document management:

information as a corporate asset

Some of the earliest pioneers of document-management systems have been law firms, whose life blood is documents. In fact, law firms were coincidentally among the earliest adopters of word processing, which led to the need for document management.

In the earliest word processors, where documents were stored on magnetic cards, it was necessary to manage those cards. Even as simple a document as a greeting letter would be standardized if it met the needs of a large number of people. If an individual modified that widely reused standard letter, chaos would ensue. In the earliest days of word processing, only the most senior, responsible secretaries (hey, this was the '70s!) would be allowed to record anything because accuracy was everything.

This manual process of controlling the important magnetic documents worked well, and a trained operator could use the cards to produce "fill in the blanks" documents that were personalized for each letter. Eventually, there were machines that held two cards, and a card full of names and addresses could be shuffled into a standard letter, the original list/merge function.

When early standalone word processors evolved into "shared systems" in the late '70s, many people began sharing common collections of electronic documents. Law firms, for example, shared collections of contracts, agreements, letters and so on. Immediately, the need arose to distinguish original from modified documents, and then to pick out various versions of documents.

As surely as the big law firms evolved to Novell networks and WordPerfect, they adopted increasingly sophisticated document-management systems to answer this key requirement. In a profession where words count for everything and literally define the product, it is imperative to maintain the integrity of the documents.

This chapter provides an overview of the key issues of ownership, permissions, version control and audit tracking.

Document management as a specific discipline is dissolving into the larger world of increasing acceptance of digital documents. As discussed in earlier chapters, the rudiments of document management are included in the Document Info fields of Acrobat, Word and most other modern applications. As we move rapidly toward universal document sharing, this crucial concern will be blended into the general milieu. But the critical function of providing for the integrity and unique identity of individual documents will never diminish in importance.

Document Management Principles

Common sense applied to document management quickly arrives at a few basic rules:

1. Documents must be secure; access must be firmly controlled.

2. Given access permission, user changes must be tracked in an audit trail.

3. Any document/version must be traceable anywhere on the network.

4. Supervisory functions are usually available across the collection.

In order for globally dispersed users to confidently use the digital collection, very tight controls must be in place. A source document must be inviolable. The envelope of document-management information is the best first step to security.

We discussed in earlier chapters the elements of document management. Early, forward-looking applications like Adobe Acrobat, Microsoft Word and many earlier programs all built Doc Info fields into their file structures.

Those early Author, Subject, Title, Keywords, Date Created, Date Modified and other Meta-info fields included in the file structure offer the foundation for sophisticated document management.

For more information on document management, check out these sites.

Open Text Web site:
http://www.opentext.com

Documentum Web site:
http://www.documentum.com

PCDocs Web site:
http://www.pcdocs.com

Saros, a Division of FileNet, Web site:
http://www.saros.com

Classic Doc Management Fields

Assuming that anyone reading this chapter desires to share digital documents among large communities of users, I confess I've taken a shorthand view of the issues involved. All of the following fields are dynamically tracked in a true document-management system. Audit trails are available for each document, monitoring each access and action that has occurred in the life of that document. In any document-access system that involves changes to the documents themselves, tracking the overall processes that have occurred on the document is critical.

In Chapter 4 we discussed the Doc Info fields and their function as rudimentary document-management tools. There are similarities and differences between the fields for managing documents and the system of permissions and audit trails for actually controlling the documents, as shown here:

Document Management Field	Acrobat Field
Name	Title
Category	Subject
Owner	Author
Description	Keywords
Created	Created
Last Modified	Modified
Size	File Size

Document-management systems control documents through a series of permissions and audit trails. All levels of access to a document can be controlled either by an individual user or by groups. For example, the supervisor usually has over-riding capability to manipulate documents, whereas individual users may be limited to certain functions such as See Document Citation, See Document Contents, Modify Contents and so on. In addition to user permissions, the versions of the documents are often stored separately, along with an audit trail of all accesses and operations.

Status

This field describes the current status of a document, like a book in a library, or a movie at the video store. The field determines if the document is reserved to a particular user and, if so, the period of time it will be unavailable.

