Thurow, S. (2007). *Search Engine Visibility, 2nd edition.* New Riders. (revised 2016)

Optimizing PDF documents

Search engines have become increasingly efficient at indexing different types of documents. Google, for example, can index 13 types of documents (in addition to HTML formatted documents) that include Microsoft Word, Microsoft Excel, Microsoft PowerPoint, rtf (rich text format), and Adobe PDF documents. Other search engines can index PDF documents as well.

PDF stands for portable document format, which is a universal file format that preserves fonts, colors, graphic images, and formatting of any source document. Many Web site owners like to create marketing brochures, media kits, and how-to manuals in PDF format and make them available on the Web. Figure 2-48 shows a typical Web page brochure formatted as PDF.

Many Web site owners like to have PDF documents on their Web sites because they want to preserve the exact look and feel of a printed piece. For example, let's say you would like your online brochure text to display in the typeface Avant Garde. In order for the online brochure to appear in this typeface, your site's visitors must have the Avant Garde font installed in their computers. If your visitors do not have this font installed, your online brochure will look different than what you intended. Therefore, many online brochures are formatted as PDF documents.

PDF documents can achieve top search engine visibility when formatted correctly. In fact, some top search engine results are PDF documents as shown in Figure 2-49.

The building blocks of successful PDF document optimization are the same ones for HTML file optimization:

1. **Text component.** PDF documents should contain the words and phrases that targeted searchers are likely to type in to search queries. In other words, PDF document text should use the user's language.

2. **Accessibility component.** Search engines and site visitors should have easy access to PDF documents. Additionally, PDF document content should also communicate relevancy, a sense of place, and clear information scents.

3. **Popularity component.** Encouraging external, third-party link development to PDF documents.

4. **Behavior component.** PDFs satisfy both informational (learn or read) and transactional goals (download, print, send attachment in an email).

KEYWORD-RICH TEXT

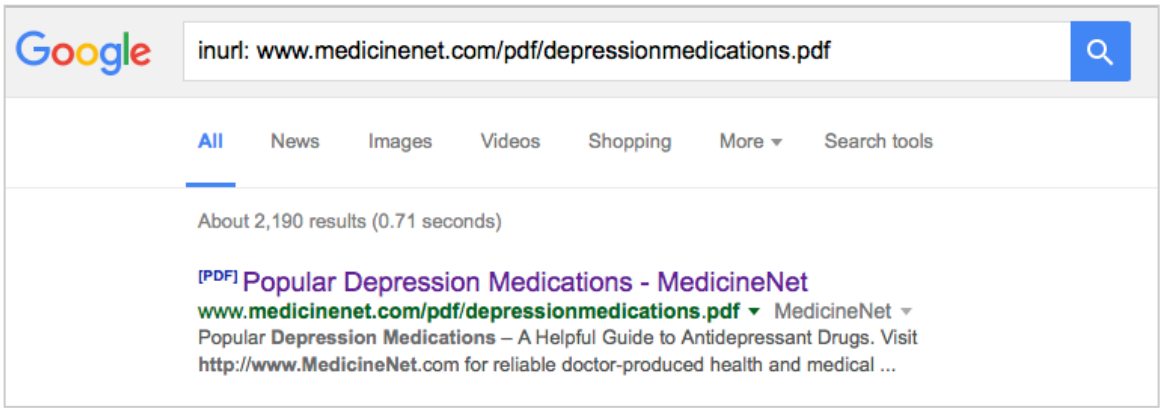
In order to make your PDF documents search friendly, the documents must contain actual text, not a picture of text. There are multiple ways to determine whether a PDF document contains text that Web search engines can index:

1. Inurl: and cached feature in search engine results pages. Copy-and-paste.
2. Document properties attribute in Adobe Acrobat and Acrobat Reader.

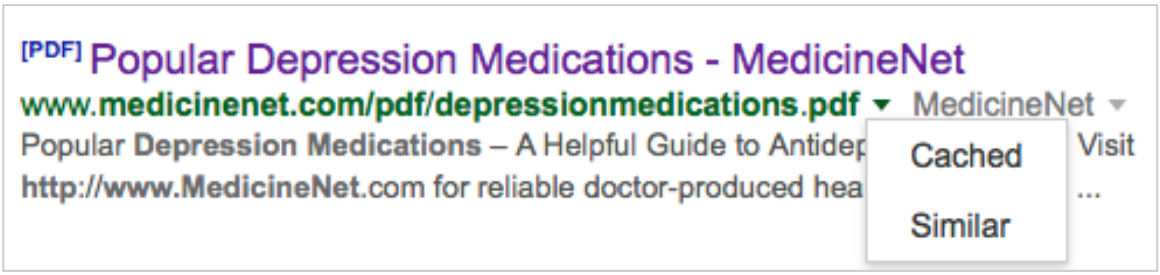
Inurl: and cached feature

If your PDF document is included in a search engine index, you can use the “View as HTML” feature to see the text that search engines are using to determine relevancy.

