acrobat catalog:

creating the keys to instant access

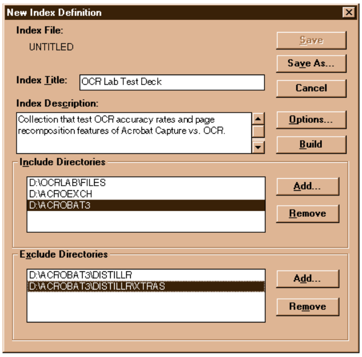
chapter five

Transforming Vast Collections Into Retrievable Information

Acrobat Catalog is the indexing process that converts a large collection of electronic files into a coherent and searchable database of documents. The process itself is simple: point the Catalog application at a directory or a set of subdirectories of PDF files, and the software does the rest.

The end result is an index of the entire collection that can be accessed by the Search program. The index and the files can reside on a network or Web server. Or the entire collection, along with a licensed copy of the search engine and Acrobat Reader, can be published on CD-ROM.

Acrobat Catalog and Search employ Verity text-search technology to index and retrieve information in PDF collections. This means that a single desktop user of Acrobat 3 can create very large libraries of files that can be searched with all of the information-retrieval capability of the Verity search engine. And this single user can publish collections on networks, CD-ROM and the Web. Compared to typical HTML hyperlinked collections, large indexed databases offer more stable long-term performance because there are no links to be broken. As long as the index is not changed, large collections will be reliably accessible through both index and full text-search and retrieval.



New Index Definition

Index File: UNTITLED

Index Title: OCR Lab Test Deck

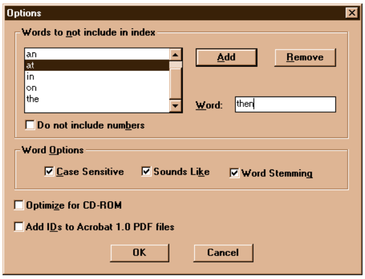
Index Description: Collection that test OCR accuracy rates and page recomposition features of Acrobat Capture vs. OCR.

Include Directories:
D:\OCRLAB\FILES
D:\ACROEXCH
D:\ACROBAT3

Exclude Directories:
D:\ACROBAT3\DISTILLR
D:\ACROBAT3\DISTILLR\XTRAS

These simple Include and Exclude Directories boxes allow many directories and subdirectories to be Catalog-ed into one searchable .PDX index.

The indexes created by Catalog offer a full set of features for searching the document collection. On the Web, most document collections are organized into somewhat hierarchical arrangements, and the publisher provides links among the documents for ease of surfing the information. For example, the Adobe Home Page at http://www.adobe.com provides a number of paths from the top page.


Options

Words to not include in index

an
at
in
on
the

Word: then

☐ Do not include numbers

Word Options

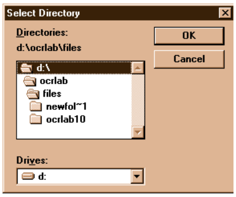
☑ Case Sensitive ☑ Sounds Like ☑ Word Stemming

☐ Optimize for CD-ROM

☐ Add IDs to Acrobat 1.0 PDF files

This very simple screen has powerful effects on future searching of this collection. For example, if none of the Word Options was checked, future users would not have access to these search enhancements. Another testament to total user control of the database index is the “Words to not include” menu, which allows the publisher to declare certain terms to be of no value for searching.

The home page of a Web site is the first page that is presented to a user. In most cases, this top page contains links to many other sub-pages, which comprise the bulk of the contents of the site.


Select Directory

Directories:
d:\ocrlab\files

d:\
ocrlab
files
newfol~1
ocrlab10

Drives:
d:

Acrobat Catalog adheres to common file structures.

The limitation to this approach is that there is only one way to find things, which is the fixed set of links designed by the Webmaster who manages the site. For this reason, most large commercial Web sites include a content search facility, often available through a button on the top page. The content search engines are free for the asking on the Web. Adobe started using the Excite! engine in early 1996. The limitation of such content-based search tools is that inexperienced users often have trouble writing a productive query that retrieves all relevant documents and only relevant documents.

By combining the limiting effect of Document Info and Date fields with full text retrieval, Search provides a simple interface for creating well-focused queries. Verity began shipping its free SearchPDF Engine in mid-1996, which is specifically designed to handle Catalog-indexed collections of PDF documents on the Web.

Document Management 101

Many people lose track of the files on their own PCs as time goes by, and this tendency increases by a couple of orders of magnitude when many users share the same files on a network. Every user has felt the frustration of looking back over old directories and mentally kicking himself for not remembering what the file was called or where it was saved. Most modern applications, from Microsoft Word to Adobe Acrobat, offer users the ability to add Meta-information to their files. Thus, we hope in the future that many paths of inquiry will lead to lost documents rather than confound the user with the silent opacity of forgotten or too-often-used file names.

