

# From Paper to Web

- Ch. 1: The Past - A Paper Legacy
- Ch. 2: The Present: Making Existing Digital Documents Accessible
- Ch. 3: The Future: Digital Documents
- Ch. 4: Acrobat Exchange: An Architecture for Instant Access
- Ch. 5: Acrobat Catalog: Creating the Keys to Instant Access
- Ch. 6: Decisions in Indexing
- Ch. 7: Acrobat Search
- Ch. 8: Enhanced PDF Collections on the Web
- Ch. 9: Advanced Navigation for Superior Information Access
- Ch. 10: Organizing Digital Documents
- Ch. 11: HTML Documents: Creation, Editing, Management
- Ch. 12: Advanced Searching Techniques
- Ch. 13: Tapping the Web: A Fountain of Information
- Ch. 14: Document Management: Information as a Corporate Asset
- Ch. 15: Publishing on CD

## How to Make Information Instantly Accessible

**Tony McKinley**

---



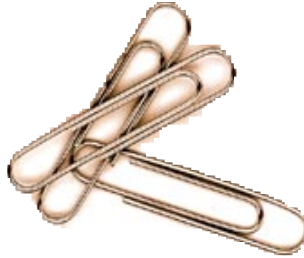
ADOBE  
PRESS

Adobe Press books are published by Macmillan Computer Publishing USA. From Paper to Web, the paper edition, is available at bookstores, both physical and virtual. ISBN: 1-56830-345-9. \$45 US.

the past :

a paper  
legacy

chapter one



## The Users And Tools Have Gone Digital, But The Information Is Stranded On Paper

"By 2004, the pile of information on your desk will be 30 percent paper and 70 percent electronic, compared to 90 percent paper today," states Dr. Keith T. Davidson, Executive Director of Xplor International.<sup>1</sup> Dr. Davidson illustrates our business dependency on paper by citing statistics such as: "Executives spend as much as three hours every week looking for missing information, the average document is copied 19 times, 200 million pieces of paper are filed away each day."<sup>2</sup>

For more  
information,  
visit the  
Xplor web site  
at

<http://www.xplor.org>

This expensive and time-consuming medium of paper is growing: more paper documents are produced now than ever before. This isn't just a problem for legacy information, but an ongoing requirement to handle increasing volumes of paper in today's information-dependent business environment. The Gartner Group estimates that 5.46 billion office documents are produced each year, 59 percent of which are accessed and retrieved manually.<sup>3</sup> Put another way, the Gartner Group estimates that 170 miles of new files are generated every day.

# Paper And Microfilm: Imperfect Mediums

Compared to electronic alternatives, paper information is much more expensive to create, reproduce and store. In a study conducted by KPMG/Peat Marwick that analyzed conversion from paper to electronic documents, it was conservatively estimated that Adobe Systems would save \$950,000 per year<sup>4</sup> with a full-scale change to the new paradigm.

Due to its physical nature, paper is limited in distribution by the number of copies generated. The fact that virtually every office and even every department has a copier is a testament to this limitation. Add to that manual updates to previously distributed documents, and you multiply the single task of document management by the number of users.

Conventional filing systems provide only one index field for the file, which is the paper file folder tab. To efficiently store and retrieve files in a file cabinet, all users must understand the indexing scheme. More important, files must be returned to the proper position in the cabinet because it may take hours of manual searching to find a misfiled document. It has been estimated that 7.5 percent of paper documents get lost completely.<sup>5</sup> This risk of misfiled and unreturned documents is virtually eliminated in an electronic file system because the documents themselves are never moved.

Paper isn't particularly useful as a groupware tool, either; changes to paper documents need to be republished on paper to spread the news. An electronic document can be annotated by many users, and the latest updated version of the document is always available to all users on the network.

Microfilm was the earliest means of preserving documents mechanically through photography, one of the ancient wonders of the Industrial Revolution. Through the use of high-power optics, entire books could be faithfully captured on a small amount of film. The negative of the film could be easily reproduced, providing a new distribution medium. Microscope-like readers displayed the pages and allowed users to peruse these miniature documents.

This new microfilm system had undeniable benefits and annoying drawbacks. On one hand, it was cheaper and easier to ship a few ounces of microfilm to distant offices than it was to ship a few hundred pounds of paper. At the remote office, a single file drawer could match the storage capacity of an entire file room, and a single administrator could manage an entire library of documents.

Users found microfilm to be fairly clunky. If you work for the FBI, the IRS or any other large user of critical source documents, using microfilm is infinitely better than wading through the comparable wilderness of paper. And microfilm works fine if you always search by a strictly defined field like Purchase Order or Customer Number.

The primary advantage is the simple physical compression of large collections of files. Unfortunately, access to information on microfilm is much more opaque than access to information in books on shelves or cabinets full of files. Access to pages on microfilm is usually strictly serial, and it is best to know exactly which page you are seeking. Without a dedicated computer-assisted retrieval system, there is no way to query or search the content of microfilm. Since the images must be displayed one-by-one on a viewer, the only practical means of access is the index.

## Digital vs. Paper Documents

We refer to books, file cabinets and libraries as comparisons to digital information. Specifically, we compare digital documents to conventional books. Books are highly evolved information transmission vehicles that illustrate the benefits of 500 years of human ingenuity and development. Books contain illustrative contents, from the finest typography to the most glorious graphics and photography. Books also contain highly advanced navigation and finding aids, such as table of contents, index, glossary and footnotes.

In their physical form, books are durable and compact vessels of information. We use books as a basis for comparison because simpler documents do not possess the features to compare to the richness of a digital document.

### Nominal Definitions Of A Page

---

A text page = 2,000 characters

An image page = 50,000 characters

These sizes are widely accepted as the rule of thumb for page calculations.

These figures are for discussion purposes only. The actual page sizes should always be measured and used in planning any actual implementation.

# Tools For Global Distribution

The earliest online databases were pure text databases, with no graphical content. The early LEXIS & NEXIS databases, provided by Mead Data Central, offered information in the fields of law and news, respectively. LEXIS guaranteed to have Supreme Court findings online within 48 hours of publication in Washington, D.C. In the beginning, all documents were manually rekeyed into the system, and eventually OCR was used to speed up the process. These methods gave Mead Data users rapid access to important information as soon as possible.

## Weightless Shipping: Comparing Paper To Digital Documents

### Pages On Paper Compared To Same Number Of Pages On CD-ROM

A typical CD stores about 650 MB (1 MB= 1024 x 1024 bytes= 1,048,576 bytes) of information, which equates to 340,000 pages at 2,000 characters per page, or about 170,000 single-spaced, double-sided pages.

If these duplex pages are printed on 20-pound bond paper, the stack of paper required to match the information capacity of a CD would weigh 1,700 pounds.

For reference, 170,000 sheets of 8.5 inch x11 inch 20-pound bond paper weighs 1,700 pounds.

Through the world-spanning communication facility of the Internet, digital documents can be made available for global instant access. For example, Adobe Systems distributes software on the Internet, and thousands of programs are downloaded every day.

An evolution in paper documents occurred when laser printers replaced impact printers, and another leap occurred for electronic documents with the development of graphical user interfaces. Just as laser printers provided more flexible fonts and page layouts, GUIs gave developers a much richer palette with which to paint.



# New Languages Are Born For The Web

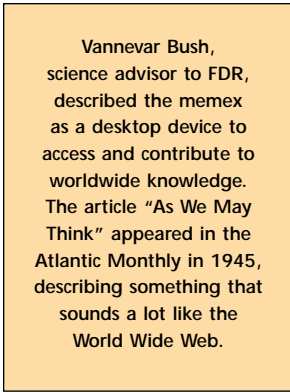
Another evolution came with presentation languages, and the birth of the World Wide Web is attributed to one such language. HyperText Transport Protocol (HTTP) introduced the fantastic ability of "Click to Go." Remember, the Internet existed long before the birth of the Web; it was just difficult to get around without a solid UNIX background.

