

## Search Engine PDF Indexing

A Datalogics, Inc. White Paper

---

### Problem

One of the most common uses of the Adobe PDF Library is to extract text from PDF files. These PDF files come from a variety of sources and in a number of forms, but the PDF Library can handle all validly constructed PDF files.

When using the PDF Library to index PDF files from internet sources, however, a common problem emerges. Many web sites will not allow search engines to download more than the first N kilobytes of any type of file, including PDF files. Though many PDF files may fit within this limit, there are many that cannot. Downloading the first N kilobytes of a PDF file will truncate PDF files that do not fit within this size limit and result in invalid PDF files being obtained. These invalid PDF files are not able to be opened by the PDF Library. Search engines and other clients that wish to index these must take an alternate approach.

This paper describes an approach for indexing the text from the first page of a PDF file through incrementally downloading portions of the PDF file automatically using the Adobe PDF Library.

### Background

PDF files are comprised of four components – a header, body, cross-reference table and trailer – present in the file in that order. All components must be present for a PDF file to be syntactically valid. Some PDF files, known as *linearized* PDF files, contain a miniaturized cross-reference table for the first page and other information for efficient random access to pages over slow links. Most linearized files will additionally pack the objects for the first page of the PDF close to the start of the file for efficient first-page access, though this is not true for all linearized files.

Within the PDF file body are a series of numbered objects. Each object represents either a portion of the content of a PDF or a resource (a font, image, etc.) used by the content of the PDF file. These objects are present in arbitrary order in the PDF file – for example, all of the fonts and images used in a PDF may be present in the file before the content of any of the pages. Page contents may not be in display order in the PDF file; these can be stored in any order, can be mixed in amongst the resources used for the PDF pages

and can even be spread out such that the contents of any PDF page exist in many different locations in the PDF file.

Because of this file format flexibility, it is impossible to take only part of a PDF file and open it as a valid PDF for text extraction or other purposes. This is not a suitable general-purpose strategy for indexing PDF files.

## Solution

The Adobe PDF Library affords users the flexibility required for search engine indexing by allowing users to randomly access PDF datastreams on remote servers. Through this, search engines can avoid downloading PDFs to intermediate files; instead, only the data necessary for text extraction is requested.

Here's how: the PDF Library supports user extensions to its PDF input and output routines. The PDF Library ships with a default set of input and output routines to read and write PDFs to files – called a *filesystem*, though not limited to using files. Through this facility, PDF Library users can work with PDF documents stored in memory blocks, in databases and from other sources including web (http:) servers. Users take advantage of this facility through easy to craft extensions of the input and output routines used by the PDF Library.

For the search engine indexing PDF files, this solution is ideal. The user supplies a filesystem that directs PDF open and read requests through URL data transfer functions such as libcurl or OS-specific APIs. Using this filesystem, access to the PDF is carried out through the same API calls used for working with a PDF file on disk; no special coding is required.

The PDF Library, using this filesystem, issues requests to the remote server for only those portions of the PDF that are needed to fulfill the user's text extraction requests. No intermediate download of the PDF to a physical file is required, and only the data needed for extracting the text from the pages of the PDF is requested from the remote server.

*Author: Matt Kuznicki  
April 14, 2009*

For questions or more information, please contact us at:

Datalogics, Inc.  
101 N. Wacker Drive #1800  
Chicago IL 60606 USA  
+1 312 853 8200  
<http://www.datalogics.com>

Copyright (c) 2009, Datalogics, Inc. All Rights Reserved.

Datalogics, the Datalogics Logo and all Datalogics product names are either trademarks or registered trademarks of Datalogics, Inc. All other trademarks are the property of their respective owners. Reproduction of this document in whole or in part without the express written consent of Datalogics, Inc. is prohibited.

This document and related materials and information are provided "as is" with no warranties, express or implied, including but not limited to any implied warranty of merchantability, fitness for a particular purpose, non-infringement of intellectual property rights, or any warranty otherwise arising out of any proposal, specification, or sample. Datalogics, Inc. assumes no responsibility for any errors contained in this document and has no liabilities or obligations for any damages arising from or in connection with the use of this document.