Version

As documents are checked out, modified and returned in new forms, various versions are tracked, with the number of versions kept determined by the administrator or individual users. In areas of collaborative use of pooled documents, it is crucial to identify individual versions to avoid misspent time and energy on out-of-date documents.

MIME Type

MIME type is crucial for Web applications because it allows a user to take advantage of his own installed software for optimum viewing of all retrieved files. MIME types are recognized by the browser by their file extensions. For example, if the user's browser is properly configured to recognize the MIME type, it will open Word to view a .DOC file, Acrobat to view a .PDF file, etc. Otherwise, unregistered MIME types will be simply stored on disk. While it is desirable to receive certain files directly to disk rather than display them, such as compressed files, it is best to configure the browser to handle all expected MIME types. In Intranet business applications this will be less of a problem because the universe of possible file types will be predictable, unlike the Web itself.
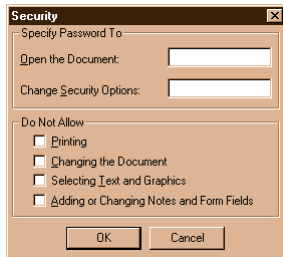
Doubling security, each PDF document can be password-protected
with multi-level permissions.

One of the biggest bonuses any Webmaster or Intranet Sysop can give to their users is a simple list of pre-configured MIME types. That way, the browser will be fully configured for at least the anticipated business documents. No users will complain that they couldn't read a spreadsheet or word processor file or some other commonly accepted business document format. They may complain that they can't read the latest 3-D VR space or some other fun stuff, but that's on them.

Another enhancement offered in the new Intranet business suites of software is an included viewer. The viewers can display and usually print "views" of 40 or more common document formats, or MIME types. While these can not be edited, the information is available to users without the source application.

For example, a browser configured to recognize the MIME types .DOC and .RTF as Word files will launch Microsoft Word on the user's computer. If Word or another compatible word processor is not installed on that computer, a viewer could be used to display the file.

Assembling Virtual Docs On The Fly

The supreme advantage of high-tech document-management systems is the ability to assemble up-to-the-minute reports from all of the best sources of original information.

Rather than create a large, aggregated document where many files are copied into one megalithic mass, you can create an elegant set of references within your short summary. This function is comparable to the way a word processor can assemble many separate chapters and sections into one long document for printing.

The hyperlinks of HTML take this feature one step further, where each document is composed of a number of links, and each link refers to a dynamically changing document, which may in turn be updated separately or draw its contents from remote sources.

You can make your points in confident shorthand and refer all doubting Toms to the URLs from which you arrived at your conclusions.

This ability to create dynamic documents through included references will never diminish in value. This function of dynamically created documents through organized links is fundamental to the value of the Web and Intranets.

Workflow: Assembly Line Style

Workflow is the ultimate brute-force document-imaging application where the main benefit is moving images of documents over a network rather than moving paper documents between numbers of people and offices. Workflow, as originally realized in a couple of billion dollars' worth of business applications, concentrates on the organized movement of document images among the various business functions, treating documents as so much raw material in a fixed business process.

In this process, there are various nodes, corresponding to programming branch points. For example, if a flood of documents comes pouring in through a high-speed scanner, the various batches of images can be handily distributed to any number of human processors to enter all the data and resolve all the questions.

In theory, workflow starts in the mailroom, where all documents are scanned and sorted into batches, after which all documents flow through an evenly distributed cycle of human decision procedures.

If the documents themselves can be organized into batches and carry their own unique index and status information, the complicated task of document handling may conceivably be automated. In fact, many very complex processes involving document handling have already been efficiently automated through workflow. But only well-thought-out and carefully analyzed high-volume, repetitive operations return a value on this investment.