Suddenly, with the development of HTTP, hypertext links were embedded within the documents. A user reading a remote document can simply click on a highlighted word or phrase and be instantly connected to another computer. Thus the documents themselves are linked, and the global wiring that accomplishes this miracle becomes invisible to the user.

The document language of HTTP is HyperText Markup Language (HTML). HTML can be considered the original word processor of the World Wide Web. The concept of hypertext itself is relatively ancient, going back to the dawn of computer time, as described by Vannevar Bush 50 years ago. But in addition to the crucial functionality of the links in hypertext, HTML was developed as a page composition language, which included a wide range of text attributes and graphical capabilities. In the long run, HTML added much more in terms of document connectivity, rather than in terms of the presentation of Web documents.

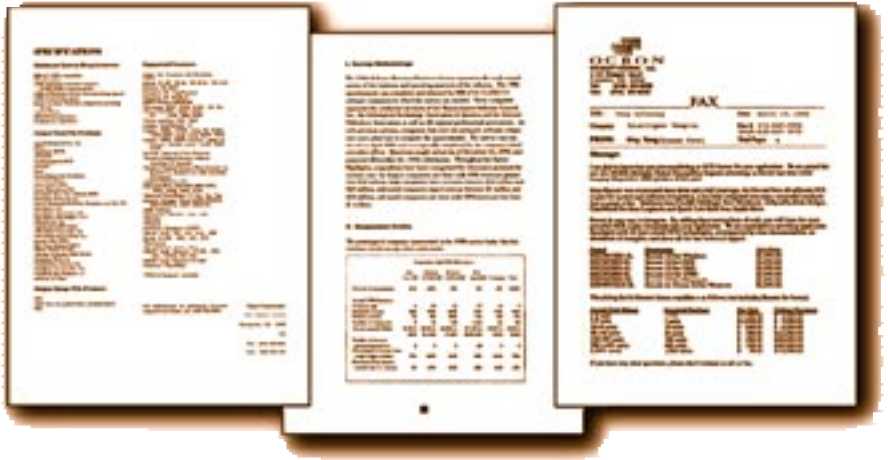
Even the most simple paper documents contain elements that have always been troublesome to represent on a computer monitor, and they have been even more troublesome to capture through a scanner. A signature on a letter is a very basic example of this "richness" that has separated paper from electronic documents. In the same way that a signature validates the letter, the lack of the signature on a reproduction may invalidate it.

When we take one small step further in the "richness" scale of common paper documents, we consider the broad universe of page composition and typesetting. Starting with basic reports that include charts and graphs, and moving into the complexity of books, newspapers and magazines, we encounter an entire "language" for the presentation of information. Using this "language," a tremendous amount of information is represented in a very dense, but instantly understood, form.



**Vannevar Bush, science advisor to FDR, described the memex as a desktop device to access and contribute to worldwide knowledge. The article "As We May Think" appeared in the Atlantic Monthly in 1945, describing something that sounds a lot like the World Wide Web.**

Adobe Systems introduced Acrobat and the PDF format to bring this richness to electronic documents. The Portable Document Format makes electronic documents much more familiar to users who grew up in a world of paper. PDF files are designed to be the analog to PostScript files in the sense that they can be used on virtually any output device. Whereas the output devices for PostScript are laser printers, the output devices for PDF are graphical user interfaces on virtually any hardware and software platform. PDF also retains all of the PostScript abilities to re-create rich hard copies through printers and faxes.



These samples  
are also available  
at

Columns, tables, letterhead: All are preserved with PDF format.

<http://imagebiz.com/PaperWeb>

The goal of PDF is to recapture the rich layout and presentation of information in a form that is equal to the presentation already out there in the vast realm of information residing in paper documents. To keep this evolution in perspective, think of the languages as generations that inherit all the capabilities of the former. ASCII was the first, HTML was the second, and PDF is a third generation language for electronic documents.

**Portable Document Format is designed to bring rich composition to electronic documents, and Acrobat Capture offers a direct path from paper to PDF.**

**Paper is a physical universal format;**

**PDF is an electronic universal format.**





To achieve global reach, a document will be duplicated and copied to multiple sites on the Web.

---

**Mirroring:** The same collection of documents resides on many servers on the global Internet. For example, Adobe, Microsoft and many others mirror their software releases on many sites on the Web to accommodate the greatest possible number of users with the best possible performance.

If you would like to get a copy of the Acrobat 3 Reader for Windows 95, simply point your browser to the following URL (Universal Resource Locator):

<http://www.adobe.com/acrobat/>

You will begin the download process. When the file `READER.EXE` is completely received on your computer, use the `RUN` command to choose this self-installing application. By following the attached instructions, it only takes a few minutes to configure Acrobat 3 Reader with your Web browser, usable as both a stand-alone program and as an integrated online viewer on the Web. When PDF files are downloaded, the Acrobat 3 Reader is immediately invoked to display the files.

For access to all of the software offered by Adobe, browse its Technical Support Library

<http://www.adobe.com/supportservice/custsupport/tsfilelib.html>

# Benefits Of Electronic Documents

Searching makes electronic documents superior to paper documents. A full text search capability enables the user to search the entire document for words of interest, and can fill in the gap left by indexes. Even the most extensive indexing systems are limited by the goals of the original index scheme.

- **An Index is (a few descriptive words) about the document;**
- **Full Text is (each word contained within) the document.**

**Comparing  
Telecommunications  
Methods In Pages**

To move the same  
**340,000 pages** discussed  
in the previous section,  
the time requirements  
over common  
telecommunications  
methods

---

**28,800 modem**  
@ **1.44 pgs/min**  
*63 hours*

**ISDN modem**  
@ **5.6 pgs/min**  
*15.75 hours*

**T1 Line**  
@ **77.2 pgs/min**  
*70 minutes*

The primary attraction of a text database is the ability to search for information by performing a simple word search. At first glance, this appears to be the ideal way to retrieve documents from a database. However, a poorly designed search may retrieve either no documents or far too many documents.

Even the earliest text-searching engines included tools to refine the search and increase the user's productivity. Advanced Text Search techniques are covered in Chapter 12.

The limited life span and gradual decay of paper can be overcome by transforming books into digital form. Some optical media claim 100-year durability, as magnetic tape has claimed for many years. Lacking time machines to verify the claims, archives can still be confidently built based on the digital format in which the data is encoded, whatever the physical media. Once digital, that electronic item can be converted to new media as it develops in the future.

The most dramatic advantage of digital over paper documents is the ever-increasing liveliness of the online medium. Mouse movements trigger sounds, clicked icons initiate videos, and interactive programs allow the user to move in a 3-D virtual

reality. Beyond gimmicks, interactivity offers a tremendous breakthrough in technical and educational documents. While books are limited to a few illustrations and step-by-step instructions, a digital document can provide instant access to complete visual, audio and 3-D demonstrations.

# Problems On The Digital Road

The road to instantly accessible information has some potential delays and complications.

All of the technology is in place: powerful CPU chips, wireless communication links to the Internet, more efficient batteries, better screens, voice input. The portable Web TV is selling well. But the price is high for the latest equipment, most of which will be considered outdated within six months of purchase.

It's important to carefully assess the costs and benefits of any large scale digitization project before buying equipment on site. In addition, consider the costs of staffing and training when evaluating a solution. Look for a system that's both flexible and compatible on multiple platforms, today and in future years. The long-term accessibility of your information depends on the planning you do now.

## Complicated Scanning Names

Regular desktop scanners run at truly prodigious speeds, sucking in images of paper pages and crystallizing them as digitized files. One step up, like the Mustang GT upgrade in a Ford, and the scanners have tons of extra image-gobbling horsepower.

It is simply good sense to figure out what you hope to do with that thing before you pump the pedal and stomp on it.

