

## Internet Publishing Review

by Péter Jacsó

# Tools for Unearthing PDF Files

Google provides a superior Adobe PDF searching experience

There has been much talk about the Invisible Web in the past 12 months. Beyond the databases—even the freely accessible ones—that generate a temporary HTML output in response to a query of a collection of records stored in a proprietary format, Adobe PDF files represent a large domain of the Web that is invisible to traditional search engines, and thus, to users.

“  
The latest innovation by Google will significantly enhance access to worthy sources of information by handling the PDF files that are available to the public.  
”

The latest innovation by Google will significantly enhance access to worthy sources of information by handling the PDF files that are available to the public. This is an important development as an increasing number of sites include PDF documents, and often without any accompanying HTML file that would at least alert users about the availability of PDF files that are relevant to their research.

But haven't there been search engines that could extract keywords from PDF files? Yes, there have been, but there is a big difference. Those are engines that can be used for your own Web site or intranet, but they are not Web-wide crawlers that are sent out to the Web to collect information about PDF documents. Seemingly, Adobe's Search PDF site came the closest to the service that Google has recently developed, but a closer look at both reveals Google's superiority.

### The Adobe Search PDF Site

Adobe teamed up with AltaVista last year to create a special site (<http://search.pdf.adobe.com>) that is believed to include information about more than 1 million PDF files. However, this number should be taken with a grain of salt, as all of my test searches brought up many duplicate items. You're not searching the entire text of the PDF files, but only the text of the title, abstract, and keywords as extracted by Adobe from the PDF files.

The extraction is not always successful even from text files, let alone from PowerPoint PPT files or spreadsheet files that were converted to PDF. There are often serious problems in every aspect of the extraction process. The summaries can be very cryptic if you don't know the document (see Figure 1).

The data for the number of pages and the size of the files are often mixed up. For example, Charles Bailey's classic bibliography keeps growing, to our pleasure, but not to the tune of 258,206 pages. And, of course, its size can't be 88 bytes, and it can't be downloaded in 0 seconds. The data are true in reverse: 88 pages and 258,206 bytes. On a 56 Kbps modem it would take some time to download. This is not an exceptional mistake but a very common one at this site.

Also, the file dates are often wrong, and off by about three decades. You'll see a very large number of files purportedly created on the last day of 1969, when there was no such thing as PDF. This suggests that the dates are taken from the file description, not the PDF file itself. I recall that on some older computers, the file-creation date automatically reverted to 1969-12-31 if you did something that Microsoft didn't exactly encourage you to do.

Often, it isn't the title of the PDF document (such as Electronic Dissemination of Scholarly Work) that is extracted and displayed as title information, but the title of the journal (see Figure 2) or some other information extracted from the PDF page.

The site's search options are simple but sufficient; however, being taken to the wrong help file isn't exactly professional or user-friendly. The partner for this project

was AltaVista, which has Basic (or Main) Search, Advanced Search, Power Search, and Raging Search options. If you take the advice of the help file, you'll be surprised.

While the help file tells you that the query *mona lisa* yields the same result as *+mona +lisa* (see Figure 3), this isn't true (except for the Advanced Search mode, which is not available for Search PDF). The query *scholarly electronic publishing* finds 15,643 PDF documents. If this seems too good to be true, it is. The Basic Search mode assumes that documents that match ANY of the words in your query are a hit, so you must use the plus sign (+) to indicate the words that must be present. The query *+scholarly +electronic +publishing* retrieves 104 items, and the exact phrase search "scholarly electronic publishing" (that is, between quotes) finds 13 items.

Even with these limitations, and in spite of the rather sloppy implementation, it was worth using Search PDF (smartly, that is) because it could find documents that traditional search engines couldn't. But earlier this year it ceased to be the only show in town when Google introduced its PDF search option.

### Google's Search Enhancement

If you didn't use the Web for a day or two back in February you may not have heard about this new feature in Google. Other developments by Google—like the purchase of Deja.com; the quick and simple phone number look-up; and the automatic presentation of English translations of French, German, Spanish, and Portuguese pages—

“  
Adobe PDF files represent a large domain of the Web that is invisible to traditional search engines, and thus, to users.  
”

overshadowed the PDF feature, which is definitely worth your attention.

There are several features that put Google's PDF search far above the other PDF search options. One is that it started with 13 million PDF documents—this makes you sit up and take notice immediately. Next is that finding PDF documents is part of the main Google search, not part of a separate site or process. The third is that Google searches the PDF documents in their entirety for your query term(s). And the fourth is that it offers to display the ASCII-text version of the PDF document as an option.

