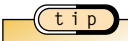
acrobat exchange:

an architecture for instant access

chapter four

Assuming that your organization or corporation has decided to provide broad online access to information in documents (or you have!), the decision moves on to what type of documents will convey the information.

Traditionally, information has been shared in ASCII appearance, as plain text in applications such as Lotus Notes, and on the Internet in Newsgroups and Listservs. With the widespread adoption of Acrobat PDF, information can now be shared online in PDF, which conveys considerably more information than plain old 80-column ASCII.



Design elements in written language go back to the dawn of civilization, to the elegance of hieroglyphics. Cuneiform was a simple language of lines cut into clay, of just the facts, like ASCII. Compared to the simple and limited earlier universal language of ASCII, PDF is an ineffably richer universal language.

When documents contain almost any sort of formula, from simple math common to financial papers to complex equations found in scientific documents, PDF can provide a more accurate representation of the information than plain text can. In addition, artwork and the actual page composition are often critical to the delivery of information—all elements that can be preserved and conveyed with PDF files.

Foreign language documents, including European, Asian and Arabic alphabets, are often unrecognizable ASCII codes in HTML because browsers usually are not set up for anything other than simple alphabets and character sets. Any PostScript application that can use PDFWriter or Distiller can create universal PDF documents that retain their original appearance on the Web, allowing any browser to view the "exotic" document.

Finding Aids: Author, Title, Subject, Keywords

"Meta-information" is information about information. As such, the Info fields in PDF provide a built-in card-catalog-style set of index fields to provide continuity to present information management standards and practices. Card catalogs provide information about collections of books in libraries, and most modern applications from Word to Acrobat provide these time-proven ways of handling big collections of documents, relying on a demonstrably successful method of access to information in huge collections.

One of the intrinsic features of a PDF document, whether it is a simple image or a big file, is the availability of the Info fields for tracking, managing and searching documents in large collections.

Every PDF Document Offers Four Classes Of Data

SYSTEM INFO

System-generated fields, such as Date Created and Source Application

DOC INFO

User-generated fields for Title, Author, Subject and Keywords

CONTENT INFO

Information about the content, such as File Size and Optimized status

ENHANCED NAVIGATION

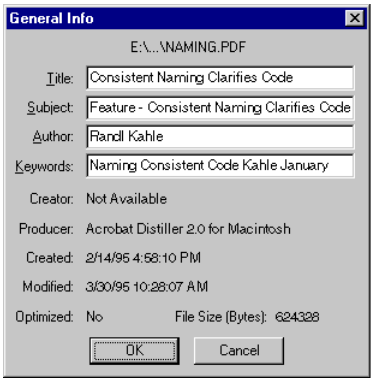
Bookmarks, Links, Thumbnail Views, Articles

Because these fields are part of every Acrobat document, authors and publishers can use them to great advantage. With careful forethought, these few, simple fields can be used to provide an extensive catalog of the contents of a large digital library.

All of these fields are very large-capacity text fields, leading to the possibility of very large indexes. In a pure design sense, the main challenge would be for the publisher to decide ahead of time the top few levels of organization, and especially how to enter the data into those levels. Ideally, the terms entered should be the ones that future users will be interested in searching.

Acrobat General Info Fields: A Foundation For Instant Access

The first four Information Fields can clearly identify the contents of a PDF file: title, subject, author, keywords.



Document information fields are vital tools for quick retrieval.

Title

This is the actual title of the document, as it would appear in a paper publication. This title can be as long as the author desires, but a practical limit of about 32 characters is suggested so that users can easily see the entire title on the screen.

This is not the simple FILENAME.PDF, usually, but it can be used for that.

Subject

The Subject field should let the user browse through a simpler layer of information contained in the article than is described by the Title.

Author

In simple collections, such as small companies or single departments of companies or universities, a simple author name will be plenty. In larger collections, it is very often helpful to at least give the user the chance to sort by last name and first name to separate a particular author's contributions in the event that many authors have the same last name.

At a minimum, it is helpful that a collection follow one convention of listing authors' names. That is, whether full names or initials are used, or whether the names appear

in normal or reversed order (i.e., William Shakespeare or Shakespeare, William), it is easiest for the user if the same convention is used throughout the collection.

In pure Acrobat databases, with the built-in Acrobat Exchange Text Search, the user has the ability to find information contained in a field without strict adherence to any conventions. However, a definite set of rules allows a user to enter more precise, and therefore more effective, query terms.

Keywords

Keywords should be registered as an approved and widely shared set of defined terms. An undocumented list of Keywords is the same as text searching and offers little additional finding value. By their nature, Keywords should be commonly accepted by expert users as uniquely defining some particular concept in the specific database.

Keywords can be especially useful because an expert publisher can provide additional finding aids to documents. For example, Keywords may define ideas or issues that are relevant to a particular document, even though that document does not actually include the Keyword in the text content.