Very basic scanner controller software will follow even the least-informed user's bidding and run that machine at maximum controllable speed. Low-level software will spit out a stream of images, and this pile of images will be perfectly usable by any C Programmer. Such snowblower-like output of naked digital files is not very handy for a normal person to deal with.

If the images are going to be captured through a very low-level interface, the user must be careful and consistent with file names. These file names are often very constrained in length because low-level controllers insist upon room within the file name itself for incremental numbering.

### The Office Of Technology Assessment: A Virtual Agency Gone But Not Forgotten

The OTA was established by an Act of Congress to provide an unbiased analysis of technology across all industries and disciplines. About 775 reports were generated in the 23 years that OTA was in operation, and these were all distributed in printed form, courtesy of the Government Printing Office.

To call the products of OTA Research Projects "reports" may be misleading to readers. The published documents often looked more like text books. The average document was 80 to 100 pages and in-

cluded photos, illustrations, charts and free-form art.

The Washington Post announced the closing of this taxpayer-owned think tank in the Sept 28, 1995 edition: "The Office of Technology Assessment closes Friday, the first government agency eliminated under the new Republican revolution." Richard Nicholson, Executive Director of the American Association for the Advancement of Science, was quoted as saying, "There used to be a time that knowledge was power. Now it seems like Congress has decided it's a nuisance."

When the closing was announced, the agency was inundated with requests for complete collections of the entire library of OTA Reports. In the end, 85 sets

#### tip

### Incrementally Numbered Page Image File Names

Given Name: FILENAME.ext

File Names: FILE0001.ext ... to FILE9999.ext

Document-scanning software often dedicates the last four characters of file names to a four-character numeric field to track page numbers. This allows one document to be up to 9,999 pages long. While this may seem like a good idea to a programmer, a scanner operator is best advised to keep scanned batches to a limit of 50 to several hundred pages.

The reason for processing small batches of images are manifold:

- Simpler recovery from any failure on one batch
- Smaller files to store and read from disk
- Smaller documents to handle in workstation memory
- Smaller files to move over any network

One 500-page book may be scanned in 10 batches of 50 pages each. Actually, there are only 25 paper pages in each batch, but because they are printed on both sides, they add up to 50 page images.

of books were scrounged together and sent out to repositories around the world.

This would have been the sad end of a noble endeavor if not for the lifesaving capabilities of a new technology. The OTA will live on as a "virtual agency" on CD-ROM and the World Wide Web. Acrobat Capture technology provided a means to transform paper documents into rich electronic documents at an affordable cost.

"Our major objective is to preserve a legacy of OTA," declares Peter Blair, Assistant Director for the Industry, Commerce and International Security Division of OTA. "These are very long shelf-life documents," Blair explains, "and our goal is to provide a research tool for our traditionally demanding users."

"We started distributing this information, bought and paid for by the taxpayers of this great nation, as raw ASCII on an FTP server in the early days of the Internet," says Blair. "We are now distributing an entire history of the agency on a five-CD-ROM set, including a history of the way OTA worked in our original mission."

In addition to CD distribution, the OTA Studies are now mirrored on a myriad of robust sites on the World Wide Web. A fragile paper collection has attained global digital immortality.

If we are scanning Moby Dick, these batches might be called

<b>Moby0001.tif</b>	<b>Pages 1-50</b>
<b>Moby0051.tif</b>	<b>Pages 51-100</b>
<b>Moby0101.tif</b>	<b>Pages 101-150</b>

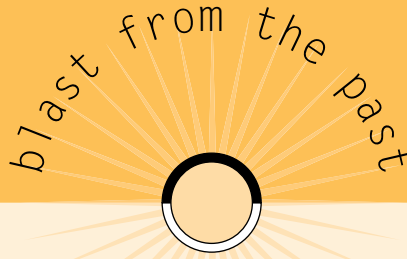
Of course, they could just be called Moby1, Moby2, Moby3, etc., but in this example the file name itself lets the user know which page the batch starts on.

At the end of the process, the document might be called MobyDick.pdf.

Note that the 8.3 convention of FILENAME.ext is by far the most widely acceptable file name format on the Web, and for that reason it is often recommended to Mac and UNIX authors.

Since these files are being handled with such minimal information attached, it is important that they be managed through the system as the blind, clumpy files they are. Files can be handled and cleaned out just like temporary files and unique names, so they should be deleted after one or two warnings.

If the documents are going to be captured and archived under a more orderly system, it is important to find the most common denominator of understanding in the basic index fields.



## Alan Kay's Original Dynabook

The Dynabook, as described by Kay in 1968, and by some accounts modeled in cardboard, was about the size of a three-ring binder and could fit into a bookbag. It had a flat screen display and wireless communications. The screen was envisioned as a touch-sensitive Liquid Crystal Display that could serve as a keyboard when necessary.

The vision has been almost realized several times, by Xerox, Apple and other bold pioneers. Today's most important development, predicted long ago by Alan Kay, is an Intelligent Agent software, which may open the door between human and computer information processing. The availability and quantity of information is already drowning even the most devoted users. It's time to stop surfing and start sailing on seas of info.

The future Dynabook can go to the beach, or into the bathtub, or out to a peaceful park bench, and still be connected to the global digital library, and the global community of connected users who are within instant reach.

# Competing With The Comfort Of Paper

Paper documents are so familiar that the intelligence built into them has become transparent, invisible. In the 500 years since the development of movable type in the Gutenberg Press, printing has taken on all of the rich nuances of spoken language. The very appearance of words and paragraphs tells us a great deal about the information embodied therein. Headlines lead us around a number of stories, at once telling us the main point of interest and defining the location and shape of the article. Within the text, punctuation shows us the flow of the facts, and symbols like quote marks sharply define the action and actors.

This is a vast, even infinite, topic, and this book examines the depth of information embedded in the appearance of the document. However, the easiest and first way to archive and disperse a document is to make copies of it.

Of course, to old-fashioned purists, paper books will still have their allure. They can be easily scribbled on and otherwise marked, and they can be read in almost any light or environmental conditions. Books can contain your father's signature, the weight of his hand and the strokes of his handwriting. It will be at least a human generation or two before digital books can carry such physical presence and emotion, if they ever can.



# Working With Digital Documents

A key consideration is the large file size of scanned images. The rule of thumb is to estimate 50K per page image. It should be noted that this is based on a simple text page, and the file size of pages including graphics and fine text could be much higher. In addition, the textual contents of the page will make up another 1,000 or more characters in overall file size. Once again, the estimate of 2,000 characters of text would be on single-spaced, typewritten page. A typeset page with small text could contain two to five times more characters, while a simple memo or letter may be only a few hundred characters. In addition, to provide text searching capability, an index is included with the page or document, adding more characters to the size of the file.

Acrobat Portable Document Files are much smaller than the aggregate file described above while still retaining the appearance of the original. Acrobat Capture retains only the image of graphical elements that can not be converted to text. The page layout and the appearance of the text itself are preserved, including type styles and font attributes such as size, bold, italic and underlined. This conversion to text greatly reduces the overall file size.

Another important issue of digital documents is the comparison between content and appearance. Let's review the benefits of image and text.

1. Image files faithfully reproduce the look of the original page and document = Appearance
2. Text files allow word-by-word searching of the data in a document = Content



The ideal solution would provide an accurate depiction of the *appearance* of the document, which can be navigated through, and a complete text version of the document for searching the *content*.

Adobe's Acrobat files are the Internet version of PostScript files. PostScript became a global de facto standard because it formed a common language that could reproduce richly composed documents on a variety of hardware and software platforms.