The key to feasibility of this scheme is that someone makes sure to take advantage of the built-in structure by filling these value added fields with intelligently organized information.

The PDF structure offers many of the core elements of document management information. Documents are specifically identified through these multiple fields, and future users of these collections will be able to assemble very specific subsets of the collection for a unique purpose or audience.

T Elementary version control, one of the key functions of traditional document-management systems, is available in the Date Fields that show when the document was Created and Last Modified.

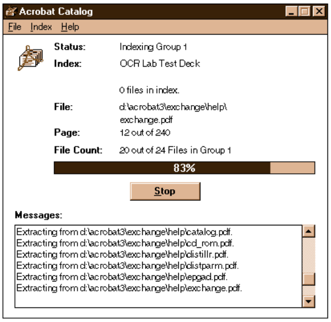
Combined with the Security feature, a basic document-management system is built into the Portable Document Format. Two primary functions of document-management systems are available: Access to the document is limited to password-group individuals, and version control is determined by the Date Fields.

blast from the past

Document-management software grew up around network word processing. Some of the heaviest word processing users are law firms, where one of the primary products consists of words on paper, in the form of legal documents. As most of us know all too well, legal documents tend to be long and complex and evolve through many versions during their lifetimes. In a law firm, the factory where these intricate devices are fabricated, many individuals are involved in the process, and many subcomponents are assembled in the final product.

When typing moved to word processing, the individual operators still had physical control of the documents and components, if only on floppy disk. When word processing moved to the file-sharing environment of a network, a critical requirement arose to control all of the versions of the documents and subcomponents. An entire niche industry grew up known as document management, and the early leaders included SoftSolutions, PC Docs and Saros. These document-management systems brought traditional controls to the potentially chaotic world of electronic files.

On a global contract, many offices will work with the same set of documents. It becomes vitally important to control each document in the physical sense, restricting access of certain users to reading, copying, editing and so on. More fundamentally, it is crucial to maintain the integrity of the final document and all of the sources and versions that contribute to the ultimate document.



Status: Indexing Group 1
Index: OCR Lab Test Deck

0 files in index.

File: d:\acrobat3\exchange\help\exchange.pdf
Page: 12 out of 240
File Count: 20 out of 24 Files in Group 1

83%

Messages:

```
Extracting from d:\acrobat3\exchange\help\catalog.pdf.
Extracting from d:\acrobat3\exchange\help\cd_rom.pdf.
Extracting from d:\acrobat3\exchange\help\distillr.pdf.
Extracting from d:\acrobat3\exchange\help\distparm.pdf.
Extracting from d:\acrobat3\exchange\help\epgad.pdf.
Extracting from d:\acrobat3\exchange\help\exchange.pdf.
```

Acrobat Catalog in action, charging through a collection of directories.

Full Text Retrieval For Total Access

Full Text Retrieval (FTR) technology corresponds to a sort of Popular Mechanics view of using computerized information. Once the stuff is in the computer, most users assume they should be able to just enter their question and get their answer. It's hard to argue with that premise because we all wish that everything would be as simple to use as Star Trek led us to believe. "Spock, give me all the information we have about life forms on this planet," Captain Kirk would command. In seconds, Spock would come back with all the relevant data.

High expectations, fueled by sci-fi, are almost realized through the FTR engines, especially on the Web. A seeker of truth can really learn a lot by pursuing certain ideas through several queries because the varying results of each query can be incorporated dynamically into a more precise query. However, while the results are spectacular, they are never achieved with the ease of Spock because most users are still limited by a lack of knowledge of all of the power of the available tools. But it won't be self-limited for long thanks to the user-friendliness that is increasingly engineered into the Web search engines, as well as the "intuitive" techniques woven into the desktop operating systems. Some day soon, a commonly accepted query language will be adopted as a standard, and information retrieval will be as simple as math and algebra.

tip

The harder you work, the luckier you get.

Index searching and FTR, the two polar extremities of document searching, are perfectly complementary. Full text searches tend to return too much content that is irrelevant, whereas index searches tend to miss relevant material because the user is not familiar with the index structure. Referring to Dr. Salton's criteria for measuring retrieval effectiveness, precision and recall, Acrobat Exchange offers the potential for great performance by giving users tools to directly adjust both parameters.

FTR = Full Text Retrieval, refers to a user's ability to search the contents of the documents rather than just the index.

Precision = Retrieve only those documents relevant to the query.

Recall = Retrieve all those documents relevant to query.

Acrobat Catalog creates an elegant, multifaceted document database that can be served up on the Web. With Verity's freely distributed SearchPDF server software, a breakthrough in rich, searchable collections is expanding quietly on the Web.