A modern publisher might assign the Keyword "dynabook" to V. Bush's article "As We May Think" because that article contains historical concepts that are relevant to "dynabook." Of course, since the article was written in 1945, it does not contain this word, and therefore the article would never be found in a search for "dynabook" without the Keyword entry.

Health-ifying thousands of great recipes might be possible by adding the Keyword "carob."

All of those devil's food and wacky cakes offer plenty of good chocolate-eating fun. These original recipes for chocolate delights might be re-created with a chocolate substitute, but the original recipes would, of course, never mention "carob." By adding the Keyword, the author offers future bakers many ideas.

PDF Document Info

Then there's information related specifically to the PDF document.

Creator

The Creator field lists the program that created the document. The value of this field is that reproducible bugs will be able to be retroactively fixed, right down to the version number. It is incumbent upon authors to do their best with using the latest, legitimate version of any document creation program.

Producer

This is the actual driver or low-level document builder that wrote the file. Once again, going forward, expect that tools will come out that can automatically refine electronic document systems to make them more useful in the future.

These driver programs are destined to become history, and it will be beneficial to know their idiosyncrasies for future tweaking and translating.

Created

This is a basic file characteristic, and it enables users to quickly find the most recent files, or any file that appeared within a certain time frame.

Modified

This is another extremely basic file characteristic that allows users to make sure they are working with a very specific version of a file, whether it be the latest version or some other specific version of the file.

Additional Fields Customize Document Appearance

The five additional controls in the Document Info file allow the author to determine exactly how the document appears to the end user. The five settings areas are called Open Info, Font Info, Document Security, Base URL and Index.

These five areas are the basis for a simple digital library, and all of these settings taken together serve the purpose of a digital card catalog augmented with a number of finding aids and user convenience features. All of these fields are easily accessible to the casual user and are completely documented in the “pdfmark Reference Manual.” They are widely used in managing PDF documents in third-party products such as text and RDBMS packages.

Open

This option allows the author or publisher to serve up digital documents in a designed and chosen appearance. When the user of the documents clicks on and pops open a digital document, the publisher can control the appearance of the document very directly. The document may open in Full Screen, Bookmarks or Thumbnails mode. In addition, the Page Number, Page Magnification and Page Layout can be specified.

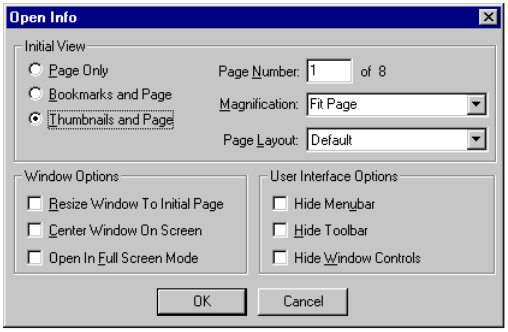
Initial View

Page Only: This setting offers the simplest screen, and just the pages themselves are displayed.

Bookmarks and Page: This setting splits the screen and adds the navigational Finder aids of pre-defined Bookmarks. This feature allows the user to quickly browse-at-a-glance to areas of interest and click to go there.

Thumbs and Page: This setting splits the screen and adds the navigational Finder aids of pre-defined Thumbnail Views. You can quickly browse-at-a-glance to areas of interest and click to go there.

Additional Fields: Page Number, Magnification and Page Layout are additional self-explanatory fields.



Under the Magnification Pull-down menu:

In addition to the Toolbar button choices of Fit Page, Fit Width, Fit Visible, there are multiple Zooms available at percentages of 50, 75, 100, 125, 150, 200, 400 and 800.

And, under Page Layout, the choices are Default, Single Page, Continuous or Continuous Facing Pages. The electronic publisher is faced with page composition and typesetting options, which eventually become obvious.

Windows Options

Windows options include resizing your window to fit the initial page, centering the window on screen, and full screen. Full screen takes up the user's maximum screen presentation to reproduce the original page at maximum page size settings, most closely representing the printed page in complexity and richness.

User Interface Options

User interface options include hiding the menu, hiding the toolbars, and hiding window controls and font info. Font info tells the user which fonts are included in the document, what types of fonts are being used (Type 1, True Type), their encoding scheme, and whether the entire font or a subset has been embedded in the file.

tip

Embedding a font within a PDF ensures excellent fidelity to the author's design by removing the possibility of the user's system substituting fonts derived from his own desktop. However, each embedded font can add 40K to a PDF document. Embedding a subset can reduce this overhead.

Document Security

Basic password control and access rights are built into the Acrobat Document Info structure. This level of security serves the needs of most digital library applications where access to the network itself is secure. That means that a secure library within a password-protected network will offer document security.

