decisions in

indexing

chapter six

Having a pile of digital documents isn't any better than having paper if you can't instantly retrieve the information you want. To achieve instantly accessible information, you'll need to carefully consider your indexing plan to accommodate the needs of both current and future users. With tools like Acrobat Catalog, you can establish consistent and effective indexing criteria.

How Will Future Users Access the Collection?

Future users will experience your document collection in many ways, and it is important to consider all of their goals when creating indexes with Acrobat Catalog. Though it is tempting to check off every option and take advantage of every possible feature, overall user satisfaction has to be considered. Just because your car can go 120 miles per hour doesn't make it a good idea to take advantage of that capability in every situation. By the same token, not every capability of Acrobat Catalog should be used "pedal to the metal."

tip

While UNIX, Mac and Windows 95 all support "long file names," the large number of users and computers out there still make this old convention a serious consideration.

Remember, to handle any files with this currently dominant convention, all long file names are truncated. When "My Documents" is reduced to "mydocu~1," information retrieval capabilities may be severely degraded. Therefore, it is still a good idea to use conventions like FILENAME.EXT file names.

All of the word search options come at the price of overhead in the index, which affects everything from overall index and collection size to the response speed of the search engine. Although users may not care about the former, the latter is a paramount concern.

If future users will rarely or never use certain options, you should not burden your collection with them. If the Case Sensitive, Stemming or Sounds Like options aren't worth their weight, they should be ignored.

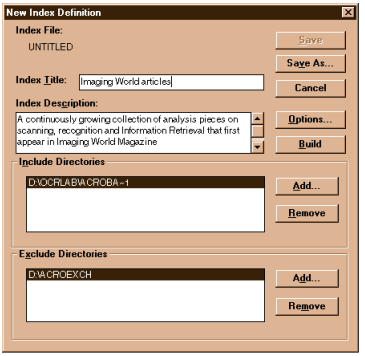
Effects Of Options On Index Size

Index Size compared to Text Collection	*10-30%*
Remove up to 500 Stopwords	*10-15% Reduction in Index Size*
Remove Numbers	*10-20% Reduction in Index Size*
Remove Word Stemming	*10-20% Reduction in Index Size*
Remove Case Sensitive	*05-10% Reduction in Index Size*
Remove Sounds Like	*05-10% Reduction in Index Size*

Stopwords

Stopwords are words that are not indexed, and therefore not searchable, in a text database. The reasoning for stopwords is that they convey little value for the burden they place upon the database. Typical stopwords include articles like "the," "a," "an," as well as prepositions such as "to," "in," "from" and other common words. By removing them, not only is the text index smaller, but the number of irrelevant retrievals is often reduced.

However, if your users will be searching for "From Here To Eternity," they would get hits only on "Eternity." Or if they were searching for "On The Road," they would find every document with the word "Road."



On the other hand, if your database includes terms that are simply too common to provide any search value, the Stopwords function allows database customization.

Meta-information such as Index Title and Description is entered here, and all of the Index Options are available as pulldown menus. The Build button creates the Index.

Numbers

In the modern world of brand names and revision levels, it might be difficult to choose not to index numerics. For example, it wouldn't be possible to search for "Windows 95" or even "July 4."

There may be certain disciplines where numbers will never be used in queries. For example, searching for taxonomic classifications may never require numbers. And if there is an alpha-only primary field or key in this index, searching numbers may be superfluous.

Options

Case Sensitive, Stemming and Sounds Like Search options require a certain amount of overhead in the index. This overhead affects both the speed of searching and the size of the overall indexed collections. If future users are not going to enjoy significant, outstanding benefits from these enhancements, they should not be larded on to the system. For details on these search features, see Chapter 13.

Dynamic Or Static Collection - How to Manage Updates

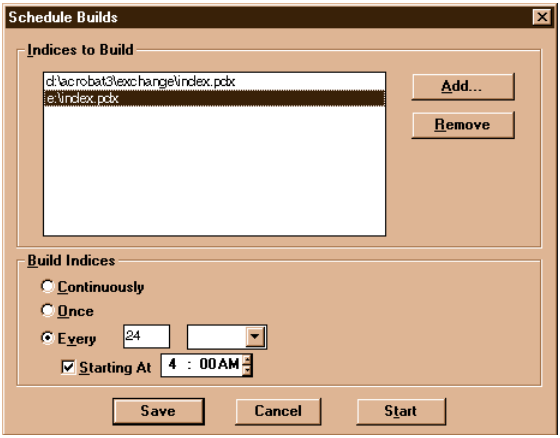
In the ideal system, every new document is instantly incorporated into the index. The Continuously option provides for this luxury. A folder can be constantly watched or be in perpetual processing, with Catalog grabbing every PDF file that shows up and adding it to the Searchable Index.