At a time when most Web sites offer either preconceived hyperlinks or some loose search capability, Web sites with large collections of PDF documents and SearchPDF offer a very organized, stable structure. Acrobat PDF and Verity combine to form a very strong alloy of technology.

Full Powered Boolean Search, Without the Hassle

Boolean search is popularly depicted as uncool, as a first-generation approach. New engines promise concept searching and natural language interfaces to very large collections of information. They encourage a “just type in whatever might relate to what you are looking for” approach.

Boolean Logic In Query Terms

Boolean logic is a model for information retrieval that combines query terms into complex relationships. The base Boolean operators are And, Or, Not. These operators allow the user to build multi-term queries with specific relationships between the terms.

For example:

Find “blackbird” And “stealth” — Finds spy planes

Find “blackbird” Not “thrush” — Disregards bird watching

The fact of the matter is that today’s concept engines do a great job at expanding the query, which means they tend to retrieve a ton of documents. This is often counterproductive because the user now must plow through tons of “hits” to find information of real value.

Anyone who has tried to use FTR search engines productively will admit Boolean queries make a lot of sense. The drawback, or sacrifice, is that the user must make the effort to learn a little simple arithmetic to begin to appreciate the full power.

Concept search engines are simply automating basic Boolean search operations. The concept search engines rewrite the user’s query by expanding the original terms. Obviously, semantic and statistical concept search theory represent some of the leading developments in this field, but most business applications will be very well served by the simple yet powerful Boolean search capability available in the Search feature.

	For Greater Precision		*For Greater Recall*	
	Refine The Search		**Expand The Search**	
1	**AND**	A AND B AND C	**OR**	A OR B OR C
2	**PHRASE**	”A B C“ (in that order)	**Wildcard**	
3	**NEAR**	A NEAR B NEAR C	**Word Stemming**	
4	**NOT**	A, B, C NOT D	**Thesaurus**	
5	**FIELDS**	(Title **Contains** Acrobat)		
6	**Combine Terms**	A AND (B OR C)		
7	**Review Results**	Refine Terms for New Search		

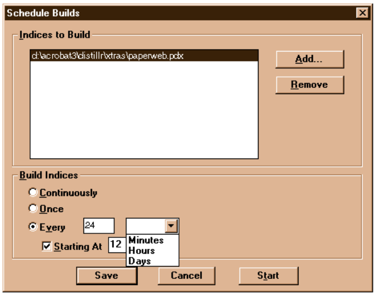
Field/Full Text Search Combined

The limit of success for searching a specific field index database is the user's knowledge of the field structure and contents. If the user does not know the specific appearance of the query target info, in terms of exact spelling and particular field location, it is impossible to reliably find information in a traditional field specific database.

At the same time, a full text retrieval search will often be far clumsier than an index search. It is difficult for the user to guess at the proper terms as they might appear in the text, and query expansion techniques often return large numbers of irrelevant documents. The time spent wading through large hit lists is counterproductive in the information-seeking process. The combination of using document information fields and text search in a query is the most productive way to use this knowledge-expansion technology.

tip

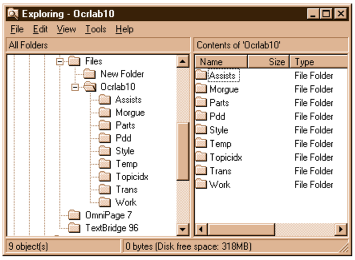
The answer to the question of how to index a document database is always: "As many ways as possible. I have no idea how future users might use this collection!" But still, this question must be asked: Do you want to sacrifice labor and overhead to add more accessibility to your collection?



Indices to Build: d:\acrobat3\distillr\xtras\paperweb.pdx

Build Indices:
○ Continuously
○ Once
● Every 24
☑ Starting At 12

"Builds" or new indexes can be created according to the users' needs. A Build re-indexes all of the PDF files in a targeted set of directories and can be scheduled to be performed at all intervals from just Once to Continuously, with Minute, Hour and Day frequency definitions available.



Contents of 'Ocrlab10'

Name	Size	Type
Assists		File Folder
Morgue		File Folder
Parts		File Folder
Pdd		File Folder
Style		File Folder
Temp		File Folder
Topicidx		File Folder
Trans		File Folder
Work		File Folder

9 object(s) | 0 bytes (Disk free space: 318MB)

For ease of operation, it is helpful to have all of the related files in one directory to get the most out of Catalog and Search.

Optimizing Build Options

Because index size directly affects performance speed, and because this issue is by far the most important issue for the end user, it is vital to understand the Build options. There are two options that remove words and numbers from the index. And there are three options that provide query expansion through wildcards, word stemming and thesaurus; and two that refine a query through case match and proximity.

