(BJL)
- upgraded to log4j-1.2.9(BJL)
- include package-list for javadocs(BJL)
- [1037145] Infinite loop in PDFParser.parseObject(BJL)
- fixed error where spaces before integers was causing parse errors(BJL)

3. Version 0.6.7 (10/09/2004)

- Added the following command line applications (BJL)
- Revamped the way character spacing and font information is obtained(BJL)
- Improved location information about a character drawn on the screen.(BJL)
- Changed the PDFStreamEngine.showString to showCharacter to support the newly improved location information. This will now only show one character at a time.(BJL)
- Fixed bug in PDDocument.isOwnerPassword and isUserPassword that was using the wrong length for the encryption key(BJL)
- Upgraded to ant 1.6.2(BJL)
- Upgraded to checkstyle-3.4(BJL)
- Upgraded to JUnit-3.8.1(BJL)

- Upgraded to lucene-1.4.2(BJL)
- Integrated patch(1016603) for issue 943319 to fix parsing of open office documents(BJL)
- Patch:985347 No longer throw exception for "No 'ToUnicode' and no 'Encoding' for Font"(BJL)
- Patch:996191 Fixed case statement with missing break(BJL)
- Patch:996781 Fixed null pointer exception in acroform fields(BJL)
- Renamed DecryptDocument to DocumentEncryption to support encryption and decryption(BJL)
- Added load/save/encrypt/decrypt convenience methods on the PDDocument class(BJL)
- COSWriter now attempts to keep object numbers from parsed documents and writes 'free' entries in the xref if necessary(BJL)
- Added the ability to set the word separator on the PDFTextStripper(BJL)
- Fixed issue where PDFBox would throw an IOException if a PDF was incorrectly missing an endobj tag(BJL)
- Fixed 918220 where PDFBox would freeze when parsing certain cmap files(BJL)
- Added initial colorspace support(BJL)
- Fixed issue where AppendDoc was throwing ClassCastException(BJL)
- Fixed 1013163 Can't parse filters that use filter abbreviation(BJL)
- Fixed 1011244 Where encrypting then decrypting was causing a problem(BJL)
- renamed TextPosition.getWidth to TextPosition.getCombinedHorizontalDisplacement to better reflect its actual value(BJL)
- Fixed 919215 PDFBox now support stream replacement(BJL)
- Fixed 955043 Added support for 'ETenms-B5-H' encoding(BJL)
- Fixed 996050 Class Cast exception when importing(BJL)
- Added support for Font descriptors(BJL)
- Fixed spacing issues when doing textfield FDF import(BJL)
- Fixed 1017175 Large number converted when re-written(BJL)
- Fixed 1029873 PDFBox now allows for multiple xref sections(BJL)
- Added support for document Viewer Preferences(BJL)
- Made currentDocument and pdfDocument protected in util.Splitter to allow easier subclassing(BJL)
- Fixed 1034427 After Splitting page orientation is lost(BJL)

4. Version 0.6.6 (07/20/2004)

- Improved support for setting of checkbox fields(FDF import)(BJL)
- Added the org.pdfbox.PDFSplit utility to split a single document into many documents(BJL)
- PDFBox now ignore the Length field that is associated with a stream, it has been found to be wrong in some documents(BJL)
- Fixed bug when writing out PDF documents and the document contained an non

alphabetic character such as (or)(BJL)
- Fixed bug in PDFont where dictionary encodings where not being processed correctly(BJL)
- Fixed bug in COSDocument.isEncrypted which was comparing COSNull to the wrong object(BJL)
- Integrated patch for supporting multiple lines in the appearance stream(BJL)
- Upgraded to lucene-1.4-final(BJL)
- org.pdfbox.ExtractText now uses the system encoding as the default encoding instead of ISO-8859-1(BJL)