First, perform an inurl: type of search so that the PDF listing appears on a search results page, as shown in Figure 2-50, using Google as an example:



If you click on the Green arrow and select the “cached” option, you can see the search engine friendly text within a PDF document.



Then:

1. Select All (Command-A on a Mac, Control-A on a PC).
2. Copy (Command-C on a Mac, Control-C on a PC).
3. Paste special (unformatted text) into a text editor. Use Control-V (PC) or Command-V (Mac) to paste the selected text into the text editor.

This is the text used to determine relevancy within a PDF document.

Document properties attribute

Another way to determine if a PDF document contains text the search engines can index is to check the Document Properties dialog box. If no fonts are displayed in the Document Properties dialog box, then the PDF document does not contain text.

To check for fonts in your PDF files:

1. Open the PDF document in Acrobat or Acrobat Reader.

2. Select File > Properties. The Properties dialog box should appear with a file tab labeled Fonts, as shown in Figure 2-52. If any fonts appear in this dialog box, the PDF document contains text the search engines can index.

Tip: Avoid use of Type 3 fonts in PDF files, because they're often generated with missing or incorrect font size and character encoding information. (Source: https://scholar.google.com/intl/en-us/scholar/inclusion.html#indexing)

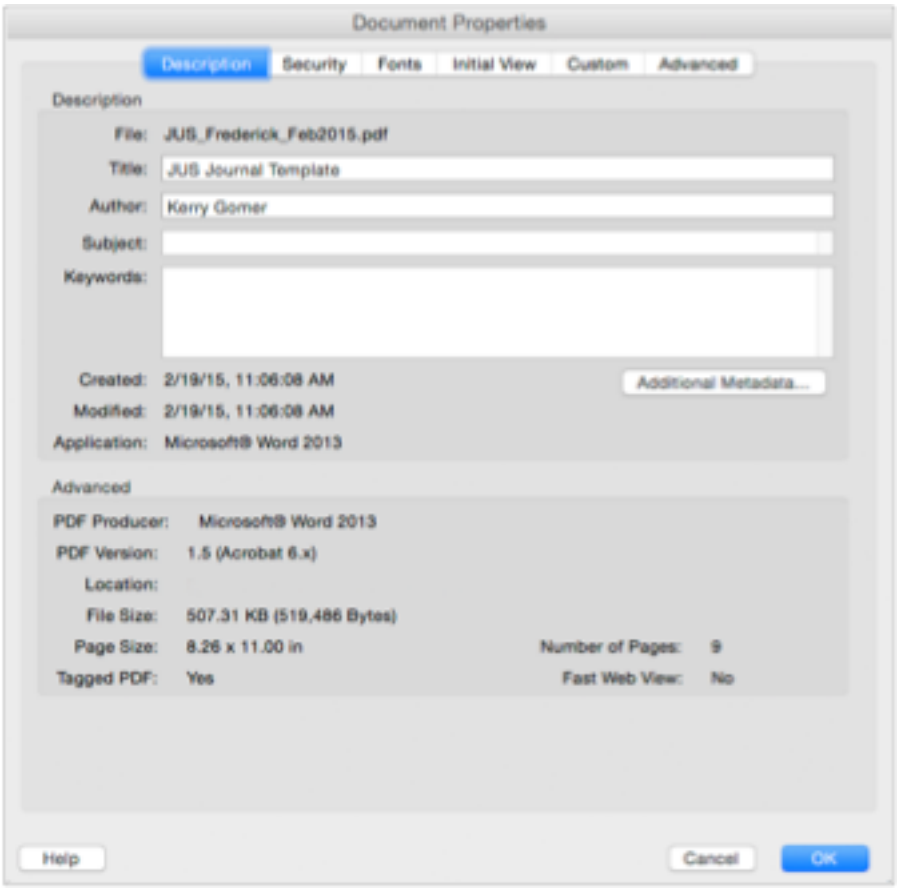
To see the specific text the search engines are able to index, use the copy-paste method previously mentioned.

PDF metadata

Adobe Acrobat does allow PDF creators to add metadata to PDF documents.