Software and hardware develop separately, but both develop prodigiously. As the many editing packages blossomed for each operating system, laser printers evolved to outperform earlier generations of very capable dot-matrix printers. Prior to PostScript, each word processor and page composer required a dedicated printer to produce even the most basic pages. Electronic documents are evolving on the same path to richer presentation that computer-generated paper documents went through in the '80s and early '90s.

tip

**Page composition has been developed over 500 years, always with the goal in mind to present information more densely and in a more orderly way. The logic behind this design trend becomes very simple if you consider paper as very precious. To conserve this highly valuable resource, scrupulous use of page space is critical.**

**Now, we have to constrain ourselves to the much less dense medium of the computer display. With a tiny fraction of the resolution to present complex images, we need to be very crafty and parsimonious in our presentation of information. Considering the patience of our busy readers, we must design our information vehicles to deliver the goods quickly.**

- Text Only:** Original pages are converted to formatted electronic text  
*Benefit:* Results in the smallest files, fully text searchable  
*Disadvantage:* Discards the image of the original page
- Image Only:** Original pages stored as full-page bitmap images without converted text  
*Benefit:* Provides exact copies of originals  
*Disadvantage:* Search and retrieval limited to a few index fields
- Image + Text:** Original pages are stored as full-page bitmap images with links to text  
*Benefit:* Allows text searching and retrieves images of originals  
*Disadvantage:* Results in the largest files

## Evaluating Access To Text

Text can be considered the opposite of image as a type of digital document. Whereas the image faithfully reproduces the form of a document, the text makes up the content of a document. Though this is not a perfect comparison, since pictures and graphs are part of the content of many documents, it serves to illustrate the primary difference between image and text.

Until relatively recently, text was the only practical content for online services due to hardware limitations. The capacity of 1200/2400-baud modems was limited, as were the user terminals. Very early LEXIS/NEXIS terminals were dedicated teletype machines, which were completely incapable of reproducing graphic images. Fax was the only answer for remote use of "images" of documents, which, of course, was still a great improvement over the mail.

If the density of the information in a computer file could be weighed, a text file would be an extremely dense file. Information, in this case, being defined as data that can be easily read by a human. In a pure text file, there is nothing but the barest of layout elements, only tabs and carriage returns, to create the presentation or appearance of the information. Almost every single character in a text file is data.

tip

A single-spaced one-page letter typed in 10 or 12 point is comprised of approximately 2,000 bytes. Since the image of the same page, a compressed 300-dpi image, will be about 50,000 bytes, text will carry about 25 times more data than the image of the same typewritten page.

---

**OCR: Optical Character Recognition, software that converts scanned images of documents to text and data; contents of documents can be searched by word and phrase.**

The single greatest disappointment of optical character recognition is that the results are not perfect. Forget the fact that many people can not read and write perfectly or type accurately. Users expect OCR to produce perfect results. Why OCR bears this unique burden is a mystery—it probably has to do with every human’s sci-fi-propelled desire to beat the computers. And OCR is one of the few things that computers have tried to do that they don’t do perfectly. (Stifle the snickering.)

In most applications of OCR, the accuracy rates will be around 95-99 percent. This means that there will be unrecognized text in the output of raw OCR processing. Even when OCR gets 99 out of every 100 characters correct, there will still be 20 errors on the average page that need to be fixed. Since these corrections are done by people, they occur at human rates rather than the super-human rates of OCR. The costs to clean up the OCR process are always the largest component of the overall costs of an OCR project.

tip

The value of the information becomes the pivotal decision point. For example, you might scan and automatically recognize 10,000 pages of information for \$5,000 worth of hardware, but it could cost \$50,000 in labor costs to clean up the scanned documents. Thus, it’s crucial to evaluate the real value of expensive 100-percent accurate information compared to 95 percent or greater accuracy on automatically generated information.



## Enhancing Access To Text

What makes digital documents better than paper documents is instant access to information within the documents. Most current document enhancement techniques are simply the computerized versions of proven traditional methods. The cut-out thumb tabs of large hardback dictionaries are the original bookmarks, cut on a beveled angle to allow readers of a 5,000-page book to arrive roughly in the right area of the book.

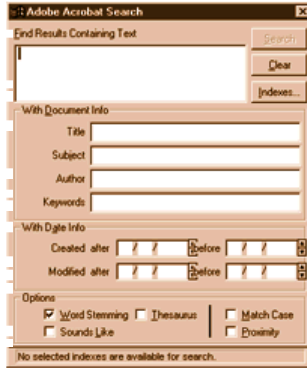
Of course, bookmarks in digital documents can be more precisely accessible, and there can be far more of them, in almost infinitely nested sublevels. Once again, digital documents are more valuable precisely because of instant access, not just the physical advantages of tiny size and easy transmissibility.

If there had to be one single differentiator between paper and digital documents, it's hypertext links. While links are built on conventional footnotes and references, they don't just reference another document, they provide an instantaneous path to the other document. These links automate the paths of reference through information, enabling study at a real-time pace. Users may follow their inspiration to rapidly pursue specific ideas in vast seas of information.

Deciding which tools to use becomes rather obvious with practice. Key fields, such as those inserted using Adobe Acrobat at the time of scanning, may be the fastest route. However, it all depends on the thought process of one very important person: the key word creator. If he or she is familiar with your operation and documents, your key words can accurately reflect your information-searching needs.

Alternatively, if you chose to do a full text search, you'll have to carefully construct your query. If it's too tightly constructed, you won't find anything. If your search terms are too vague, you'll get hundreds of hits and wind up mostly frustrated. If your search is just right, you can successfully use the tool with little work on the inputting side.

Consider the nature of your material and the types of searches you're likely to do before embarking on your scanning project. That preliminary time investment will pay off handsomely in the near future.



### Acrobat Search

#### Left

General Info available for every PDF file, some of which is entered by the author and some is derived from the applications used to create the file.

#### Center

You can customize the performance of the Search Engine under File-Preference-Search in Acrobat 3. Using any combination of criteria, you can refine the search for more useful results.

#### Right

Search results can be sorted and highlighted to your preference to expedite access to files.

It is an easy mistake to assume that full text searching is obviously better than any index database. After all, we know what we are looking for and can just search on the right terms, right?

In extremely large databases or constantly expanding databases, like various search engines on the Web, it's handy to focus the search function. Here's where investing in intelligent index information to a large collection of files pays off. Like key word insertion, time devoted to categorizing and tagging files adds immeasurably to the value of a collection of documents.

**The limit of indexing is not due to any fault in the technique or the person doing the indexing. The limit of indexing derives from the unknowable needs of future users. The reason for an index is to be able to rapidly search an overwhelmingly large database.**

Chemical Abstracts, Biological Abstracts and Current Contents are all designed to provide scientists with a very up-to-date awareness of all scientific papers being published in each particular field. These few, specialized publications keep scientists informed on literally thousands of technical and professional journals.

These tightly focused secondary publishers add value to an otherwise impenetrable avalanche of data. For example, in a typical edition of a scholarly journal, several papers will be published. The value and reliability of the information in each published article has usually been criticized and judged by a jury of peers. Only worthy material gets published.

This material goes through a second round of judges, when the editors of the secondary publishers decide which articles are worth entry into the database. This database is called a Bibliography and contains at least the following fields of data.

#### **Info that is specific to the article**

Title	Affiliation	Key Words
Author	Abstract	

#### **Info that is specific to the published journal**

Issue	Date	Pages
-------	------	-------

Even when the finest minds in the world are working on the job, and every present-day topic of interest is being faithfully tagged and preserved, there is more information here than can ever be packaged in any sort of abstract or bibliographic index.

308196-1

Amer. J. Bot. 69(9): 1410-1419, 1982.