5. Version 0.6.5 (03/08/2004)

- Fixed bug in revision 3 encryption algorithm(BJL)
- added support for CIDFontType0 glyph widths, which fixed issue with spaces being during text extraction(BJL)
- Fixed infinite loop when parsing a corrupt content stream(BJL)
- Add characterspacing + wordspacing when determining the width of a space character(BJL)
- Added support for more font types(BJL)
- refactored the pdmodel.interactive package, form fields use object delegation instead of inheritance for the widget, see PDField.getWidget and PDField.getKids(BJL)
- Fixed bug where an inheritable cropbox would cause stackoverflow exception(BJL)
- Changed usage of PDField/PDWidget to look like object delegation instead of inheritance by adding a PDField.getWidget instead of extending PDWidget(BJL)
- refactored interactive package, this will break any existing code that uses the PDField/PDAnnotation classes. You will need to adjust your package names!!(BJL)
- Now uses StandardEncoding as the default encoding(BJL)
- Bug in AppendDoc example that did not take into account groups of pages(BJL)
- PDFont now also tries the bootstrap classloader when loading AFM resources(BJL)
- added -startPage and -endPage command line options to org.pdfbox.ExtractText(BJL)
- Added support for corrupt PDFs with garbage before the header(BJL)
- Fixed bug where there was whitespace instead of garbage characters in front of the first object(BJL)
- performance improvements for the Matrix implementation(BJL)
- upgraded to lucene 1.3(BJL)
- fixed bug in cmap parser for cmap files that all ended in 'def'(BJL)
- Removed createObject method from COSDocument, COSWriter will handle all object references for you(BJL)
- Updated AppendDoc to use PDDocument instead of COSDocument and a couple bug fixes(BJL)
- PDFParser now closes the document if there were parse errors(BJL)

- TextPosition now has the PDFont that is associated with the piece of text(BJL)
- Added initial version of org.pdfbox.PDFViewer, a GUI application to view the internal structure of a PDF document. This can be used for debugging purposes at this time but may end up being a Adobe Reader like application if there is enough interest(BJL)
- Changed COSNumber/COSInteger/COSFloat interface to have both intValue and longValue(BJL)
- Added methods isUserPassword & isOwnerPassword to PDDocument(BJL)
- Added cmap files for CJK languages, please give me some feedback(BJL)

6. Version 0.6.4 (11/02/2003)

- Fixed bug which caused infinite loop(BJL)
- Fixed bug in encoding where DictionaryEncoding kept a reference instead of making a copy leading to encoding problems(BJL)
- Added PDFTextStripper.(get|set)PageSeparator, which will allow the user to output a string after every page(BJL)
- refactored text stripping code to separate the logic processing of PDF operators and the logic of extracting text(BJL)
- ran findbugs on source code and fixed a couple minor issues(BJL)
- Refactored font functionality to PDFont, some API methods are no longer available in COSObject(BJL)
- changed name of org.pdfbox.Main to org.pdfbox.ExtractText(BJL)
- added contribution of org.pdfbox.Overlay from Mario Ivankovits(BJL)
- added log.isDebugEnabled checks to log4j calls(BJL)
- added better escaping when writing COSNames(BJL)
- fixed bug where encryption dictionary is sometimes set to COSNull instead of not being present(BJL)

7. Version 0.6.3 (09/13/2003)

- Now contains the ability to import/set FDF data thanks to a contribution from Stefan Uldum Grinsted(BJL)
- No longer throw an error when stream is not followed by 0A or 0D0A to allow more PDFs to be parsed(BJL)
- Added -encoding argument to org.pdfbox.Main to control the encoding of the output(BJL)
- Remove Prev entry from trailer if it exists because PDFBox automatically clears all old entries, only an issue when modifying/saving an existing PDF document(BJL)
- Fixed bug in master password encryption algorithm for Revision 3 encrypted documents(BJL)
- COSString no longer uses UTF-8 when encoding the byte array(BJL)
- Added PDDocument.getPageCount()(BJL)

- Fixed bug in PDFEncryption where(BJL)
- Now enforces text extraction permissions(BJL)