1. Open the PDF document in Acrobat or Acrobat Reader.

2. Select File > Properties. The Properties dialog box should appear with a file tab labeled Description, as shown in Figure 2-52. Add Title, Author, Subject and Keywords metadata.



Note: Since official Google PDFs contain metadata, I highly recommend including metadata in PDFs you wish to be featured in SERPs.

Currently, search engines do not use a PDF document's metadata information to determine relevancy. Like the meta-tag description and keywords attribute in HTML files, PDF metadata is considered secondary text because search engines are able to access text content within a PDF document.

URL structure

A PDF's URL structure should be descriptive and straightforward. For example what does this URL structure communicate?

http://www.medicinenet.com/pdfs/depression-medications-popular.pdf

Users can easily determine that the PDF is about popular depression medications on the domain medicinenet.com.

Now look at this URL. What does it communicate?

http://uxpajournal.org/wp-content/uploads/pdf/JUS_Frederick_Feb2015.pdf

Here are some URL guidelines for PDFs:

- Short but descriptive.
- Contains important, relevant keyword phrases without keyword stuffing.
- Makes sense to human users.
- All lowercase (optional, but be consistent).
- Reinforces the document title or primary headline.
- Use hyphens (not underscores) to separate words.
- Minimize the use of stop words (a, an, the, or, nor, for)

PROVIDING ACCESS TO PDF CONTENT

Two reasons many PDFs with outstanding content do not receive qualified search engine traffic are lack of access and orphaning.

Lack of access

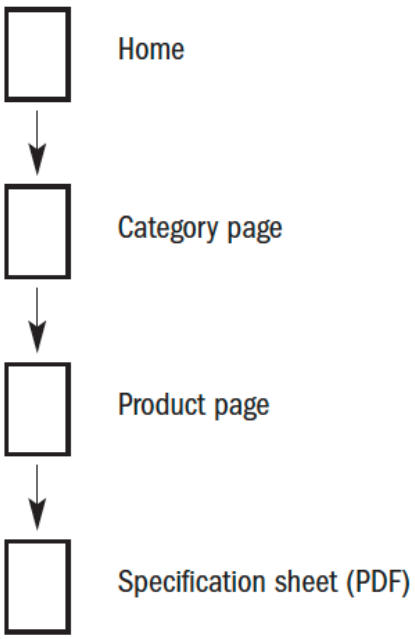
Many Web site owners will only provide access to white papers and brochures after site visitors fill out a form. The reason they do this is lead generation. In exchange for site visitors' personal information (name, email address, address, phone number, etc.), site visitors can download useful and informative white papers, for example.

Search engine spiders do not fill out forms to access Web content. So if the only way to access PDF content is through forms, the PDF document's listing will not appear in search results.

Orphaning

Many Web site owners do provide access to PDF documents. However, they do it in a way that communicates, "I do not believe this content is important."

For example, many Web site owners link to PDF content in the following hierarchical, linear manner:



All too frequently, the only link on a Web site to a PDF document comes from an individual product or service page, essentially orphaning the PDF document. This one-way information architecture and corresponding interface communicates to the search engines that you (the Web site owner) must not believe the PDF's content contains valuable information. If you believed the content were valuable, then you would link to the URL in more than one place.

One simple way to create additional links to PDF documents is to add them to a wayfinder site map. Many search engine professionals, and even search engine representatives themselves, often feel that an XML sitemap and a wayfinder site map are magic solutions for providing access to content. However, as I outlined in previous sections of this book, navigation is only one part of the accessibility component. Relevancy, a sense of place, and information scent are all equally important. Usability counts with both contextual/supplemental navigation and external, third-party link development.

Always communicate important information to site visitors before they click a link to a PDF document. First, let people know that if they click a link they will be viewing a PDF document. Acrobat Reader (a different application) will be launched if it is installed on end users' computers. Launching a new application when people do not expect it leads to a poor user experience.

Second, because PDF file size tends to be significantly larger than HTML file size, let users know the file size before they click the PDF link. Remember, meeting user

expectations leads to a positive search experience.

Finally, link to PDF documents using important keyword phrases in or near anchor text. Including only hypertext links in a wayfinder site map will help, but it is not an ideal solution for search engine optimization.

A better solution is to summarize each PDF document's overall content, using appropriate keyword phrases whenever possible. For example, on the fictional TranquiliTeas Web site, a simple way to let visitors know they will be viewing a PDF document is to make the hypertext link look like the following:

View the TranquiliTeas Organic Tea Brochure—PDF (360K).