**MORPHOLOGICAL STUDIES OF THE NYMPHAEACEAE:12.  
THE FLORAL BIOLOGY OF CABOMBA-CAROLINIANA**

**EDWARD L. SCHNEIDER AND JOHN M. JETER**  
Department of Botany, Southwest Texas State University, San Marcos, Texas 78666

ABSTRACT

Observations have been made on the pollination ecology of *Cabomba caroliniana* Gray in Texas. Flowers are numerous with morphologically similar perianth parts. The adaxial corolla spurs are nectariferous and attract small Diptera (e.g., *Nesophila cerasini* and *Hydrobia bilobiflora*). Anthesis occurs for 2 consecutive days with flowers opening about 10:00 a.m. and closing around 4 p.m. on each day. First-day flowers have short, indichneous stamens and longer pollen-receptive stigmata which arch outward over the nectaries. In 2nd-day flowers the stamens have elongated to the level of the stigmata and narrow dehiscence occurs above the nectaries. Stigmata of 2nd-day flowers are pressed together at the center of the flower and are nonreceptive to pollen. Insects attracted to 2nd-day flowers in search of nectar become dusted with pollen (due to the position and extensive dehiscence of the anthers) and as insects fly to 1st-day flowers, achieve cross-pollination by virtue of the stigmata position over the nectaries. Seed anatomy is similar to that of other nymphaeaceous genera (i.e., abundant perisperm, little cellular endosperm, a histiolar nucellar "tube," and a small dicotyledonous embryo). Pollination morphology and comparative xylem anatomy support the segregation of *Cabomba* from the Nymphaeaceae, sensu stricto. The anatomical correlations between seeds and the morphological pollination syndrome (found elsewhere in Nymphaeaceae, sensu lato), however, suggest a phylogenetic relationship.

ALTHOUGH STUDIES dealing with the reproductive anatomy and morphology of *Cabomba* are numerous (e.g., Baillon, 1871; Caspary, 1888; Raciborski, 1894a, b; Chiffot, 1902; Cook, 1906; Fassett, 1933; Golteniewska-Furmanowa, 1970; Mosley, 1958; Wood, 1959; Ranji and Padmanabhan, 1965; Padmanabhan and Ramji, 1966; Riemer, 1966; Riemer and Hainck, 1968; Gregory, 1974; and Inamdar and Alekshy, 1979) there has been little research conducted on the topic of floral biology. Tarver (1976) and Tarver and Sanders (1977) investigated the floral biology of *C. caroliniana* in Louisiana. They observed that anthesis occurs over a 2-day period with flowers opening about 10:30 a.m. and closing about 4:30 p.m. during both days. They further observed protogyny as well as changes in the floral morphology by noting that 1st-day flowers have short stamens, about half the height of the carpels, and that the carpels are bowed outward so that stigmata are oriented toward the petals. In 2nd-day flowers, however, the anthers have been elevated to the level of the stigmata, and the stigmata are contiguous at the center of the flower. Pollen release was observed to occur about midday of the 2nd day. Tarver and Sanders further observed that seeds are produced only by flowers which have been visited by flying insects, for caged flowers did not set seed. They determined that the common honey bee, *Apis mellifera* Linnaeus, was the dominant pollinator of *Cabomba* in the Louisiana populations studied, though unidentified halictid bees were also seen in the flowers.

It is the purpose of this investigation to 1) collate and amplify the literature dealing with the floral biology of *Cabomba*; 2) investigate the mechanism(s) of pollination and the relationship between floral morphology and insect visitors; and 3) compare the floral biology of *Cabomba* to the floral biology of other Nymphaeaceae, sensu lato, with the possible goal of contributing to the understanding of both the evolutionary origin and adaptive radiation of the nymphaeaceous flower and the interrelationships among the genera of the Nymphaeaceae, sensu lato.

**MATERIALS AND METHODS**—Two populations of *Cabomba caroliniana* were studied. One locale was in the San Marcos River, San

1410

- Title
- Authors and Affiliations
- Abstract
- Some journals include Keywords here
- Body Text
- Italic (& Latin)

Even the smartest librarian and the most insightful indexer can only assign code words to any present-day document based on the present day's interests. Ten or 20 years from now, or even months from now, something could happen to make an entire body of work extremely interesting again.

Researchers relying upon index information are strictly limited to the indexer's specific field of focus at the time of publication. Beyond this categorical information, the information in the source documents is not searchable.

A skilled searcher can individually retrieve documents and can individually peruse and read documents of interest. In this way, critical information may be unearthed that was never considered when the documents were being indexed, archived and filed away.

# Scanned Document Management: Access And Storage

Because image files themselves contain no data, it is critically important to assure reliable storage systems. A collection of images with no index is similar to, but worse than, a pile of unmarked microfilm rolls. A roll of film may contain a few thousand pages, while a common home PC with a 2.5-gigabyte disk might contain 50,000 images. Reliable storage is essential.

However, the heart of the matter is to ensure that index data is firmly attached to source data. A giant batch of unindexed TIFF files reduces every bit of information to a needle in a haystack. Finally, a backup of the index of large collections of image files is a critical requirement to ensure the viability of an image database.

Images are also exceptionally large files, and they usually cannot be made smaller by data-compression methods. In modems and other communications hardware, files can often be compressed during transmission. This feature allows modems to perform at greater than 100 percent efficiency. For example, a 50K text file might be compressed by 20 percent to 40K for transmission and uncompressed back to 50K on the receiving end. Thus, even though the modem ran at normal speed, 20 percent more data was transmitted. However, since images are compressed when they are created, there are few additional benefits to be gained.

These large images place unusual demands upon storage and network resources. To provide acceptable performance to the user, images should be delivered in small batches. For example, while retrieving a 20-page document, users prefer to view the first few pages while the remainder is being retrieved. This allows users to quit downloading irrelevant files before the entire document is received.

In 1995, Netscape Navigator introduced a new approach to this technique. The software downloads a very low-resolution sample of a GIF image at first, then fills in the image area with a barely recognized shape. Subsequent refreshes of the screen continue to sharpen the detail until the entire file is delivered and the final image appears. In 1996, Acrobat 3 applied this same technique to enhance PDF files, allowing more efficient access to large documents on the Web.

## **Benefits of Acrobat 3**

Download a page at a time over the Web

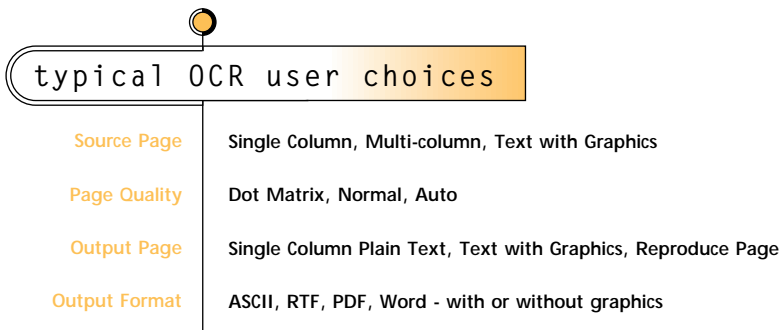
Verity SearchPDF display hit highlights in PDF on the Web

Pages download in most efficient order, text first, outline fonts, graphics, filled fonts

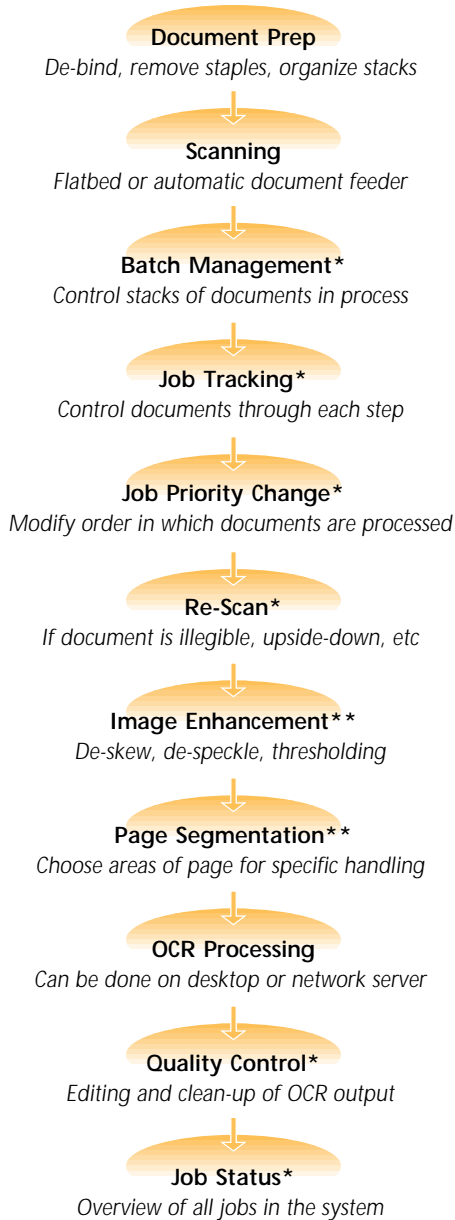


# Getting From Paper To Digital Format

The process of scanning and recognition is the gateway through which most paper-based information will be brought over to the world of instantly accessible information. It is very important to understand that the tools of scanning and recognition are not sledge hammers; they are very fine machines with subtle, powerful controls.