8. Version 0.6.2 (4/18/2003)

- Added required libraries to CVS(BJL)
- Added log4j logging(BJL)
- Added automated tests and test data for text extraction(BJL)
- Significant text extraction work(BJL)
- Modified build so that build.properties settings are no longer required(BJL)
- Added automatic handling of files encrypted with the empty password(BJL)
- Removed unimplemented decoders from filters test(BJL)
- Fixed several LZW decode bugs introduced after 0.5.6(BJL)
- Fixed bugs relating to processing out of spec PDF's with bad # escaping in the name ("java.io.IOException: Error: expected hex number" bug)(BJL)
- Fixed Lucene UID generation bug(BJL)
- Fixed GetFontWidths null pointer exception bug(BJL)

9. Version 0.6.1 (3/9/2003)

- Fixed bug in parsing stream objects which led to "Unexpected end of ZLIB input stream"(BJL)
- Changed license from LGPL to BSD to allow pdfbox to be used easily in Apache projects(BJL)

10. Version 0.6.0 (3/5/2003)

- Added PDF document summary fields to the lucene document(BJL)
- Massive improvements to memory footprint(BJL)
- Must call close() on the COSDocument(LucenePDFDocument does this for you)(BJL)
- Really fixed the bug where small documents were not being indexed(BJL)
- Fixed bug where no whitespace existed between obj and start of object. Exception in thread "main" java.io.IOException: expected='obj' actual='obj<</Pro(BJL)
- Fixed issue with spacing where textLineMatrix was not being copied properly(BJL)
- Fixed 'bug' where parsing would fail with some pdfs with double endobj definitions(BJL)

11. Version 0.5.6 (11/28/2002)

- Fixed bug in LucenePDFDocument where stream was not being closed and small documents were not being indexed (BJL)
- Fixed a spacing issue for some PDF documents (BJL)
- Fixed error while parsing the version number (BJL)
- Fixed NullPointer in persistence example (BJL)

- Create example lucene IndexFiles class which models the demo from lucene (BJL)
- Fixed bug where garbage at the end of file caused an infinite loop (BJL)
- Fixed bug in parsing boolean values with stuff at the end like "true>>" (BJL)

12. Version 0.5.5 (10/03/2002)

- Added example of printing document signature(BJL)
- Added example to print out form fields values(BJL)
- Fixed bug when appending documents(BJL)
- Various other bug fixes(BJL)

13. Version 0.5.4 (09/17/2002)

- Fixed bug in text output where '?' instead of the proper character(BJL)
- Fixed bug where sections of text were not being output at all(BJL)

14. Version 0.5.3 (09/13/2002)

- Fixed bug in 128 bit encryption(BJL)

15. Version 0.5.2 (09/06/2002)

- Catch all NumberFormatExceptions and wrap them with IOExceptions(BJL)
- Fixed bug where FDF documents could not be appended to PDF Documents(BJL)

16. Version 0.5.1 (09/04/2002)

- Now supports unicode for the document summary(BJL)
- Better support for Type0 fonts(BJL)
- Fixed bug with an empty LZW stream(BJL)
- Fixed parsing error for ID operator(BJL)

17. Version 0.5.0 (08/31/2002)

- Now supports unicode for the document summary(BJL)
- Better support for Type0 fonts(BJL)
- Fixed bug with an empty LZW stream(BJL)
- Fixed parsing error for ID operator(BJL)

18. Version 0.4.1 (07/25/2002)

- Fixed bug where .notdef was being output as document text(BJL)

19. Version 0.4.0 (07/23/2002)

- Added extract text ant task(BJL)
- Implemented AFM(Adobe Font Metrics) resource loading(BJL)
- Changed project from pdfparser to pdfbox to better reflect future needs(BJL)
- Fixed numerous bugs submitted by users(BJL)

20. Version 0.3.0 (07/09/2002)

- Added indexer for the lucene project(BJL)
- Initial implementation of PDF encryption(not working yet)(BJL)

21. Version 0.2.0 (06/03/2002)

- Added support for the various encodings(BJL)
- Improved the accuracy of the text output(BJL)

22. Version 0.1.0 (05/25/2002)

- Initial Version(BJL)