**Document Title:** Scholarly Electronic Publishing Bibliography  
**Author:** Charles W. Bailey, Jr.  
**Date:** 1969-10-01 12:54:31 **Pages:** 258206  
**Download:** 0 sec **Size:** 88 bytes  
 \* Download time assuming 56 kbps modem.

**Summary:** URL: http://info.lib.ohio.edu/sepb/sepb.coc<br> URL: http://info.lib.ohio.edu/sepb/sepb.html<br> URL: http://info.lib.ohio.edu/sepb/sepb.pdf<br> The bibliography is also available as an HTML file. URL: http://info.lib.ohio.edu/sepb/sepb.html<br> The HTML version can be searched, and it includes Scholarly Electronic Publishing Resources, a collection of links to related Web sites that deal with scholarly electronic publishing issues. HTML file URL: http://ftp.princeton.edu/pub/harnad/Harnad/HTML/harnad95.coc<br> vae.s.html HTML file URL: http://ftp.princeton.edu/pub/harnad/Harnad/HTML/harnad97<br> learned.sena.s.html HTML file URL: http://ftp.princeton.edu/pub/harnad/Harnad/HTML/harnad96<br> peer.review.htm<br> HTML file URL: http://ftp.princeton.edu/pub/harnad/Harnad/HTML/harnad92<br> interactive.js.html HTML file URL: http://ftp.princeton.edu/pub/harnad/Harnad/HTML/harnad97<br> antiquity.html HTML file URL: http://ftp.princeton.edu/pub/harnad/Harnad/HTML/harnad90<br> skywriting.html

Figure 1

**Document Title:** Journal of Information Technology Impact  
**Author:** Stamos Karamouzis  
**Date:** 1969-12-31 16:00:00 **Pages:** 8  
**Download:** 4 sec **Size:** 29233 bytes  
 \* Download time assuming 56 kbps modem.

**Summary:** Electronic Dissemination of Scholarly Work Stamos University New Orleans Louisiana, U.S.A. Abstract<br> scholars have developed and used a variety of media work. This article examines journals as a means of work. Finally, it describes the Journal of Information online scholarly journal that provides a forum for exploring social impact of information technologies. Keywords: journals, disseminating scholarly work, Journal of Information Impact. Finally, the last section describes the Journal of Information Impact, an online scholarly journal that provides a forum for information on the social impact of information technology growth in scholarly publishing coupled with the increasing subscriptions is creating tremendous economic pressure and agencies that traditionally support scholarly activities. academic libraries that struggle with journal costs a what academics need to keep up-to-date with their (Kiernan, 1998)

Figure 2

mona lisa	Finds documents that contain <b>both</b> mona and lisa, including capitalized variants (Mona, MONA, lisa, Lisa). AltaVista ranks the results to show first the documents containing the words close together, and near the top of the document.  When you do not include a plus ("+") or minus ("-") in front of any of the terms, the terms are interpreted as having a plus in front of each of them. In this example, "mona lisa" is equivalent to the "+mona +lisa" example below.
Mona Lisa	Finds documents that have both Mona and Lisa but not any other capitalized variation. When you use a <b>capitalized word</b> , AltaVista assumes that you are only interested in an exact match.
+mona +lisa	Finds only documents that contain <b>both</b> words. Be sure there is no space between the plus sign and the word.

Figure 3

PDF: [andrew.treloar.net/Research/Theses/PhD/Hypermedia.pdf](http://andrew.treloar.net/Research/Theses/PhD/Hypermedia.pdf)  
 Research/Publi- cations/APWWW95/ Treloar, A. (1995) **Scholarly electronic publishing and the world-wide web** Pro- ceedings of AusWeb95 (the first Australian ...  
[Text version](#) - [Similar pages](#)  
[\[ More results from andrew.treloar.net \]](#)

**Electronic Scholarly Publishing Bibliography**  
 ... General Interest Bailey, Jr, Charles N. "Scholarly Electronic Publishing ... books, and electronic documents ... and new publishing models. ... sepb.pdf" Darnton ...  
[www.nyuliblibrarianship.org/infob99.htm](http://www.nyuliblibrarianship.org/infob99.htm) - 8k - [Cached](#) - [Similar pages](#)