Document encryption and advanced security options can be added to the basic password and rights scheme of native Acrobat documents.

Index

This feature automatically opens a specific Index when a user accesses this particular document. On a Web server environment, where occasional users typically require only specific file access, this feature may be an unneeded luxury. However, on Intranet applications, the author may assume that a user of any article in the collection will probably require access to all of the articles in a certain section.

Articles: Reading Complex Documents On A Monitor

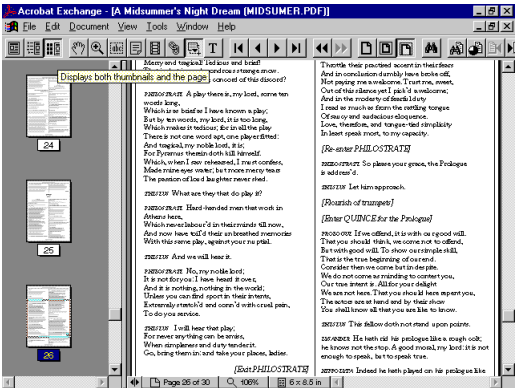
With the Article feature in Adobe Acrobat, the author can build in the simple navigation capability that the user has with his own eyes when reading a magazine. Since a complex page is often too large to read on an average monitor, the user will typically be "zoomed in" to a magnification that presents the text in the most readable size.

The Article feature allows the user to easily follow the text of an article as it is laid out on the page. For example, the end of column one will naturally flow into column two. And the text of the article, and the information content of the article, will flow around all of the various illustrations and insertions on the page. The Article tool allows the reader to follow complex electronic text as easily as he reads the newspaper or a magazine.

Superior Navigation

Navigation is different from search and retrieval of information. When searching, the user will enter specific index criteria to identify a single file or class of files, or he will perform a text search on a collection of documents. Navigation refers to the way the user can move around within the information in an orderly fashion. Navigation is done both within a document and between related documents.

On the World Wide Web, hypertext links are the most common form of navigation, and the user must depend upon the paths built into the collection by the publisher. The Portable Document Format provides several forms of navigation that may enhance the user's access to information.

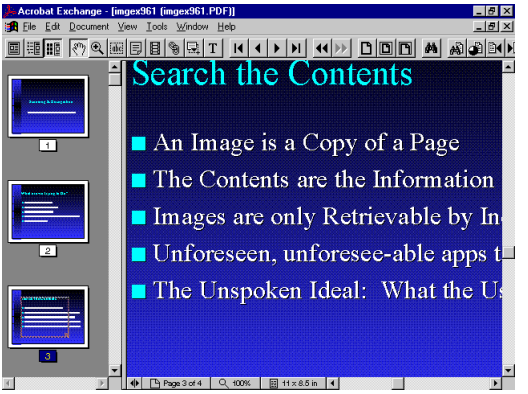


A mouse click provides quick selection of Full Screen, Thumbnails and Bookmark Views.

Thumbnail Views

Thumbnail Views are miniaturized displays of the pages within a document that are not designed to be read but instead offer a quick way to select pages for reading. Thumbnails can be automatically generated by Acrobat Exchange, and they can be viewed in a split screen beside the Page View.

Thumbnail Views are especially useful when the documents include different types of pages that are easily recognizable in the smaller pictures. For example, a technical manual often includes pages of text as well as drawings and diagrams. Rather than flip through pages one by one, Thumbnails provide an easy way to jump directly to pages of interest. Another example includes financial reports, which often include explanatory text along with charts and diagrams. Once again, the reader can quickly scroll through the Thumbnails views and jump directly to the desired information.



Thumbnails allow rapid navigation through recognizable snapshots of pages in a document. By clicking a Thumbnail View, the user goes directly to the fully readable Main Page.

Bookmarks in PDF files are functionally similar to thumb tabs often found in large dictionaries and other reference works. In books, these are rounded notches cut into the edge of the pages, with a labeled tag at the base of each notch.

Bookmarks

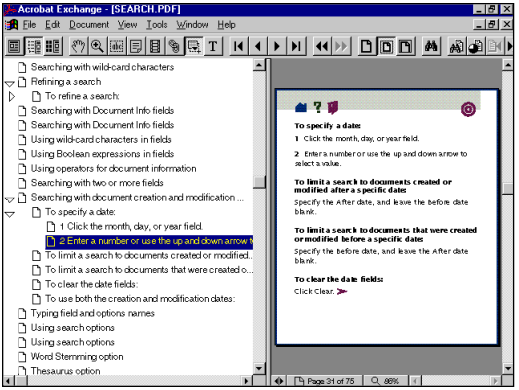
Bookmarks are provided by the publisher to allow the user to view the various topics covered within a document, and to instantly move to the desired area. These Bookmarks are selected and viewed in the same vertical window on the left of the Acrobat display.