Some Web site owners like to create subcategory pages with summaries to each PDF document. They add the category page links to the wayfinder site map (or site index) as well as making the subcategory page a part of a site's global navigation scheme.

Robots excluding redundant PDF content

Sometimes, the same content on a Web site is formatted as HTML and as PDF. To avoid duplicate/redundant filtering, use the robots exclusion protocol, the robots.txt file specifically, to let the search engines know not to crawl the redundant content.

One way I implement PDF optimization is to put redundant content PDFs in one directory and original content PDFs in another directory.

For example, the redundant content PDFs might be listed under a directory category labeled "pdfs" and the original content PDFs might be listed under a directory category labeled "pdf." I only apply the robots exclusion protocol to the redundant content PDFs, as shown below:

User-agent: *
Disallow: /pdfs/

PDF LINK DEVELOPMENT

External, third-party link development to PDF documents might be more difficult to implement due to the file format. Most people expect a link to deliver an HTML formatted document, not a PDF document.

Nonetheless, PDF documents with unique content can receive high quality links, especially white papers.

Machine speed, also known as page speed, has been a ranking factor for many years.

Likewise, users that want to read PDFs on a mobile device don't want to wait for the content to download. People don't link to or cite content that is difficult to download.

Therefore, PDF file size should be minimized. Some ways to minimize PDFs include:

- Limit the number of fonts/typefaces used in the document.
- Limit the number of graphic images used in the document.
- If necessary, reformat graphic images to be smaller in file size. You can put the higher-resolution images on your site individually.
- Limit PDF metadata to essential information (author, title, description, keywords, URL). Don't put long paragraphs of information in the metadata.

10 steps to successful PDF optimization

The same optimization guidelines apply to PDF documents that apply to HTML documents.

1. **Make sure your PDF documents contain text that the search engines can index.** Search engines are currently unable to index the text inside graphic images in PDF documents. So if you create a PDF document by using a flatbed scanner, making it an image-only PDF, the search engines will not be able to extract that text.

2. **Use keyword-rich text in your PDF documents.** The main advantages of optimizing PDF documents is that they tend to be text-heavy documents, and their URL structures are quite simple. A little bit of keyword research and keyword placement can result in higher PDF visibility.

3. **For PDF documents with multiple pages, ensure that the most important text is on the first page of your PDF document.** Be sure that the titles, headlines, and text on the first page of your PDF documents contain your most important keywords, when appropriate.

4. **Minimize download time.** In general, search engine representatives recommend keeping document file size to less than 100K, mostly for usability reasons. PDF documents are considerably larger files because fonts are embedded in them and they often contain high-resolution elements, such as photos and illustrations. Two ways to minimize PDF download time are to limit the number of fonts used and to use lower resolution graphic images in Web-only PDF documents.

5. **When appropriate, create optimized HTML pages with abstracts of PDF documents.** If your PDF documents are large file sizes, such as manuals or catalogs, consider creating HTML pages that summarize the PDF files. The abstract pages should contain 200 to 250 words of quality content within the

<body> and </body> tags. Title tags and meta tags should also contain keywords.

Additionally, whenever possible, the anchor text leading to the PDF document and words near the anchor text should contain keywords.

6. **Be sure to have links to your PDF documents in multiple places on your Web site.** Do not orphan your PDF documents in the site's site navigation system. Add links to PDF documents in your wayfinder site map or site index. Add links to PDFs in contextual navigation.

   If your Web site contains many PDF documents, categorize them and create a PDF or online publications section with a topical site map. Communicate to both search engines and site visitors that you believe your PDF documents' content contains useful, relevant information.

7. **Robots exclude redundant content.** Since the commercial Web search engines have been able to index PDF documents and many other text-based documents for a long time, creating the same content in different formats will only lead to duplicate content filtering and a lower index count. Be proactive. Exclude redundant PDFs using the robots.txt file.

8. **Accommodate both informational and transactional searcher goals.** If you want people to download your PDFs, then use the word "download" within anchor text. Always let users know that they will be viewing a PDF when they click on the link to it. Minimize download time.

9. **Do link prospecting and outreach.** People often link to PDF guides that are useful and informative. Remember, you have more control over the look-and-feel of a PDF. A guide with illustrations and checklists can help many users feel your organization is more trustworthy.

10. **Be realistic.** Not all PDF materials (such as trifold marketing brochures and annual reports) can use keyword-focused text. Some PDF documents should not have a table of contents, headers, and footers. Understand which types of PDF documents can easily be optimized and which types cannot. Focus your optimization efforts on the PDF documents that can be optimized.