It is yet another one of those "common sense" feelings about computers: They should naturally have this ability. At this point, the software can do it, and the hardware is rising to the challenge, doubling in speed every 18 months while staying at the same price.

However, as collections grow in size, even the fastest computers run a little slower. It will always be a tradeoff between user response and timeliness of database updates. In the fields of finance and war, timeliness is everything. In more conventional fields, other measures may apply.

Updating Once would be used by publishers who wish to distribute pre-organized, enhanced collections.

Using the Every menu option is the conventional way of doing database updates, popular on every platform from the earliest mainframes to every type of online host platform. Usually, these scheduled builds interfere the least with users because they are scheduled for off hours.



For these choices, consider the effects on server loading compared to the value of updates to the users. Information currency vs. access speed is a question that arises at the busiest sites.

Naming Conventions

Standards are a godsend, allowing developers to concentrate on meaningful issues rather than try to invent new publishing media. Now that virtually every PC ships with a high-speed CD-ROM drive, this media has achieved universal acceptance. It's cheap, it's easy, and it works on every platform.

ISO 9660 is in large part responsible for the fabulous blossoming of this new media. By providing one stable standard format for all CD-ROM, the cheap real estate of that 650 MB universal media on a sturdy plastic disk became the de facto standard of physical distribution.

Scan 70,000 Pages

Per Week

At 99.985+ Accuracy

case study

An Interview With Dave Abbott Of Reed Technology Information Services

The Government Services RTIS has held the data-entry contract for the Government Patent Office since 1969. This demanding contract specifies extremely high accuracy and a rigid production schedule. ICC has consistently exceeded the specifications by always seeking out and using the most effective techniques and technologies.

Since joining International Computaprint Corporation (now RTIS) 15 years ago, Dave Abbott has been a driving force in developing world-class data-capture systems. “We are always aiming to increase productivity so we can increase the workload,” he declares. By employing the most effective systems, “Resources can be allocated to intelligence and information development.”

A review of Dave’s production methods over the last 15 years provides a unique history of the scanning and OCR industry. The contract involves data entry of legal, approved patent applications into the GPO’s typesetting system. “We still use a term called ’brown bag patents‘ to refer to very long patents,” Dave explains, “going back to the days when all the patents were keypunched onto paper tapes. The tapes for a 1,000-page patent would have been moved around in a shopping bag.”

From those early days of paper tape, ICC moved on to key-to-tape and key-to-disk operations in the early ‘70s. However, the first breakthrough in data-entry productivity came in 1983 with the adoption of Dest OCR scanners.

Dave smiles with satisfaction at a technology investment that paid for itself 100 times over. The Dest 246, perhaps the most successful OCR scanner of all time, accurately read a limited number of popular typestyles at a rate of 15 seconds per page. Having recently retired the Dests due to lack of maintainability, Dave estimates, “Our cost on those scanners was down to the hundredths of a cent per page.”

Though the Dest was limited to about a dozen typestyles, 75 percent of patent applications were typed in those styles. As Project Director for the patents contract, Dave developed a custom interface to the error-correction and composition systems. “We expect our typists to produce eight pages, or about 10,000 keystrokes, per hour,” he says.

For scannable patents, the Dest 246 could potentially outperform a data-entry operator by 30 to one. And the scanner’s accuracy was equal, often superior, to data entry. That bench of Dest 246’s ranks as one of OCR’s greatest implementations ever.

In the last year, Dave has raised the standard again. “With the new system,” he says, ”our scannable documents have increased from 75 percent to 95 percent of our work, up to 70,000 pages per week.” In effect, ICC has increased the number of documents processed by more than 20 percent while decreasing the time to scan them by a factor of three.

Dave scans the documents at 400 dots per inch rather than the more common standard of 300. “It only adds 20-25 percent to the file size,“. he tells us. ”And I have never heard of accuracy *decreasing* by going from 300 to 400 dpi.”

“OCR accuracy is around 98 percent,” states Dave, “or 2,000 errors per 100,000 characters. Our contract calls for 15 errors per 100K, and we do better than that.” Though spell checkers are a common means of document correction, Dave is “cautious about using dictionaries. Humans have a better overall frame of reference, so we only use dictionaries with human review.”