[epub Mailing List Archive, Fwd, Version 21, Scholarly ...](#)  
 ... 12:35:37 - 0700: ... are useful in understanding **scholarly electronic publishing** > efforts on the Internet ... [edu/sepb/sepb.pdf](http://www.sims.berkeley.edu/colours/sus290-7/98/epub/0039.htm) > Word: <URL:http ...  
[www.sims.berkeley.edu/colours/sus290-7/98/epub/0039.htm](http://www.sims.berkeley.edu/colours/sus290-7/98/epub/0039.htm) - 7k - [Cached](#) - [Similar pages](#)

**Excited about Electronic Publishing**  
 ... Notes 1 A fundamental resource ... and Ownership on **Scholarly Publishing**, College & ... 1996 On the ... **Journal of Electronic Publishing** 5/2 ...  
[www.chief.ac.uk/university/academic/CA/C0/bib/02/CA/Courses/Batore.html](http://www.chief.ac.uk/university/academic/CA/C0/bib/02/CA/Courses/Batore.html) - 39k - [Cached](#) - [Similar pages](#)

**Electronic Scholarly Publishing, Today's Technical Alternatives**  
 ... While slower to develop ... growing areas of publishing, with several ... of an **electronic scholarly journal** there are ... Document Format (PDF), Common Ground's ...  
[www.cob.up.edu/~pbl/Ed/Ed03/EducomEP/tech.htm](http://www.cob.up.edu/~pbl/Ed/Ed03/EducomEP/tech.htm) - 7k - [Cached](#) - [Similar pages](#)

PDF: [scout.cs.wisc.edu/report/sr/pdf/sr961122.pdf](http://scout.cs.wisc.edu/report/sr/pdf/sr961122.pdf)  
 The Scout Report... an "understanding [of] **scholarly electronic publishing** ... issues, New Publishing Models, and ... Space Telescope Electronic Information ... of pdf files ...  
[Text version](#) - [Similar pages](#)

Figure 4

The mainstreaming of PDF hits in the results list is perfect (see Figure 4). You see the familiar short entries on the results list with a small PDF identifier at the very beginning. If you click on it, the PDF document will be displayed. However, if you have a slow connection or are afraid that it could be a large PDF

document that would take minutes to display, even on a cable modem, you can sample it in plain ASCII format, which is instantly displayed when clicking on the Text version hotlink in the short entry (see Figure 5).

The display is instantaneous because Google generates the text version while

7.3	Survey Results	141
7.3.1	Basic demographics	141
7.3.2	Access to technology	143
7.3.3	Use of <b>electronic publishing</b> technologies	149
7.3.4	Advantages of <b>electronic scholarly publishing</b>	156
7.3.5	Disadvantages of <b>electronic scholarly publishing</b>	167
7.4	Conclusion	176
Library Case Studies		
8.1	Introduction	177
8.2	Case study research	177
8.2.1	Overview	177
8.2.2	Case study issues	178
8.2.3	Designing the case study	178
8.2.4	Data collection	180
8.2.5	Data analysis	181
8.3	Higher Press	182
8.3.1	Overview	182
8.3.2	Origins and organization	182
8.3.3	Financial sustainability	183
8.3.4	Products	183
8.3.5	Lessons learned	184
8.3.6	Future prospects	185

Figure 5

## Premedia Technologies Forms Partnership with ScreamingMedia

R.R. Donnelley & Sons Co.'s Premedia Technologies unit and ScreamingMedia, Inc. (<http://www.screamingmedia.com>), a global provider of content solutions, have announced that they have formed a partnership that offers magazine publishers the opportunity to transform print content into electronic content and to distribute it through private or general syndication channels to meet subscribers' needs for specific information.

Using Premedia Technologies' services, publishers can convert content from application files such as Quark XPress into formats like XML or HTML. Publishers can then utilize ScreamingMedia's Syndication!Connect services and its technology platform to parse, normalize, categorize, and integrate it directly into the platforms of their subscribers, or syndicate their con-

tent through ScreamingMedia's global digital content network.

"Our publishing customers are the real winners in this partnership," said Mary Lee Schneider, president of Premedia Technologies. "By partnering with ScreamingMedia, they can develop their own private syndication networks directly with subscribers or have immediate access to all the Web and wireless clients in ScreamingMedia's network. We provide the means to convert and store their content, and ScreamingMedia provides the filtering and integration technology. As a result, publishers effortlessly gain opportunities to extend the value of their content while at the same time developing new revenue streams."

Source: R.R. Donnelley & Sons Co., Chicago, 312/326-8000; <http://www.rrdonnelley.com>.