Bookmarks can be used to indicate chapters, sections, subsections and so on, or they may be used to indicate charts or illustrations. Many electronic publishing applications, such as PageMaker and FrameMaker, can automatically generate Bookmarks from the header structure of the document.

tip

A very informative source of tips and tricks is adobe.mag, Adobe's online magazine for Web developers, online publishers and others involved in digital publishing. Its archives offer a wealth of how-to articles, and it is a great place to look before embarking upon a project. Adobe.mag is available at:

http://www.adobemag.com



Bookmarks may be up to 20 levels deep.

Links And Web Links

Links in PDF work the same way that hypertext links work in HTML. A word or any part of a PDF document can serve as a link to any other part of the document, or to another document. With Web links, this hypertext facility connects PDF documents to any URL (Universal Resource Locator) so that the links can point out to the World Wide Web.

Web links are particularly interesting because they provide a seamless union of publishing on both CD-ROM and the World Wide Web. Assuming that the user's computer has access to the Internet, whether through a dial-up connection or through a network, a publisher can exploit the *economical publishing medium of CD* and still maintain the *timeliness of an online* offering.

A manufacturer can publish a catalog on CD, providing users with very quick access to large amounts of information. Full-color photography, lavish designs and extensive illustrations can be most economically distributed on CD, whereas browsing through the same information would be very time-consuming on all but the fastest connections to the Web. However, Web links would provide updates to the latest information and any special offers that occur from time to time.

Benefits of Delivery with PDF

The entire raison d'etre for digital documents is improved access to information. But even now, in the very early stages of the Web, the sheer volume of information is overwhelming. We already can see the need for the Intelligent Agent software that Alan Kay has been talking about for the last 20 years. Many busy people don't want to surf.

Information providers and publishers can overcome the drag of non-productive searching and link-following by structuring information in ways that are easily accessible to the users. The combined Index Fields and Text of Acrobat PDF files can be used to offer powerful and coherent organization to very large collections of information.

In most cases, the time spent in creating this organized structure will be repaid a hundredfold in the time savings of the users of the information. PDF also provides multi-platform compatibility.

There are at least a few definitions of the word "platform" in the field of computing. For example, platform may refer to the operating system, such as Mac, Windows, UNIX or DOS. Or platform may refer to the medium on which the files reside, such as Novell LAN, CD-ROM or the World Wide Web.

In the first sense, the operating system platform, Web browsers are available for all three platforms. Therefore, HTML documents written for the Web are accessible from all of these platforms and can be described as cross-platform documents. However, even though Web browsers can be directed to read files from a network or local drive, or from a CD-ROM, the practice is not common yet. On the other hand, PDF documents are widely distributed on CD and over networks, demonstrating the fact that the Acrobat format is best suited for the widest variety of platforms.

Comparing PDF And HTML Print Capability

In an announcement of Adobe PrintGear, president and co-founder Charles M. Geschke made these comments regarding the Adobe vision of the future of document printing:

> *"Here at Adobe, we've been developing printing solutions for 13 years. We think we have some of the best solutions in the world for printing and imaging. We are proud to have developed some of the key technologies that have spurred changes in the market and created the market for desktop publishing.*

"Looking to the future, we see a new revolution, this one driven by changes in connectivity enabled by the Internet. These changes will make the dreams of many people, of **instant access to the entire world's library of information,** *come true. (Emphasis added.)*

"At Adobe, we're creating a series of tools that will help publishers of magazines, newspapers and other sources of information make billions of pages of information simply and easily available to you over the Internet. We know that these changes will dramatically change what role your printer plays in your day-to-day life. Expect to see a series of innovations from Adobe that anticipates these changes in printing so when you buy a printer with Adobe technology, you can have confidence it's been designed to print not just the things you print today, but those you'll print in the future as well.

"PostScript is the standard printing technology for producing high-quality output. However, PostScript is both a language for describing the printed page, as well as an interpreter that resides in the output device. Moreover, there are many different flavors of PostScript based on the many applications that produce PostScript output.

"In contrast, PDF files are highly structured, and general programming constructs are not permitted. As a result, the imaging operations are usually much simpler. Each page of a PDF document is independent of the others. The apparent arbitrariness of PostScript is eliminated, so PDF provides the foundation for a print production system that can deliver consistent, predictable results."[1]

Because Adobe Acrobat software has been built on the foundation of PostScript, the PDF file format currently offers far greater promise for printing than does HTML. HTML was developed to present information on-screen, with an emphasis on individual users having the capability to control the final appearance of the pages. If printing and controlling the quality of the output is a concern, PDF is currently far superior to HTML. And because PDF comes from the same company that created PostScript, continued emphasis on high-quality output will remain a priority.