All of the Build options affect index size and, therefore, performance speed, and the end user's needs should determine which options to employ. Features that are unlikely to be used, or unlikely to provide useful help, should be discarded.

Full-featured information-retrieval systems are resource-intensive, and adequate hardware and bandwidth are required for optimum functionality and user approval. This is not the type of application to be dropped willy-nilly on an already-busy network server.

Individual Builds, or index-creation processes, can be designed to include and exclude an array of directories and network drives. Builds can be scheduled to occur Once, Continuously (careful!), or on a time schedule to accommodate resource availability.

To assure best performance, a Purge should be performed on an index before each new Build. This can be accomplished by simply deleting the nine subdirectories, or subfolders, under the index directory or folder.

For ease of purging, and also for moving and backing up to other media, it's best to keep the index and nine subs together: assists, morgue, parts, pdd, style, temp, topicidx, trans and work.

The Process

The process for creating and maintaining the Catalog will vary depending on the scope of your digital document collection. Again, preparation is critical for creating an ongoing dynamic collection that makes information readily available.

tip

Encourage the use of the Author, Title, Subject, Keyword and Date Fields because they make the contents of big collections much more accessible. If you're searching a set of articles or memos, you want to be able to narrow the search down to the smallest subset you can define. If, for example, you want all the memos authored by Bucky Fuller that concern the keywords tensegrity principles and spaceframes, written between 5/47 and 5/74, this type of refined search will be much more rewarding and will illustrate the real speed and power of instant access to information.

1 Determine user requirements and areas of improved access to information:

Publish an author’s guide to proper and accepted use of Doc Info Fields;

Options that affect Index Size, Speed and Search Features:

List of Excluded Terms

Include or Exclude Numbers

Case Sensitive

Sounds Like

Word Stemming

2 Carefully estimate required resources for Catalog and Search:

(Do Not Scrimp HERE!)

Processor resources should be generously allotted to this intensive task;

Disk space for documents PLUS 50-80% overhead for index.

3 Create a new index in Catalog:

Create Index Title and Description;

Choose Directories, or Map Network Drives,

where PDF collections are stored;

Schedule the Build (Index Creation Process) for Once, Continuously

or Timed;

Choose Options based on User Requirements for Exclude Terms and

Numbers, Case Sensitive, Sounds Like, Word Stemming.

4 Document Index features and options in effect to help future users be most efficient.

Summary

The two prime approaches to information retrieval, which are fielded index and full text searching, are available with Acrobat Catalog.

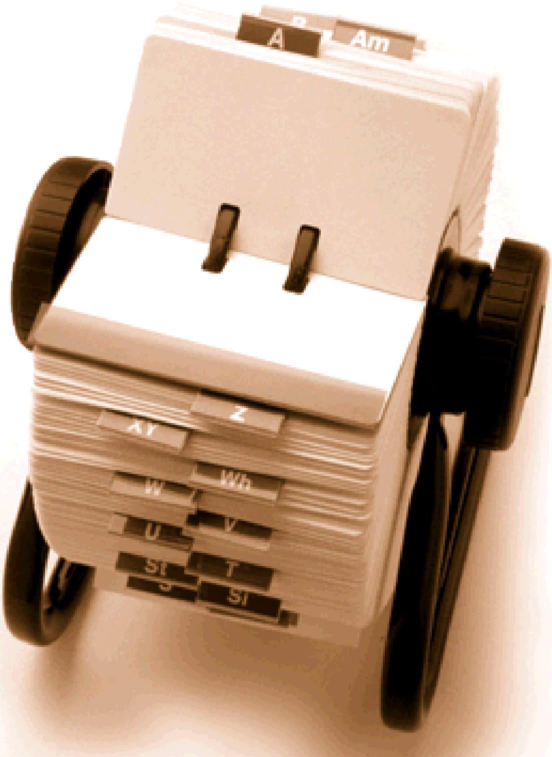
Index retrieval through the Document Info fields provides specific access to individual files through traditional document-management methods.

Text Search offers the ability to query the entire contents of the files, in addition to the Document Info fields.

Combined index and text queries offer the synergy of both techniques, as long as the end users are familiar with the data likely to be found in the Document Info fields.

To take fullest advantages of these capabilities, all authors should add Document Info to offer another means of navigation, potentially as helpful as Bookmarks and Thumbnails.

Document database architects and publishers (such as yourself) should always provide simple guides suitable for first-time users that tell them what they can expect to find in the Document Info fields and text of the collection. A few simple hints can point the best way to success in each particular index and text-retrieval collection.


A
Am
Z
W
Wh
U
V
St
T
Si