PDFBoxProposal

PDFBox

Abstract

PDFBox is an open source Java PDF library for working with PDF documents.

Proposal

The PDFBox library allows creation of new PDF documents, manipulation of existing documents and the ability to extract content from documents. PDFBox also includes several command line utilities. Future development plans include extending PDFBox with advanced data extraction and high level PDF creation functionality.

In addition to PDFBox, this proposal also covers the FontBox and JempBox companion libraries. FontBox is a Java font library used to obtain low level information from font files. JempBox is a Java library that implements Adobe's XMP specification. All these components would be incubated as a single Apache PDFBox podling project.

Background

The PDFBox project started in 2002 and was originally written by Ben Litchfield in 2002 and currently lives on SourceForge. The initial purpose of PDFBox was to extract text content to be indexed by the Lucene search engine. In addition to text extraction the library also supports a low level API for PDF creation and manipulation. In the past, several developers have helped develop specific features in PDFBox but none have continued once their specific needs where met.

In 2006 discussions began with the FOP team to collaborate on a single PDF library within the Apache organization. New projects have expressed interest in advancing the functionality of PDFBox.

Recently, Tika also expressed interest in advancing the content extraction capabilities of PDFBox.

The FontBox and JempBox libraries have no dependencies to PDFBox, but their primary purpose is to support PDFBox and the development community is largely overlapping. It makes sense to include all three libraries in a single project.

Rationale

The PDF document format is a common format found on internet and across industries as a way of sharing documents. Several Apache projects utilize PDF technologies but there is not a single independent PDF library within the Apache organization.

The Apache XML Graphics project (FOP/Batik) has a write-only PDF library and is in need of PDF parsing functionality. Many features overlap those of PDFBox. This is currently a duplication of effort, bringing PDFBox into Apache and combining our efforts will result in a more robust PDF library that will be able to support many more use cases for working with PDF technologies.

FontBox, FOP and Batik all contain font loading/handling code that could likely be merged into a single common library either within the PDFBox podling or outside it.

Initial Goals

The initial goals are:

- Advanced text extraction techniques
- Increase community involvement
- Cooperation with existing Apache projects such as XML Graphics
- Increasing support for PDF document features
- Adding a high level API for document creation
- Adding a streaming API for document creation
- PDF/A creation and validation functionality
- Review licensing of both bundled and external dependencies
- Manage export control notices for cryptographic features
- Figure out how to handle font handling code across FontBox, FOP, and Batik
- Replace JempBox with Adobe's XMP library

Current Status

Meritocracy

Not all initial committers are familiar with the meritocracy principles of Apache. It is expected that the committers that are not will learn the meritocracy rules and they will be followed through the life of the project.

Community

PDFBox has existed for several years on SourceForge and has an active community and continues to grow each day. There are hundreds of existing projects that utilize the current version of PDFBox.

Core Developers

Ben Litchfield is the main developer on this project although it is expected that developers from a variety of existing Apache projects will become part of the team.

Alignment

The ability to search PDF documents is a basic requirement for any enterprise search solution. PDFBox provides the basic content that is needed for content indexing. This functionality aligns with the those of Lucene, Nutch, Tika and UIMA and all users of these projects will benefit from continued development of PDFBox.

PDFBox shares similar font loading and handling needs as FOP and Batik, and the code in the FontBox companion library could well be merged with similar code in the other projects.

Known Risks

Orphaned products

PDFBox has been in development for over 5 years. The rate of development has varied, but the PDFBox user community has grown each year. PDFBox implements the PDF specification, which is highly utilized by companies across the world. The need for a PDF library is strong and is unlikely to change in the near future.

"Competing" formats

In recent times, additional paged document formats have been developed (or are in development) that have similar goals/functionality:

- Microsoft's XPS (XML-based ZIP container, proprietary core-functionality)
- Adobe's Mars (XML-based ZIP container, largely based on open standards, extending them where necessary)

Inexperience with Open Source

All developers have experience with Open Source projects.

Homogenous Developers

The initial set of committers is diverse and the project is likely to attract new developers.

Reliance on Salaried Developers

PDFBox is not the primary job for any of the initial committers.

Relationships with Other Apache Products

PDFBox has relationships with the following Apache Products

- Apache Lucene Lucene users typically integrate with PDFBox to add PDF indexing capabilities.
- Lucene Nutch Nutch currently utilizes PDFBox to index PDF documents.
- Tika Tika currently utilizes PDFBox for extracting PDF content.
- Apache UIMA UIMA analyzes unstructured content and would benefit from PDF content.
- Apache FOP and Apache Batik There's an experimental plug-in (currently hosted outside of the project) for FOP that uses PDFBox to support embedding of existing PDFs in XSL-FO documents for PDF output. Both Batik and FOP have code to parse fonts which FontBox needs to do, too.

A Excessive Fascination with the Apache Brand

Many existing Apache developers are already familiar with PDFBox. PDFBox was initially written to compliment the functionality of Lucene and has worked with it's developers over the past several years. PDFBox will benefit from closer cooperation with several existing Apache projects.

Documentation

- PDFBox (http://www.pdfbox.org/)
- FontBox (http://www.fontbox.org/)
- JempBox (http://www.jempbox.org/)

Initial Source

Initial source will come from the existing SourceForge repositories of the PDFBox, FontBox, and JempBox projects.

Source and Intellectual Property Submission Plan

The initial IP submission will be done as a software grant to the ASF.

External Dependencies

The "Adobe AFM License" and the "SUN JAI" licenses described below need to be reviewed to ensure they comply with Apache license standards.

Library	License	Description		
Adobe AFM	Adobe AFM License	Resources for extracting font encoding. Bundled inside PDFBox jar file.		
Bouncycastle	BSD Variant	Support for encrypting/decrypting PDF documents.		
<ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="187e8e7d-a910-4577-8af6-1be1d956caa1"><ac:plain-text-body><![CDATA[	IKVM	BSD Variant [1]	Support of PDFBox on .NET platform	]]></ac:plain-text-body></ac:structured-macro>
junit	CPL	Unit Testing Framework		
Lucene	ASL	Provide classes for easy Lucene integration		
JAI-CMM	Sun JAI	Provides support from color spaces		

[1] IKVM itself is BSD but contains either GNU Classpath or the OpenJDK class library (both GPL with exception). This may need to be reviewed, too.

Cryptography

PDFBox implements the RC4 encryption algorithm and utilizes Bouncy Castle for additional encryption routines.

Required Resources

Mailing lists

- pdfbox-dev@incubator.apache.org
- pdfbox-commits@incubator.apache.org
- pdfbox-private@incubator.apache.org

Subversion Directory

- https://svn.apache.org/repos/asf/incubator/pdfbox

Issue Tracking

- JIRA PDFBox (PDFBOX)

Other Resources

- none

Initial Committers

Name	Email	CLA
Ben Litchfield	ben at benlitchfield dot com	No
Daniel Wilson	williamstonconsulting at gmail dot com	No
Philipp Koch	pkoch at apache dot org	Yes

Affiliations

Name	Affiliation
Ben Litchfield	Independent

Daniel Wilson	DV Brown Company
Philipp Koch	Day Software

Sponsors

Champion

- Jukka Zitting

Nominated Mentors

- Jukka Zitting
- Jeremias Maerki
- Niall Pemberton

Sponsoring Entity

- Apache Incubator PMC