Precision: Retrieve ONLY Relevant Documents

The late Dr. Gerard Salton, formerly of Cornell University, is widely regarded as the father of Information Retrieval (IR). Dr. Salton crystallized thinking in the field when he defined the two critical measures of IR as precision and recall. The first measure, precision, refers to the ability to retrieve only results that match the query term. The second, recall, refers to the complementary ability to retrieve all results in the collection that match the query.

For more information on Dr. Salton, and his unique pioneering SMART program, try this link. Also use the root URL to get to the Cornell University Computer Science Department.

In traditional database structures, where unique pieces of information are stored in strictly defined fields, precision of retrieval is very high. Of course, applying Dr. Salton's text-retrieval measures to an indexed database is a misuse of his methodology but is informative in this context because Exchange combines traditional Index Fields with FTR.

http://www.cs.cornell.edu/Info/Department/Annual95/Salton.html

http://www.cs.cornell.edu

Dr. Salton explored the less tightly defined collections of information such as text databases and in this sense did not expect to find data in the orderly structure of a Relational Database Management System. Acrobat Search combines structured and unstructured Information Retrieval methods.

Therefore, Acrobat Search has an advantage over pure text retrieval because the query can be focused upon specific subsets of documents and would therefore be naturally more precise than text searching alone, which the criteria of precision and recall were designed to measure. The PDF Author, Title, Subject and Keyword fields can be used singly or together to identify either a single, unique document or a unique class of documents, such as all documents by a particular Author. In addition, the Date Created and Date Modified fields provide precise time sampling of the collection.

Recall: Retrieve All Relevant Documents

The difficulty in retrieving all relevant documents in a collection arises from the inability of the publisher and user to think in identical ways for classification and organization. Once again, the combination of Index and Full Text information in Acrobat Exchange provides a wide and fine net for finding relevant documents in large collections.

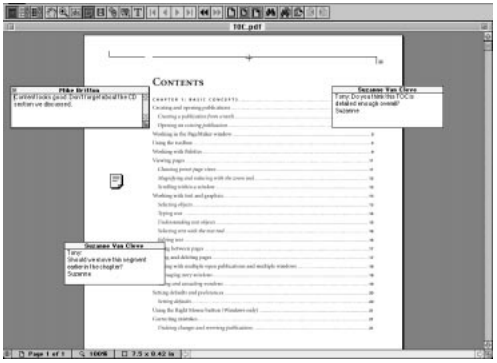
Annotations And Extra Links

As the name implies, Acrobat Exchange is designed for groups of users to share files on a network. The Annotations feature allows many individuals make comments on the PDF file and share the comments with the group. Because the notes are stored in a separate layer of the PDF, the original remains untouched for viewing and printing.

Annotations are represented by icons that resemble 3M's ubiquitous Post-it™ Notes, and they provide the same functionality. They convey a message that can be attached and easily detached from the original document. In Exchange, individual users can customize the appearance of their notes in such areas as color, label and fonts in the note. In a workgroup, the contribution of each individual is instantly recognizable.

tip

The contents of all the notes of one or many PDF documents can be summarized in a separate document.



Notes can be up to 4x6 inches in size and can contain up to 5,000 Roman characters.

From Here to Multimedia— The Create Link Tool

The Create Link screen gives the author the ability to design the appearance and action of links in PDF documents. With Acrobat 3, PDF caught up and maybe surpassed the action functions in documents available in HTML.

Each has its own place and time, of course, but the essential nature of HTML is linking, and the essential nature of PDF is presentation. PDF has made a lot more progress in HTML's field than HTML has made in the realm of PDF. No disrespect intended to the brilliant HTML designers or to the emerging capabilities of HTML that

are being rapidly expanded by many vendors. Nothing promotes rapid technical advance faster than cutthroat competition on the open market, and with Microsoft and Netscape setting the pace, things happen fast.

One of the major evolutionary changes to PDF in Acrobat 3 is the new family of actions available to the author. The new Acrobat 3 capabilities in many ways mirror developments introduced in HTML 2.0 and HTML 3.0, including forms and multimedia.

Things Readers Can Do Through PDF Links

Go To View	Execute Menu Item
Import From Data	Movie
Open File	Read Article
Reset Form	Show/Hide Field
Sound	Submit Form
World Wide Web Link	

Acrobat 3 introduced a new tool under the Tools-Create Link menu, called Sound, that allows the author to add an .WAV or .AIFF sound file to a PDF document. The procedure is very simple and menu-driven, and the Sound file becomes part of the PDF file. Any sound-capable computer will allow the user to hear the voices, music or other contents of the sound file. The author can specify the action that will play the sound, such as Mouse Click in a field, or even Mouse Enter or Mouse Exit the field.

We earlier estimated a single-spaced page of text to comprise 2,000 bytes in file size, and a scanned and Group IV compressed .tif image of the same page to comprise 50,000 bytes. At normal speaking speeds, a sound file of the same page being read aloud in AIFF or WAV format may easily exceed 1 MB (1,048,576 bytes) in size.