# paper to digital – basic steps



All functions marked with \* are automated on network systems, whereas on desktop applications all of these functions are either manual or not done at all.

Functions marked with \*\* are optional, and may be automated through scanning controls and document templates.

# Document Preparation: The Critical Foundation

The clearest way to think about the importance of document preparation is to consider the two forms of a document. You can hold a dozen double-sided pages in your hands, you can move them around, you can turn them over, shuffle the stack, flick off an imperfection; you have total control of them.

Once you put those dozen pages through a scanner, you have a 1-megabyte file of 24 separate images. Your scanner might relate them into a single stack for you, but you now have an invisible Binary Large Object, or BLOB, that the user has to identify and handle inside of the computer.

Compared to fiddling with a demanding set of computer tasks, those good old paper pages start looking real good. You could turn them over, you could smooth out creases, you could hold them up to the light ... yes, those paper pages sure had a lot going for them.

tip

**Users of digital documents will never have the simple powers and pleasures of holding onto paper, and it is up to the people doing the scanning and digitizing to create the best possible digital documents.**

Scanning a paper document and converting it into digital form should be considered a one-time event. It's not something you take your best three cracks at like the Bell Ringer Sledgehammer on the Boardwalk.

tip

**To expedite scanning, proper procedures should be in place to:**

- 1. Reliably track paper documents before, during and after scanning**
- 2. Use every available hardware & software option to cleanse and perfect each document**
- 3. Process each image in a defined, traceable job flow, with easy fixes and error correction facilities handy**

## Handling Bound Documents

To a librarian, the thought of chopping off the binding of a book is absolutely abhorrent. On the other hand, the thought of a book decaying on a shelf compared to a book on the Internet is an ever-more moving emotion.

There can only be so many books, and they will be lost over time. Users will take them out of the library and never return them. They will be so popular that they will be used up, and the bindings will fail, and pages will fall out, and eventually the covers will be lost.

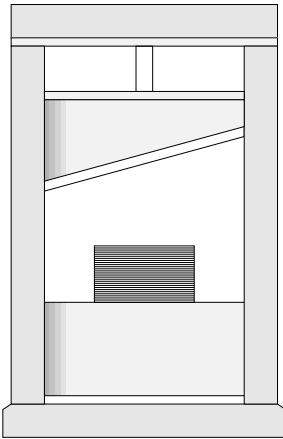
Books in electronic form are not subject to such decay. A remote user can have full use of a complete and rich version of a document or volume, while the "original" is never touched. By the same token, many more users can get to these original documents than could have ever seen the paper documents.

If all of the above arguments do not make sense, there are special book scanners that allow rare or delicate works to be handled with the least wear and tear. One of the earliest models was introduced by Minolta in 1995. It included an adjustable bed on which the book was laid face up. The bed itself was separately adjustable on each side so that as the pages were turned, the left and right sides would move slightly up or down to compensate. Mounted above the book on a tall tripod, the scanner camera looked down and used auto-focus to create clear images of the pages.

This gentle machine preserves the original book and delivers excellent page images. Compared to the alternative of pressing the books themselves down on the scanner platen, or trying to squash them on a book edge copier, the book scanner is a welcome device. However, the physical labor involved in this process is quite considerable.

If, on the other hand, at least one of the above arguments about productivity makes sense, documents should be "unbound" or "debinded" before scanning. The labor savings provided by automatic document feeders is overwhelming. The least expensive sheet feeder on the cheapest available scanner will always be several times more productive than a person shuffling pages by hand.

The best way to remove the binding is to use a heavy-duty guillotine paper cutter. It is important to keep the pages as square and straight-edged as possible to facilitate scanning. The paper cutters found in commercial print and copy shops go through thick books like butter.



An idea that makes book aficionados cringe, the book guillotine is a useful tool for scanning in the digital world.

If the documents have been bound by less-easily-dispatched methods, more labor will be required for document preparation. The primary goal should be to make the pages ready to go through the paper feeder. This means the pages should be stacked evenly, with nothing to make the pages stick together during feeding. For example, the rough holes left by a hastily removed staple will usually be enough to keep the page more or less fastened by the paper of one page being punched through the next.

These tiny imperfections can lead to maddening rework during scanning. Document preparation is hard work, but doing it well saves time and work down the line.

## Handling Single- Or Double-Sided Pages

While business correspondence documents are single-sided, published reports and other complex books are usually double-sided. Most scanners will scan only one side of the page on each pass. Software is used to simplify the process of scanning double-sided documents by incrementing the page count during scanning. For example, the tops or fronts of all of the pages are scanned first; then the stack is flipped over and scanned again. The software will count the first stack as 1-3-7-9-49; when the stack is flipped, the software will reset the counter to 50, and decrement the count as 50-48-46-2. The end result is a properly ordered set of page images from 1 to 50.

**t i p**

**For best results, always scan and OCR in batch sizes that you can stand to lose. The physical labor in scanning and creating the batch may be lost, and will usually have to be redone, if something happens to the batch during the image processing phase.**

## Paper Weights

Besides pages that are crumpled or stuck together, the most common concern in scanning stacks of documents is varying page thickness. Paper weight is the measure used to compare the weight and thickness of various paper stocks. A 24-pound bond paper is thicker and heavier than a 20-pound bond. Heavier paper stocks tend to have a grainier, or leathery, rough surface. Lighter stocks tend to be thinner and smoother.

The rollers in a scanner are set to the hundredth, even the thousandth, of an inch, to handle this variety. Rubber rollers are used to add another variable of flexibility. But all scanner feeders have limits in the range of paper thickness they can reliably handle. Card stock covers on reports, for example, are actually 70- 110-pound paper, or light cardboard. A single piece of cardstock is as thick as several pieces of paper and will often jam in the rollers.

At the other end of the spectrum, some paper is too light to reliably feed through scanners. Some very thin onion-skin paper, often used for typing academic papers and manuscripts, is only 10-pound stock, and much thinner than 20-pound bond. Glossy magazine pages, with their hard, slippery finish, are particularly challenging to feed through a scanner.

**t i p**

**If all else fails and the scanner can not be adjusted to handle the paper documents, the fall-back position is to use high-end copiers to create acceptable pages for scanning. It sounds primitive, and it is. But when all else fails, options like copy-ing start to look pretty good.**

## Fragile Handling

Fragile handling is needed for both “ancient” documents and those stored on very cheap output, such as thermal fax paper and carbon copies. Entire careers of famous chemists and physicists reside in skimpy loose-leaf binders and notebooks. Many digitization requirements become labors of love, where the work itself evokes sufficient dedication to overcome all of the difficulties.

## Mixed Orientation

Any image-processing program that uses OCR can theoretically deal with mixed orientations in the scanned images. Software can look for words to determine the baseline, and thus the orientation, of the page.

Any image-processing program that uses OCR can theoretically deal with mixed orientations in the scanned images. Software can look for words to determine the baseline, and thus the orientation, of the page.

