PDF TO HTML CONVERSION
Project Specification

Introduction

In recent years Adobe Portable Document Format (PDF) has become the major standard for document distribution over the Internet. With its roots in PostScript, a page-description language used by many printers and typesetters, PDF is traditionally based around the printed page, which gives it the advantage that PDF files look the same on on screen and on paper, regardless of the system on which they are being viewed.

HTML, by comparison, is worlds apart. Designed as a hypertext system, it is more similar to the file format of a word processor, with features such as different heading levels to aid navigation through the document. It offers far fewer features to control the layout and appearance of the page; features such as tables and frames were not even included in the original HTML specification. And web sites often look different between two different computers, or even on the same computer running a different browser.

While PDF is a good choice for distributing documents that are intended to be printed, or that must look the same on different systems, it is often more appropriate to publish documents in HTML format for viewing on the web, for a number of reasons:

- Less advanced users may find the distinction between the interface of a web browser and a PDF viewer confusing. Although Adobe's Acrobat plug-in for popular browsers aims to alleviate this problem, there are still many inconsistencies between the plug-in and, for example, Internet Explorer and other popular Web browsers.
- Although PDF files feature compression, they are, in general, still significantly larger than the equivalent HTML content. This raises a problem for users with slower Internet connections, particularly those who regularly "surf" without images.
- HTML files can have a "house style" applied to them to make them integrate better with an existing website for a more professional appearance.
- As the layout of a HTML file is flexible (and dependent on the system on which it is being rendered), it is much easier to edit HTML files, and even make drastic alterations, without harming the appearance of the page.

The Problem

The problem is to investigate the features that HTML provides for page layout and appearance, and how they can be used to replicate the content of a wide variety of PDF files. This includes attempting to imitate the layout of the original PDF as accurately as possible.

The investigation will result in the creation of a program that will process the PDF file and generate a series of HTML pages and images from it. Because of the nature of the file formats, it will not be expected to work perfectly with every document. The conversion program will have a number of user-selectable options to define how various aspects of the page are to be treated, depending on the source PDF file and on the intended result.

A further improvement of the program will be to build in "intelligence" to make it recognize various types of layout or image and perform the conversion in the most appropriate way, without any input from the user. Such a feature would be useful for batch processing, on a low-specification Internet terminal or even in a search engine website (*Google* allows you to display its cached PDFs in HTML format although its method of conversion strips out all the graphics).

Apart from the *Google* website, there are already two applications that attempt to solve this problem. The first one is the commercial program pdf2html by Archisoft. Part of the research in this project will be to download a trial version of pdf2html, to use it to convert a number of varied PDF files, and to see if there are areas where improvement is possible. The other solution, also named pdf2html, is an open-source program that simply uses Ghostscript to rasterize each page of a PDF file into the PNG bitmap format. It then creates HTML files that display each PNG image. Although this process ensures that the images match the PDF files exactly, the resulting files are very large and most of the benefits of the HTML format are lost.

Objectives & Methods

The objectives are divided into three phases. Apart from the dependencies detailed below, the objectives in each phase are not dependent on each other and can be worked on in parallel.

The first phase consists of preliminary research into the two formats to achieve a sound understanding of both file formats and the selection of an appropriate language.

Phase Two includes researching into the actual conversion itself, and will involve designing and writing code. Once objectives 6 to 9 are completed, objective 10 can begin, which will use some code as well as other information obtained from the previous objectives.

The project will be regarded as successful if Phases One and Two are completed. Phase Three includes extra functionality that will be worked on if there is sufficient time at the end of the project.

PHASE ONE	
1	Get familiar with PDF and HTML syntax
2	Research HTML's facilities for controlling layout and position of text and images
3	Research into graphics formats for web use (GIF, JPEG and PNG)
4	Research existing solutions to problem
5	Select a suitable programming language

PHASE TWO	
6	Research into extracting text from a PDF and placing it on a HTML page
7	Find a method to extract graphics from a PDF and convert to a suitable format and resolution for web use (this will involve finding a library or other software to perform the conversion)
8	Find a method to rasterize remaining elements of the page as graphics
9	Research into converting more complex layouts (e.g. columns and newspaper-like layouts)
10	Produce a design of the conversion software
11	Develop the conversion software

PHASE THREE	
12	Research the more advanced PDF features, such as tagged paragraphs and annotations, and their inclusion into the converted HTML
13	Make the conversion more "intelligent" by autodetecting certain settings depending on the type of document being converted
14	Modify the conversion software to accept a PostScript file as input

I intend to make use of existing open-source libraries where possible; for example to convert the graphics into a lower-resolution format appropriate for Web use, for previewing the PDF and HTML files and for creating the actual user interface of the converter.

Provisional Timetable

Week	Objectives 1	2	3	4	5	6	7	8	9	10	11	12	13	14	Other
1															Project specification
2															
3															
4															
5															
6															
7															
8															
9															Progress report
10															
CHRISTMAS VACATION															
11															
12															
13															Attend WWP Lectures
14															
15															
16															
17															
18															
19															Present -ation
20															
EASTER VACATION															
21															Final report
22															

Resources

I intend to use the following hardware:

- Home computer (Athlon XP 1800+ with 512MB RAM) running Windows XP Professional and Red Hat Linux 8.0
- Computer Science Department workstations (Sun Ultra 5s running Solaris and Pentium-IIIs running Red Hat Linux 7.0)

The extent to which I use the Department's hardware depends on the language selected and its availability of Unix/Linux compilers or interpreters. Some tasks, such as researching Archisoft's pdf2html software, will need to be carried out on my personal computer.

As both sites have a permanent Internet connection, it will be easy to transfer files between them at regular intervals in case of fire, theft or other unforeseen circumstances.