Another innovation of Acrobat 3 was the bundling of the Apple QuickTime viewer, which had already become popular on many fronts, including such best-selling CDs as *Myst* from Broderbund. Using the same mouse techniques described above for adding sound, an author can attach video files to PDF documents. However, unlike sound files, video is always stored separately from the PDF.

QuickTime and other video files are very large. For example, the Weezer "Happy Days" video, included on the Windows 95 CD under FunStuff, has a running time of about 4 minutes and consumes about 30 MB of file storage. This poses no problem for CD publishers, but every Web and network publisher has to take a long, hard look at transmission speed and load to move a file that large.

Sacrifices

For the average user, entering the data into the Document Information Fields is like shining your shoes. Yes, they look a lot better and you're glad you did it when you're out in public. But in private, it's a nagging chore. Professional publishers add this value without fail.

Considering the benefits to be enjoyed by future users, such shoe shining should be an absolute duty. Many documents created in word processing and other office automation packages may already contain this information, and it can be automatically entered by PDFWriter or Distiller. Even when the information resides outside the format, third-party products are available to automatically populate these fields from a separate file.

Recipients Need Acrobat Reader To Use PDF

It's hard to predict how the chore will be handled in the future, but for now the users must download new versions of the Reader software as it is introduced. To the happy beneficiaries of free software, such as users of Netscape Navigator, Acrobat Reader, et al, it is a simple chore to surf to the proper URL and download the latest stuff. To the broad market, all of these technical details are unacceptable impediments to enjoyment of the online world. Therefore, to accommodate this much larger market, we can expect future software to be self-downloading, at minimum inconvenience to the user.

The basis for this functionality is already in place with the Registry function of Windows 95. When a user contacts a Web server, the server will query the user's PC to determine if all of the necessary software components are present. If something is missing, the servers of the future will automatically update the user's PC with the necessary software.

tip

The issue of trust arises when our computers connect over the anonymous Internet and have software automatically downloaded to change our own machines, especially at home. Trust is not a technical issue, it's a social one.

The Dreaded 'Learning Curve'

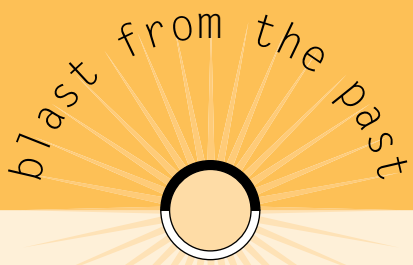
I often wondered why the best prospects for new technology always seemed to buy the stuff only after it had already been on the market for a few years. For example, major law firms were some of the last to switch to Windows. It was only after many, many years of learning that I realized they were hesitant precisely because they had already invested in earlier generations of office automation.

Not only did the stuff still work, they often had tens or hundreds of people who were skilled in the technology. The real cost of technological change has little to do with the cost of hardware and software. The cost of change is in training and learning, the cost in people time to transition to a new technology. Imagine if we had to totally relearn how to drive our cars on totally redesigned highways every five years. Delays would be expected.

The larger the network or number of users, the more difficult and expensive it becomes to move to a new way of doing things. A second level of chaos is introduced if the changes are partially phased in over a large network because different versions of files inevitably clash in all but the most meticulously designed transitions.

The great advantage today is that the leading software environments are designed to be cross-platform. Windows designs to be cross-platform by being the only platform, Mac finally designs to be cross-platform, UNIX is widely revered.

The two leading universal document formats, HTML and PDF, are being handled by millions of users on the Web. Both formats are designed to blend into the woodwork, and both run smoothly and familiarly on Windows and Mac PCs. Because these document types have been designed from the ground up to run on all popular platforms, it is likely that the learning curve will be drastically reduced. People will be familiar with the technology, and in society's eyes Instant Access will be more like programming your VCR than programming a computer. A little complicated, but anybody can do it.



Present And Future Readers

Anybody who has been using computers for more than 10 years has old files lying around. They might be Wordstar documents on 8-inch floppy disks that were written on CP/M computers, or they might be 5.25-inch floppy disks with Xywrite files written on an old DOS PC. All of these digital files have something in common with earlier generations of files that were stored on 9-track tape in IBM's old EBCIDIC language; what they have in common is that both the media and the file formats are virtually inaccessible now.

Like PostScript, PDF developed a rapidly building undertow of support. Its adoption as a standard comparable to PostScript is already a foregone conclusion. Modern technology in the form of telecommunications has given us the ability to rapidly move and preserve our files, virtually freeing us from the dead end of hardware changes. And PDF offers a stable, standard format that will generate the self-propagating support necessary for long life.

case study

What to Index for the Future

At Biological Sciences Information Services, they have been publishing both citations and abstract information on the life sciences since 1926. In their case, the Keywords assigned to certain articles may not even appear in the articles, but the Ph.D.s who will search this information will be pursuing certain concepts or trends, and the Keywords will bring certain articles to their attention. At Biosis, Ph.D.s actually do the Keyword coding!