2% (OCR works)

Any image-processing program that uses OCR can theoretically deal with mixed orientations in the scanned images. Software can look for words to determine the baseline, and thus the orientation, of the page.

7% (OCR fails)

Any image-processing program that uses OCR can theoretically deal with mixed orientations in the scanned images. Software can look for words to determine the baseline, and thus the orientation, of the page.

15% (human has difficulty)

Examples of skewed text

Orientation of the page is absolutely critical in applications where there is no control over the incoming images. A fax server is the perfect example of this requirement, where literally hundreds of different styles of fax machines are the sources of the input images, controlled by unknown people. No fax server would ever be designed to expect the pages to come in right side up. However, this feature requires a lot of CPU processing power.

More expensive scanners, designed for high-production operations, offer single-pass double-sided scanning, where both sides of the page are scanned at once. These scanners are driven by software that automatically keeps track of the images being produced by each of the two sets of Charge-Coupled Device (CCD) arrays, and orders the images into a single, consecutive file. This duplex capability has recently become available on less-expensive scanners.



**Charge-Coupled Devices are the eyes of modern scanners; they translate reflected light to digital information. CCD arrays are miniature rows of these light-sensitive devices that track and translate every pixel (picture element) on the page into white, black, gray or colored information.**





## Handling Graphics And Illustrations

To just touch upon the topic of scanning richer images, in gray scale and color, it is important to consider the combination of these images with text in the source documents.

There are several collisions between rich graphic scanning and text.

1. Rich color graphics that look beautiful on the average computer monitor are far different from text on a screen. The most common color picture format is called GIF, or Graphic Interchange Format. It's so popular that GIFs have always been handled like native files on the Web, also called inline graphics.
2. GIFs were designed to show full-color graphics, mostly scanned photographs, on the typical PC monitor. The typical VGA monitor has a resolution of about 72 dpi, so resolution above that just deteriorates the picture on the monitor.
3. Text at any resolution less than 200 or 300 dpi will produce poor OCR results.
4. Except for programs that are designed to work with HP AccuPage and other similar software, OCR requires binary images. Color images are not suitable for OCR.

You'll have to decide what features are most important when establishing scanner settings.

## Scanning Techniques

Document scanners share all of the same optical and paper handling machinery used in copiers and fax machines, so the digitization of paper pages is a common and reliable technology. Because the actual "magic" of transforming black and white bits into computer images is being done on ever-cheaper silicon, scanners will be cheaper, sharper and faster with each new generation of desktop processor.

## Small, Medium & Big Scanning Configurations

<b>HP ScanJet IIIc</b>	~ \$ 1,500 (including feeder)
Resolution	200 - 600 dpi Optical resolution, up to 2400 dpi extended
Scanning Speed:	4 Pages per Minute
Comment	Clearly superior for gray & color scanning applications
<b>Fujitsu SP 10C</b>	~ \$ 1,500 (w/ SCSI Interface)
Resolution	200, 240, 300, 400 dpi
Scanning Speed	12 Pages per Minute
Comment	Clearly superior for faster document feeding and scanning
<b>Fujitsu 3093</b>	~ \$ 6,500 (w/ SCSI Interface)
Resolution	200, 240, 300, 400 dpi
Scanning Speed	27 Pages per Minute
Comment	Durable, productive scanner for department use
<b>Fujitsu M3099</b>	~ \$28,000 (w/ SCSI)
Resolution	200, 240, 300, 400 dpi
Single-side Speed	80 Pages per Minute
Double-side Speed	120 Page / Images per Minute
Comment	High-speed throughput, enhanced image processing

The raw output of scanners can often be enhanced to make the finished images more suitable for specific needs. All of the following processes can be performed in either hardware or software, usually depending upon overall throughput requirements. Especially in very high-volume installations, dedicated hardware solutions may be more productive than software running on standard workstation processors and SCSI interfaces. The point is, these processes are indispensable enhancements to any document scanning and digitization system.

The most important adjustments you can make are the brightness and/or contrast settings. Brightness is directly comparable to setting the f-stop in a camera, controlling the amount of light that hits the film. In a scanner, a thin, long strip of light-sensitive computer chips passes over the image. These chips are the CCDs that measure the amount of light reflected from a given spot to determine if that spot is black or white. In color and gray-scale scanners, the CCD actually measures the wavelengths of the reflected light to accurately reproduce shades of gray or color.

Adjustments to the brightness setting change the receptivity of the CCD. This allows the user to precisely control the output of the scanner. The ideal brightness setting brings out the best in the material of interest and can drop out unwanted background.

Contrast is so similar to brightness that many scanner programs combine these two adjustments into one. While brightness makes the entire image lighter or darker, contrast lets the user tune the scanner for the difference between black and white, or any other gradient of light.

Adjustments to contrast can sometimes be used to remove speckles in the image because it requires more or fewer pixels to be recognized as a speck. Higher contrast may see five or seven touching black pixels as a speck on the page image, while lower contrast might ignore so few pixels and portray the entire area as white.

## Scanner Specifications And Settings

**Resolution:** The finest detail that can be discerned by the scanner, usually measured in dots per inch (dpi). Typical resolutions are 200, 240, 300, 400, 600 and finer. OCR is usually done at 300 dpi to provide enough detail for accurate recognition, without capturing too much detail and ending up with overly large files.

John Solomon, a seasoned veteran in the field of document digitization, won the contract for the OTA conversion. Solomon, Vice President of Input Solutions (Gaithersburg, MD), has been dedicated to the field of high-end document recognition since his years selling the first intelligent OCR machines, invented by Ray Kurzweil.

Solomon's observations on this project are very enlightening: "We tested at 200, 240, 300 and 400 dpi. One thing that we consistently saw was that higher resolution scans yielded higher OCR accuracy with Adobe Capture." Solomon is someone who focuses sharply on OCR accuracy, but he also

described another benefit of higher resolution scanning. "Graphics clarity improved at higher resolutions, and moire patterns on halftones virtually disappeared.

"The Fujitsu 3099A has 100 levels of brightness, where most scanning software offers about 15 levels of adjustment. The 3099 also has 100 levels of density, and we were lucky to be able to take advantage of this extra-fine level of adjustability. We found an individual with a great eye for how the scanner would see a page. Because the typical document in this project was long and dense, we tested each book before scanning. The operator's visual judgment let us run just a few test scans to optimize performance for accuracy and speed."

**Image Type:** The scanner's ability to recognize binary (black & white), gray scale or color. Binary-only scanners often have the ability to dither the image to create a digital halftone of a gray or color image. OCR and most commercial document imaging applications use binary images. Scanning in gray scale and color is an art form unto itself.

**Scanning Speed:** The number of pages per minute (ppm) than can be scanned, usually through the ADF (automatic document feeder). On most scanners, higher resolution results in slower speeds. A scanner rated at 20 pages per minute at 200 dpi may only do 12 pages per minute at 400 dpi.

## Straighten Images Via De-Skewing

Image de-skewing assures straight images through a feature called edge detection. This is a function that looks for long straight lines, most particularly at preset parameters for the page. Edge detection sees long areas of black pixels forming a long border and determines whether this border is close to the horizontal or vertical axis. The software then rotates the entire image to match the 90 degree orientation of that edge. This feature is sometimes expanded to include border removal, a process that deletes that entire area of black pixels that was seen as the "edge" in the de-skew stage.

The process of digitizing OTA reports began with the “de-binding” of the documents, accomplished by an electric guillotine slicing off the spines. The resulting stack of pages, usually double-sided, was then fed through the high-speed scanner. The Kofax 9275 Image Processor de-skewed the images on the fly, but the throughput was far below the scanner’s advertised speeds at 200 dpi.