Advanced Search Preferences Search Tips

# Google

[inurl:pdf "scholarly electronic publishin](#)

[inurl:pdf "scholarly electronic publishing"](#)

PDF: [scout.cs.wisc.edu/report/sr/pdf/sr961122.pdf](http://scout.cs.wisc.edu/report/sr/pdf/sr961122.pdf)  
 ... understanding [of] **scholarly electronic publishing** ... Issues, New Publishing Models and ... Space Telescope Electronic Information Service ... viewing of pdf files from ...  
[Text version](#) - [Similar pages](#)

PDF: [www.vtt.fi/inf/nordep/proceedings/epsem96/FIN.PDF](http://www.vtt.fi/inf/nordep/proceedings/epsem96/FIN.PDF)  
 ... Recent trends in **scholarly electronic publishing** David F ... protog for automated PDF links using delayed ... Journals are: **Electronic Publishing (EP-odd** ...  
[Text version](#) - [Similar pages](#)

PDF: [www.unc.edu/~ors/publications/fpcw6\\_generallibrary.pdf](http://www.unc.edu/~ors/publications/fpcw6_generallibrary.pdf)  
 ... St. Clair, Dennis, Nancy K. "Back to the Future: At Last Librarians Chart a New Course in **Scholarly Electronic Publishing**." *Against the Grain*, Vol. 9, No. 5 ...  
[Text version](#) - [Similar pages](#)

PDF: [zeus.slais.ucl.ac.uk/idb/ee/edline/bookmark.pdf](http://zeus.slais.ucl.ac.uk/idb/ee/edline/bookmark.pdf)  
 ... Liberties Union) 1.8 Oxford Text Archive 3.16 acronyms 3.42 **Scholarly Electronic Publishing** 3.26 AGPS (Australian Government Publishing Service) Studies in ...  
[Text version](#) - [Similar pages](#)

Figure 6

crawling the Web, not when you click on the hotlink. It may not be perfect typographically, but it's more than adequate to see how relevant a document is by scanning, for example, its table of contents pages—even if the page numbers don't always line up perfectly. The matching query terms are automatically highlighted by Google (using different colors for the different words in the query). The matching criteria are strict, so hyphenated versions (such as "publish-ing") are not highlighted on the page.

If you rely only on *The Guardian* for technical information you may accept that "there does not seem to be a way to limit a Google search to PDF files." Well, there is. All you have to do is add to the usual query the qualifier *inurl:pdf* and your results will be limited to PDF documents. While the query for the exact phrase "scholarly electronic publishing" finds 4,060 items (both HTML and PDF documents), the query *inurl:pdf* "scholarly

electronic publishing" will retrieve 85 PDF documents (see Figure 6).

Considering that Google is the search engine of choice for many searchers, and it is one of the three largest search engines, users should benefit from this PDF search enhancement. Many government documents, research papers, and legal regulations—as well as a great number of forms—are available only in PDF format. Among the Web-wide search engines, only Google can find them. All these features together make Google's development an excellent idea that is also superbly implemented. Among the many failing dot-coms, Google keeps flourishing—to the delight of its users.

*Péter Jacsó is associate professor of library and information science at the University of Hawaii's Department of Information and Computer Sciences. His e-mail address is [jacsop@hawaii.edu](mailto:jacsop@hawaii.edu).*

## Arbortext Expands Open Standards Support with Its New Epic Editor 4.2 Software

Arbortext, Inc., a provider of XML-based software for publishing to multiple media, has announced the upcoming release of the next version of its Epic Editor software. Epic Editor 4.2 will offer expanded support for software developers who require the openness and power of JavaScript, Document Object Model (DOM), and eXtensible Stylesheet Language (XSL). According to the announcement, with version 4.2, Arbortext continues its support for open industry standards, allowing software developers to continue using the programming languages they are most comfortable with, while still being able to publish from a single source to multiple media types, including Web, print, CD-ROM, and wireless.

ware as required by their organization," said David White, director of product marketing at Arbortext. "Especially for the medium and large companies that form the bulk of our customer base, we're always striving to satisfy their need to combine the stability of open standards with the power of a software product that scales for current needs and future requirements. Epic Editor 4.2 was developed to provide application developers the ability to achieve that goal."

Epic Editor 4.2 will ship in July. All Arbortext customers under maintenance are eligible to receive the Epic Editor 4.2 upgrade at no charge; the upgrade will be automatically shipped to them.

Source: Arbortext, Inc., Ann Arbor, MI, 734/997-0200; <http://www.arbortext.com>.