Document Workflow

Mailroom

Content Modifiers — Analyst / Marketing / Finance

Publishers — Editor / Designer

Final Web-Ready Document

The key is the transaction. If a set of documents is closely coupled to a business transaction, where real money changes hands, it makes sense to look into workflow. If it's just a once-in-a-while thing, where you are fishing for benefit/payoff, workflow is just a red herring and should be ignored.
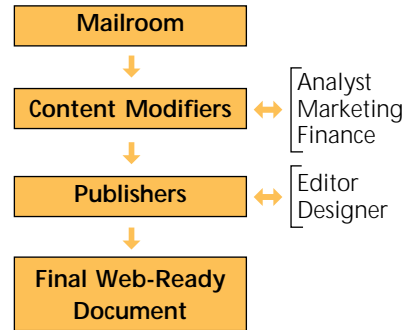
Complex Doc Management Via RDBMS

To understand the desirability of managing documents in a database, you only have to ask a Webmaster what his job is like. Documents constantly change, with new documents replacing old documents, or usually just succeeding them in an ever-growing aggregate. It falls to the Webmaster to track all the links between the new and old and maintain order. In computer science, this task was cracked long ago in Relational Database Management Systems (RDBMS).

In very large collections of documents, you want to retrieve information rather than search for information. By their very size and volume, it becomes impractical to follow hypertext links to desired information. The oft-stated goal of "three clicks and you're in" is virtually impossible to achieve once the number of documents grows to

the thousands and beyond. So large collections are untenable not only from the Web-master's system-management point of view, but also from the end user's convenience and efficiency point of view.

The solution to both of these bookend requirements is to incorporate RDBMS tech-niques and technology. These are the engines that drive the big business applications and routinely handle tens of thousands of users and millions of transactions. Think of the automated teller machines that sit out in the cold and rain and sleet and snow and reliably debit your account no matter which bank it may be in and within sec-onds hand you a short pile of cash. That complex transaction requires far-flung com-munications and transactional processes between very large financial databases, and it takes less than a minute. These transactions are performed, logged and verified by huge online systems. They are usually running under proprietary megalithic in-house programs or on commercial databases such as IBM DB2, Oracle, Informix, Sybase, SAP and others.

With increasingly sophisticated security software on the Internet, the big database vendors are bringing their power to Internet applications. For example, Informix offers a series of Datablades to allow new functionality that can be built within the robust architecture of its RDBMS. One announced Datablade allows the Excalibur information-retrieval capabilities to be applied to data and information residing within Informix databases.

This convergence represents several breakthroughs in efficiency. This is the economy of supporting one combined system for information retrieval rather than running par-allel systems with all of their recurring costs and burdens. For example, in the past all of the business applications would be run on the "Big Iron" and RDBMS software, while "soft functions" such as search, document management, workflow and collabo-ration would run on secondary, or parallel, systems. Combining these functions on the primary systems currently supported by IS budgets makes a very persuasive cost and procedural argument.

Another advantage is that free-form information collections can enjoy the benefits of transactional speed and verification, including automatic instantaneous updates, which are common in RDBMS applications.

The major benefit of this convergence is that all users may utilize all of the processing, communications and stability of the larger system, enjoying the synergy of conver-gent purposes. It makes common sense to have all the documents and all the data available to the users in one collection and through one interface.

Summary

Web-accessible document-management systems allow large and widely dispersed groups to refer to, modify, copy and otherwise employ remotely shared collections of documents. This ability to dynamically share the creative and editing processes is referred to as collaboration software on LANs, WANs, and the Internet and Intranets.

Workflow is a concept from the document-imaging industry, where complex, repetitive document-handling processes were automated through network-based transfers of electronic documents. The prototypical workflow applications were employed in insurance claims processing, manufacturing work orders and other document-based applications. As an evolutionary adjunct of document management, workflow is the programmed movement of documents through a specific process.

The entire process can be automated in a workflow system, and any participant in the process can be notified when his participation is required, and he can also view the current status of the process. In repetitive processes, workflow offers enormous improvements in efficiency through automation.

The ideal document-management application is the repetitive query that generates a constantly updated report. A number of "views" of extremely complex operations can be constantly updated and instantly available. All key reports from R&D, Marketing, Accounting, Personnel and so on can be generated from a single dynamic document-management process that always reports up-to-the-minute views of the entire process.