The processing of the documents was done with the Pentium Scan Station and the Fujitsu 3099A, feeding a queue on an HP NetServer, which had a 100 Mhz CPU, 64 MB of RAM and dual 4 GB drives. A Pentium/133 with 32 MB of RAM and a 2 GB Fast SCSI hard drive was designated as the Capture conversion server, running over Novell 4.1. Solomon noted that the Fast SCSI offered dramatic improvement in

throughput on the Capture server. Editing and cleanup was done on 486/100s and P75s.

“Our target was 200 pages per day for our editing and cleanup operators. Our best operator hit 373 pages per day at the end. A lot depends on the source document, and a lot depends on trying hard,” Solomon declares.

Five CDs, containing 23 years of pure research into technology and the future, can sit in a single tiny player. That five-disc player on a PC Web Server provides wide-open, free access to the rich reports. The Office of Technology Assessment will live on in perpetuity as a “virtual agency,” mirrored on servers around the world, part of the global system that evolves from the fertile World Wide Web.

## Setting Resolution

Document scanners offer varying levels of resolution when scanning, which is measured in dots per inch. Each dot, or pixel, represents the light reflected from a minute point on the page. The proper resolution for each application is dependent on two factors:

1. How the documents will be used
2. The processes the documents will go through

Image Compression		Comparisons
DPI	Bytes/Sq. Inch	Uncompressed* (8.5 inch x 11 inch page)
200	40,000	0.44 Megabyte
300	90,000	1.05 Megabyte
400	160,000	1.77 Megabyte

\* Most documents are compressed in practice, but relative sizes remain fairly constant.

In actual practice, image compression is routinely employed to reduce file size, and the general rule of thumb is 50,000 bytes per image, or roughly 20 images per megabyte. The raw image file is compressed most often in an international format known as CCITT Group IV. However, a 300-dpi image will always be larger than a 200-dpi image.

Because even compressed images are such large files, the lowest possible resolution should be chosen. Lower resolution results in smaller files, reducing storage requirements on physical media and reducing transmission time and network load.

Returning to the first criteria for choosing a resolution, the user of the images must be considered. In the commercial document imaging industry, where pages are scanned into workflow and document management systems, 200-dpi images are the most widely used. The users of these documents require an easily readable reproduction of the page, and 200 dpi offers a very good copy. For example, Fine Mode Fax is 200 dpi.

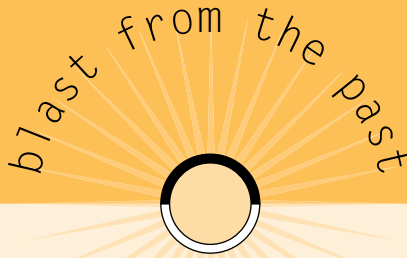
Optimal resolutions are:

- 200 dpi images are fine for viewing
- 300 dpi is recommended for OCR
- 400 dpi and above is restricted to special requirements

## OCR Robots Speed The Process

To make the information on the pages accessible, optical character recognition can be performed on document images, converting digital copies of text into actual computer text. OCR is a robot typist, designed to reduce the need for manually rekeying information from paper sources. OCR is always faster, and in some cases it is more accurate than human typists. Once OCR has converted document images to text, they may be searched for key words and index fields, greatly improving access to the information compared to simple images.

OCR software should be considered as more of a robot than a software program. The distinction is that a software program, like a word processor, spreadsheet or telecommunications program, is designed to automate a repetitive series of computer-based tasks. Optical character recognition exists to take in paper documents, in all their infinite variety, and bring them into the computer's digital world. OCR is built from thousands of rules, but the ultimate variable is infinite, paper documents.



In the very early Kurzweil Intelligent Scanning Systems, a number of expert systems were used to perform character recognition. However, the system had to be "trained" to produce the best results for each document.

In addition to the crucial Brightness settings, the operator was presented with three choices at the beginning of each job:

### **Do 0 (zero) and O (uppercase O) look alike ?**

This would tell the software to emphasize string-checking to assure that alpha characters followed alphas, and numeric characters followed numeric characters.

By so doing, mixed alpha-numeric strings, such as part numbers, became virtually impossible to recognize accurately. But at least the word "book" would be spelled with "o" instead of "O."

### **Do 1 (numeral 1) and l (lowercase l) look alike ?**

A "yes" answer to this question would tell the software that a serif font was being recognized. It would also point the string-checking features to look for unique instances where a "1" would normally be found. For example, if an ambiguous character appeared followed by a period, "l." or "1." it would be recognized as the numeral one. So in a long alphabetical listing, the order would appear as a-b-c-d-e-f-g-h-i-j-k-l-m-n-etc.

### **Do L (uppercase L) and l (lowercase l) look alike ?**

Opposite of Number 2, a "yes" answer to this question would tell the software that a sans-serif font was being recognized, where both upper case "l" and lower case "l" are represented by a plain, vertical line. In this case, an upper case "l" would not appear in the middle of string of lower case letters.

# Fine Tuning OCR

In the earlier section on scanner settings, we discussed brightness and contrast and how they can be used to optimize scanning. There are two types of easily distinguished OCR errors that give a clue that the brightness or contrast settings can be improved: broken characters and joined characters.

**t i p**

**The adjustments that fix problems at one end of the “Too Light - Too Dark” spectrum cause problems at the other end. For example, if the fine characters in italic text are breaking up, the proper correction is to lower the brightness, or decrease the contrast. These same adjustments may cause bold text on the same page to run together. The only solution is to make a choice based on the importance or preponderance of italic or bold text.**

It is not always possible to correct OCR errors by adjusting contrast and brightness. Remember, OCR accuracy is always dependent upon the quality of the original. Having said that, hand the best possible image over to the OCR robot.

**t i p**

**An old-fashioned magnifying glass is an excellent low-tech tool to double-check scanner adjustments. In situations where the problem of broken or joined characters can not be fixed after many settings changes, it can be helpful to take a close look at the original page. Many times the print on the page itself will be actually broken or crashing.**

One reliable method to determine how big to make the size of undesirable specks, to be filtered out by the despeckle feature, is to measure the smallest feature of the subject text. For example, in many fonts, the smallest discernible feature is the dot on a lower case “i.” If the dots on the “i”s are disappearing, the despeckle feature is set too large and is removing more than it should from the page.

Drop-out colors are colors that the scanner does not see. Familiar forms are often printed in very specific shades of pink or green or blue, each with a particular Pantone Color. These colors reflect a wavelength of light that is not picked up by the CCD array of a particular scanner. For example, many early, inexpensive Japanese scanners used a yellow-green light for scanning. Many popular yellow highlighters would be completely invisible to these scanners.



Using the same optical wavelength trick, Kurzweil scanners designed for the office offered red drop-out. This feature allowed lawyers, for example, to mark up a contract with a red pen. The marked-up document could still be scanned accurately because the invisible red marking would not interfere with OCR. When the file went to word processing for clean-up, the typist could use the red marked-up pages to make the desired corrections.

In the section on brightness and contrast, those controls are used to perform this same trick on background colors. Paper tends to yellow and brown as it ages, and this discoloring can sometimes interfere with OCR. By adjusting brightness and contrast to emphasize the text, a form of “drop-out” is achieved by minimizing the color on the paper.

In fact, even gray-scale processing done on images is a form of “color drop-out” where all the least-informative shades of gray in the image are discarded, in the interest of presenting the best possible image to the OCR process.

## Summary

The benefits of digital documents far surpass the effort required to create them. With an ever-growing array of tools, our ability to instantly access information is increasing exponentially.

1 News Release, 2/28/95, Xplor International, The Electronic Document Systems Association

2 News Release, 2/28/95, Xplor International, The Electronic Document Systems Association

3 Gartner Group Briefing, 3/10/95, Jaime Popkin, Electronic Workplace Technologies

4 Adobe Acrobat Product Benefits Study, 9/9/94, KPMG Peat Marwick LLP

5 Source: Lawrence Livermore Labs, Coopers & Lybrand