“We have found self-directed work teams to be a tremendous source of innovation,” Dave explains. “We have scheduled training courses in Object-Oriented training for both our operations and programming staffs because learning is the foundation for innovation. We want to provide our people with intellectual tools as well as hardware and software tools.” Dave’s devotion to understanding and using technology is obvious as he says, “We are constantly streamlining our processes through these tools. They are all enablers.”

David K. Abbott is Vice President for Reed Technology and Information Systems in Horsham, Pa.

ISO 9660: The worldwide physical standard for CD-ROM: how many tracks, how many sectors, what size sectors, how much error correction, how much 'lead-in,' how much table of contents ... all very stringently controlled by International Standards Organization ISO 9660.

Andy Moore, Moore's Imaging Dictionary, 2nd Ed., August, 1995, Flatiron Publishing.

The Process

The actual process of indexing your collection is straightforward if you've already assessed your user needs.

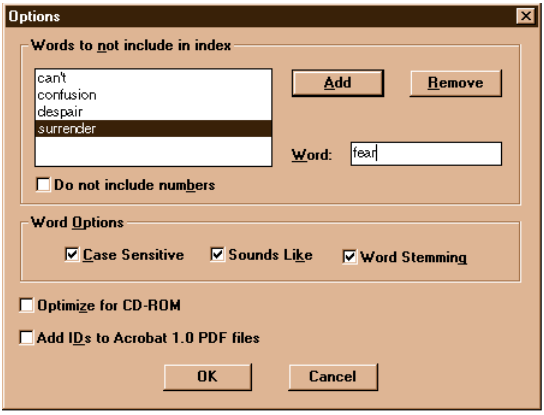
steps

1. **Gather PDF Files For Your Collection** → *Keep Relative Paths*
2. **Finalize Document Information And Navigation Requirements** → *Bookmarks, Hyperlinks, Thumbnail Views*
3. **Help Users To Grasp The Collection** → *Document Site With Maps, Search Options, Descriptions*
4. **Use Catalog To Build The Index** → *Use Fast Machines For Servers*
   Set Preferences For Peak Performance
   Optimize For CD-ROM
   Purge For Efficiency
5. **Serve Your Index On A LAN, Web Server Or CD**

Gather PDF Files For Your Collection

The best way to build an Acrobat Catalog collection is to put all of the files into one folder, with all related files in subfolders. This conventional directory tree structure provides both efficient file access performance and convenient portability.

Acrobat Catalog directs its indexing function at pre-defined file structures that are specified in the index definition process. The subsequent index and the nine support directories should ideally be stored in the same folder with all related files in the collection.



Sphinx-like simplicity of interface offers extensive database tailoring choices.

The beauty of keeping all of the documents and organization files in one folder structure is that the entire collection is easily portable. This means that the hyperlinks within a Catalog-indexed collection will maintain their full functionality when they are moved to new media, whether that be a CD, a new disk array or a Web server.

Finalize Doc Info And Navigation Enhancements

Like everything else, the work you do is equal to the value you create when it comes to building a Catalog Index. To be truly better than paper, our new documents must take advantage of all of the built-in potential for advanced functionality.

At a bare minimum, the Title, Author, Subject and Keyword fields should be filled in to provide future utility to any and all potential users. The System fields of Date Created and Modified and so on will be automatically indexed.

All bookmarks and hyperlinks should be finalized before the Catalog process, although they can be added later. The value of one agreed-upon version of a document collection can not be overstated, and multiple versions of an index inevitably lead to confusion.

Help Users Grasp The Collection

In addition to employing effective design rules in building your collection, you can help the user by explaining the rules you have followed. These rules include the type of information in each of the document info fields. Specific usage should be described, and any custom fields must be explained.

For example, the doc info fields could be used for legal documents

Author	*Lead Attorney*
Subject	*Client Matter #*
Title	*Client Name*
Keywords	*Related Parties*

Any use of Catalog options should be explained, to help users understand the search techniques that will work in this collection. If stopwords are used, it would be helpful to include a list to avoid wasted queries. Certainly, if numbers are not included in the index, the users should be warned.

Following the example of many advanced Web sites, the folder structure might add to user convenience with a description of the contents available for top-level browsing. This approach gives the user a bird's-eye view of the entire collection before he begins to search and retrieve.

tip

Always design collections to emphasize user convenience. All the information in the world is useless if no one gets to it.