The emphasis must be on serving future users. If lawyers are going to be handling this collection, use legal conventions. If biologists are going to be using it, organize the info the way the scientists will use it.

In other words, use only those data fields that are most helpful, but be sure to exploit these fields to the fullest extent.

See Process Map on page 100 for a step-by-step guide.

Decisions To Make

Let's not reinvent the wheel when it comes to designing digital documents because paper documents have a half-millennium of traditions and lessons learned. But we must not fail to give our future readers every possible advantage of the new document. It is at this nexus, where paper information becomes digital, that we have the historic and singular opportunity to offer our users a new way of reading, researching and learning.

Perhaps in the future, a society at home in a digital world will think more along the lines of Bucky Fuller than Aristotle. Bucky often said that he could not afford to remember anything he could look up in a reference book. He wanted to reserve as much of his considerable mental processing capability for creative and synergistic thinking, and he zealously strove to focus his mentality on the most productive processes.

Perhaps the just-in-time philosophy of total quality management will be applied to knowledge and learning in the future, as it is now applied to manufacturing where parts and material only show up when they are required by the production process.

In this scenario, people in the near future will come to trust and rely upon the availability of information, and they will free their minds to concentrate upon the creative tasks at which human thinking excels, and which no computer to date can accomplish. The computers do their jobs, we do ours.

In looking forward to such a world, it is the duty and burden of every publisher of digital information to offer up products that lend themselves to these new and as yet still experimental approaches to using data and knowledge.

Sometimes the rate of change outruns the sensibility of change. The examples are myriad. In the auto industry, when automatic transmissions became available, many

people became convinced that manual transmissions had been rendered obsolete by automatics. In the '60s and '70s, manual transmissions became a special option on American cars. Of course, manual transmissions are a lot more fun to drive and give you a lot more control of the car. When the imports arrived in the '70s and '80s, manual transmissions reassumed their rightful share of the market.

This allegory concludes with the observation that the forms and conventions of digital documents are in a state of vibrant flux right now, and sometimes valid traditional forms are ignored for trendy but unproductive presentations. Every digital publisher and digital librarian should see the new media not only in the light of what it can do that paper cannot. Every new media presentation should also be judged on what it doesn't offer that a book containing the same information does.

As an example of the rate of change of new media, consider the 3-D visualization software called the Virtual Reality Modeling Language (VRML). The prototype VR browser appeared in a sort of game called Labyrinth, which was produced by Mark Pesce and Tony Parisi. Labyrinth came out in February 1994, and by October of that year VRML 1.0 was recognized at the Second International Web Conference.

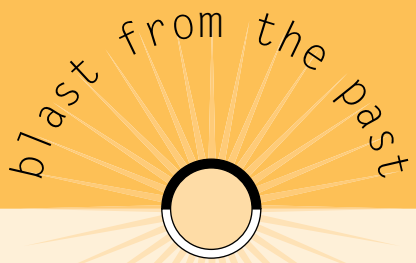
Silicon Graphics stepped up and took responsibility for VRML 1.0 as a standard, published programming environment and released its VRML development environment called WebSpace in April 1995. The recently introduced Cosmo from SGI takes VRML thinking and capability to new dimensions and new forms of object interactivity.

The lesson learned in the example of VRML is that it went from dream to reality in less than two years! Technology cycles are intensifying and shortening in quantum leaps. In this mystifyingly rapid evolutionary cycle of new Web technologies, we must try to put documents into the most stable formats. The documents themselves must be ocean-going canoes with outriggers with oars and sails that can ride through oncoming waves of technology changes.

How Much Is This Info Worth To Convert?

Books are sitting on shelves, chockful of meticulously organized and presented information. I'd like to think digitizing all of the books on library and university library shelves, all the rare books in the world, all of mankind's accumulated visions of truth…but I know that business and legal documents will be the first afforded this royal treatment.

There are several digital library projects currently active on the Web, and the trend is in place for future global libraries, as originally envisioned by Vannevar Bush and Bucky Fuller and Alan Kay and the rest. But this technology will first be embraced by



Because the idea of a universal document seems so natural to most people, the technical details are usually overlooked. Because of the way computers were presented to us as a society, we assume computers can do all jobs better than all previous means. This is a thought bubble that pops almost as soon as you begin to really try to use a computer on the Internet.

