ENHANCING AN OPEN SOURCE PDF DOCUMENT WRITER

BY

Howell C. Balolong

Aileen A. Gullos

Jan Alexander T. Lao

SCHOOL OF ARTS AND SCIENCES

ATENEO DE DAVAO UNIVERSITY

MARCH 2005

ENHANCING AN OPEN SOURCE PDF DOCUMENT WRITER

An Independent Research

Presented to

The Faculty of the Computer Studies Division

Ateneo de Davao University

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science major in Computer Science

by

Howell C. Balolong

Aileen A. Gullos

Jan Alexander T. Lao

SCHOOL OF ARTS AND SCIENCES

ATENEO DE DAVAO UNIVERSITY

MARCH 2005

TABLE OF CONTENTS

ABSTRACT

The increasing use of Portable Document Format files has highlighted the need for better PDF document writers. There are many available PDF document writers nowadays. However, writers that have good functionalities are often proprietary.

This study concentrates on enhancing an existing open-source PDF document writer by using the most desired features of other writers. Furthermore, this study includes a survey on the users of the features that they want to be included in the enhanced writer.

The possible ways of implementing the major enhancements were also studied. One way is by using PDFMark operators. Another is by making use of third-party components such as PDF libraries. The group used PDFBox, which is an open-source Java PDF library, for the creation of bookmarks. The same library was used for the extraction or deletion of PDF pages from an existing PDF file.

Keywords:
PDF, PDF Document Writer, Java PDF Library

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The Portable Document Format (PDF) file format is commonly used nowadays. In fact, it is the de facto standard for the secure and reliable distribution and exchange of electronic documents and forms around the world.

The following statements are some of the common problems that PDF addresses. Recipients can't open files because they don't have the applications used to create the documents. With PDF, anyone, anywhere can open a PDF file using a free PDF reader. Another problem is that combined paper and electronic archives are difficult to search, take up space, and require that the application in which a document was created be available for future access. PDF files are compact and fully searchable, and can be accessed at any time using a free PDF reader. In addition, Interactive hyperlinks make PDF files easy to navigate.

A PDF document writer or creator allows users to convert printable files into their equivalent PDF files, preserving the original format of those files. However, the available PDF document writers at present are bounded by a number of limitations. These limitations include rendering of fonts and images, user interface, automatic conversion of Web addresses beginning with *http* and *www* to PDF hyperlinks and options for PDF document creation. A few of these limitations may have been addressed by existing PDF document writers which are proprietary and are therefore not available for use by all. Also, most

advanced features are only present in proprietary products. The group took the challenge of creating an enhanced writer for public domain by building on existing open-source products and learning from the strengths of proprietary products.

1.2 Statement of the Problem

This study sought to answer the general problem: How can an open-source PDF document writer be enhanced using the most desired features?

Specifically it sought to answer the following questions:

1. What are the important features of currently available PDF document writers?
2. What are the features that users are looking for in a PDF document writer and what open-source PDF document writer can be enhanced?
3. What are the current tools and libraries which support these features?
4. How can these features be incorporated in an open-source PDF document writer?

1.3 Objectives of the Study

The general objective of this research was to enhance an open-source PDF document writer using the most desired features.

The specific objectives were:

1. To identify the important features of currently available PDF document writers

2. To survey the features that users need in a PDF writer and to identify the open-source PDF document writer that can be enhanced
3. To identify the current tools and libraries that support the features
4. To incorporate the features in an open-source PDF writer

1.4 Scope and Limitation of the Study

The study focused on enhancing an open source PDF document writer. Widely used writers were evaluated and compared to identify the features that could be added to the enhanced writer. A survey was conducted wherein users were asked to rate the said features and to cite the changes that they would want to see in the new writer. Libraries that might help in the development of an enhanced writer were also studied. An assessment of open source PDF document writers was also done.

1.5 Significance of the Study

The study is significant for users of PDF writers. There are powerful PDF document writers available but most of these writers are proprietary. Some free writers may also have great features but lack other important features that are present in other writers. By creating a free writer that combines some of the important features of existing writers, users will be given a much better option in creating PDF files.

Another significance of this study is the contribution to the open-source movement. The group will be enhancing an open-source PDF writer. Through

this, the existing writer will further be developed, thus better software is created. Moreover, other programmers can make use of the enhanced writer and use it for further studies.

1.6 Definition of Terms

1. Portable Document Format (PDF) – The de facto standard for the secure and reliable distribution and exchange of electronic documents and forms
2. PDF Document Writer – a software that allows users to convert printable files into their PDF file equivalent, preserving the visual integrity of the file
3. Open-Source Software – any computer software whose source code is either in the public domain or, more commonly, is copyrighted by one or more persons/entities and distributed under an open-source license such as the GNU General Public License (GPL)
4. Electronic Document – a document that is made with an electronic agent that retained the format in which it was made, sent or received, or in a format that does not materially change the information contained in the document that was originally made, sent or received
5. Postscript Language – a computer language with the primary purpose of generating graphical images (including text) in a device independent manner. It describes these graphical images by describing how they

are drawn with lines and curves and fillings of areas, by placing bitmap images or bitmap image

6. PDFMark operator- used in PostScript code to represent PDF features
7. Ghostscript – a set of software that provides an interpreter for the PostScript language and the ability to convert PostScript language files to PDF (with some limitations) and vice versa
8. PDF library – a library that allows developers to create, read and modify PDF files