Building the Index

Never make the mistake of choosing your information retrieval server as the place to save money just because it is actually just a big computer sitting in a room gathering dust and generating heat. Large text database collections should be hosted in environments with lots of free random access memory (RAM) and fastest possible bus connections to large-capacity storage.

The Acrobat Catalog User Guide suggests having at least 10 times the amount of RAM as the file size of the largest document that will be indexed. For example, 24 MB RAM is recommended to index a document of 2.4 million words.

Compared to the earlier examples in this book of 333 words per single-spaced page, 2.4 million words could be estimated at a nominal 7,200 pages. Since most of the new PCs come with 32 MB of RAM, a 9,600-page document should fit into memory comfortably on a midrange PC for speedy Catalog indexing.

tip

Nominal pages as defined in this book are 2,000 characters per page, or 333 words per page, equivalent to solid, single-spaced typing. This is a relatively dense document format, somewhat less dense than books but much more dense than most common business and legal documents.

The moral of this story is that the capacities are at once incredible and still realistically limited. The Intel P7, the PowerPC and the Ultra Sparc, fast memory buses and cheap RAM may minimize all of these concerns in the immediate future. Performance is, of course, enhanced in high-powered operating system environments. Still, economic design choices never change. If no one will ever use the frills, don't waste the space.

Set Preferences For Peak Performance

Even for those of you who never touch .ini files, let alone consider editing them, please do consider it just this once. The Acrobat.ini file contains the parameters that precisely control the Catalog indexing functions, and they can be tweaked with just a little effort. The .ini file is ASCII, so it can be easily edited in Notepad or DOS Edit.

These simple adjustments are explained for Windows and Mac users in the Catalog 3.0 Online Guide. They allow the publisher to best serve the users, whether it be a dynamically updated database or a tightly streamlined index for CD-ROM distribution. It's worth learning because it serves the primary goal: making the information easily accessible to the user.

Optimize For CD-ROM

This option organizes the files so that the index information is optimally accessible, providing for the quickest possible searches. This option is only one step in creating a collection that is optimized for CD-ROM.

The 650 MB capacity of a blank CD seems gigantic, but multimedia such as sound, and especially video, will consume large chunks of storage. The rich graphical content of new documents also creates big files, and the search and navigational capabilities add their required space. Plan ahead!

According to the Catalog Online Guide, the "GroupSizeForCDROM = 4000." In English, this suggests that 4,000 documents is the maximum number that will be reliably indexed under the Optimize for CD option.

If you are writing large numbers of short documents, you should consider using automated bookmarks and links to provide alternative navigation and retrieval methods.

Purge For Efficiency

The Acrobat Catalog indexing process is incremental, so if a collection is being re-indexed as new documents are added, the file space consumed by active and inactive indexes grows continuously. To improve response time and to provide additional disk space for new information, it is important to Purge the indexes as often as necessary to always arrive at the most efficient system.

tip

"A faster way to purge an index is simply to delete the nine subfolders of the index folder: assists, morgue, parts, pdd, style, temp, topicidx, trans and work."

Such a "purge" could be accomplished with the DOS "DELTREE" command from the Index Directory on down.

Courtesy of the Acrobat Catalog Online Guide

Serve Your Index

The Catalog-indexed collections can be served on most media. The mix of Acrobat products available includes

LAN *Multiple User licenses of Acrobat 3*

Web *SearchPDF on Server, Acrobat Plug-in on Browser*

CD *SearchCD and all Acrobat Readers on disk*

tip

Managing Folders And Drives

Keep your file hierarchy consistent when using indexes, otherwise your system won't be able to find the original files to retrieve. The whole idea is that you can just pick up the top folder and move the whole collection elsewhere (even to other drives), keeping the relationship of the contents intact. The idea of subdivided but cohesive files is actually just like the hanging Pendaflex folders full of manila file folders full of paper-clipped documents that we all know and use.

Summary

The key to accommodating the needs of our future users is providing a speedy and productive means of retrieval. The nature of this information should determine the way it is indexed.

- Is pure speed of retrieval likely to be more important, or should utmost flexibility be built in for future researchers?
- Is this collection designed for ongoing, dedicated users who will learn all of the available functions, or is this information meant to be available on a hit-or-miss basis to casual users?
- Is file space or bandwidth a consideration, such as on a network or a Web site? How many CPU clicks can be devoted to option-enhanced text searches?

The best answer is generally not the "kid in the candy shop" response. The publisher must consider the way the information is currently accessed, and how that access can be improved in terms of response time and information retrieval options.

part 3

searching digital content