In the old days, there was only one standard: ASCII. The standard 128 characters of basic ASCII were almost universally shared by all the BUNCH. In the early days of big computing, there was IBM and the BUNCH, comprised of Burroughs, Univac, NCR, ControlData and Honeywell. The BUNCH shared the language of ASCII for files, while IBM used an expanded language called EBCIDIC for files, which has since receded into history. ASCII, that limited set of characters derived from typewriters and telegraph machines, has prospered. HTML is ASCII.

commercial publishers and businesses that have a real stake in the information and can see the value of instant access in their profit & loss statements.

Every potential user will look at a bookshelf full of information and ask himself the simple question: How much can I afford to spend to make all that knowledge available on my Web?

Sometimes gross measures help to put the new media in perspective. A brochure or white paper published on the Web is instantly accessible, and the end user can choose to incur the cost in toner to print the document for his own use. For the user who really needs the information, the cost of paper and toner at a couple of cents per page is certainly worth it. The primary advantage is that only qualified prospective clients get the marketing materials.

Not only are printing costs virtually eliminated, but the intellectual density of the piece can be increased because of the very focused audience that will read it.

tip

On the Web, there are no printing or postage costs.

A traditional paper advertising campaign involves writing, typesetting and printing, and then the cost of mailing the material to a hopefully interested audience of prospective clients.

What used to cost a dollar to mail to a targeted list, only to be recycled or thrown away, is now available to a new, pre-qualified, interested community of active users on the Web. High-quality documents can be printed by any Web user on his own equipment.

The Process

Any author or publisher can easily migrate to the new world of universal format publishing, and the most important consideration is carrying forward all possible benefits of the present processes. Most often the only step that needs to be changed is the output option. Everything upstream of output is secondary, unless the publisher can take advantage of one or more advances in digital publication to add navigation or index value to the published information.

In an ideal world, every author has filled in all of the document information fields in the source applications, and field data can be translated directly into the PDF Info fields or perhaps into HTML Meta tag fields. The primary directive is to capture as much information as possible from the original document.

Meta Tags: Definable fields in HTTP headers that can add traditional index fields in document collections. Organizations that have defined specific Meta tags include those that serve math, physics and computer science disciplines. These Meta tagged documents are the HTML equivalent of other indexing schemes, with virtually unlimited variations on Author, Title, Subject, Keyword, Accession Number, Citation and all of the other traditional fields.

Of course, to be useful, these fields should be openly published and accepted in the target user community. Carrying on the conventions of paper journals makes sense.

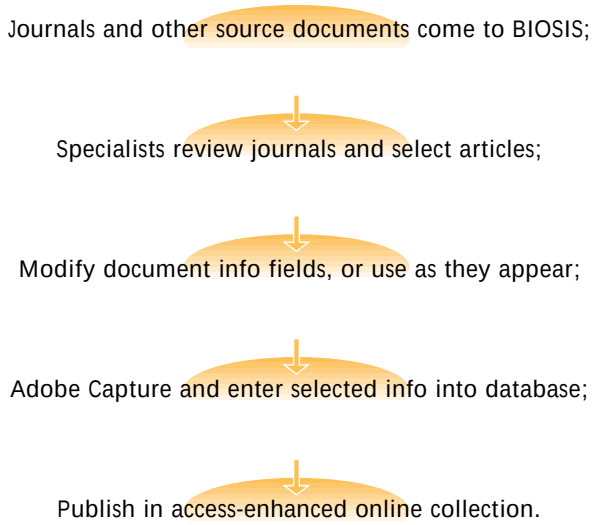
If the original document lacks this information, it is still important to enter as much data as possible before committing the PDF document to a widely published collection. Of course, the value of this added information vs. cost of acquisition should always be considered. The only hard and fast rule is what will best serve users in the future.

There are software products available that can apply a set of data fields to a set of documents, whether they be PDF or HTML.

For batch scanning purposes, a data entry sheet could be devised that would include the Author, Title, Subject and Keywords fields. The sheet would be either typed or handwritten, and either a person or an OCR program would capture the data.

This sheet would either be retyped by a data entry operator or converted by OCR into the Author, Title, Subject and Keywords fields. The result of these processes would be an ASCII file that could then be read into the fields in the Acrobat PDF files by a third-party product from Ambia, Inc. or some other developer.

process: secondary publishing



Summary

The key to building any new collection of documents is to accommodate the needs of our future users. What is the nature of this information? Most important, how will people be led to just what they are looking for?

How can you lead the users to information when the they don't know what they are looking for? The publisher must take the traditional lead and proceed as if he knows what the future user needs.

The work put into a digital document should not exceed its practical worth. If a set of documents will never be accessed by more than a single field of data, two fields of data are an indefensible luxury.

What does it cost to make all of the information instantly accessible?

All of these concerns can only be answered for each specific application. General recommendations are guidelines, common sense questions that might avoid silly mistakes.

footnotes

1. PDF for Production Printing, Adobe White Paper, March 1996.