Web Archival

Enabling the portability of documents across your business.

November 2002

By now practically every business uses the Web in some fashion to support business operations. Such universal network entities as email and Web partner portals have dramatically changed business forever.

Utilization of Web tools is still developing and spreading to new areas of business. This evolution is driven by a series of trends that stand to further incorporate these new tools into everyday business. Two primary market trends that influence a company's decisions to purchase and make use of Web-based systems are overall technology adoption and business process reengineering associated with distributed data access.

Interestingly enough, though the Internet has existed for more than 30 years and the Web for 14, businesses' adoption with respect to how they manage information is still in its infancy; and trial-and-error remains the most common way of applying these technologies to business operations.

Business Information Explosion—The Problem

Businesses have literally millions of documents that support everyday processes. Most of the time, these documents end up being stored in a room full of file cabinets, on employees' computers, or, even worse, lost forever. In each of these cases, a great quantity of time is consumed trying to file away documents correctly and then to locate them again when they are needed.

IMR™
INFORMATION MANAGEMENT RESEARCH
www.imrgold.com

The macro business trend of enterprises expanding their internal systems to allow customers and partners access to business data, regardless of locale, is one factor compounding document filing and access problems. Businesses are processing information and supporting business processes in new ways, and this most often includes use of the Web as a means to support distributed access. However, most companies have not truly assessed the way in which these Web-based applications are being used or how they are integrated. The net result is that this new era of business use of Web technologies dramatically increases the amount of information that a single individual must manage on a daily basis.

What about the Web? How does it fit into the document management equation? What if an international business has employees whose workplaces are spread across many different locations? How can they easily access the documents they need? These are a few of the issues that companies are grappling with in the "age of e-business." Most companies deal with questions of productivity, streamlining processes, and reducing costs. Web-based solutions offer answers to these questions.

New tools are needed that are easy to install and use, that leverage the power of the Web, and that allow a typical office worker to quickly solve problems on their own without specialized training or complicated and expensive software engineering.

In this research paper, we will provide an overview of:

- The impact of the Web from both a technology and a business perspective
- The resulting influence on document systems and management
- Web archive and retrieval and its advantages
- Web archive solutions
- Likely future scenarios

Web Impact on Technology and Business

The 1990s saw the birth of business adoption of Web technology tools: This decade will see the integration and maturation of the Web as a mission-critical support system for business operations. There are two primary "impact zones" of the Web within business: The first is the overall growth of Web technologies presence within IT infrastructure, and the second is the impact of those technologies and their influence on business processes. Both of these macro trends continue to expand, shaping other areas of business — including how documents are captured, stored, retrieved, and used.

There are two primary "impact zones" of the Web within business: The first is the overall growth of Web technologies presence within IT infrastructure, and the second is the impact of those technologies and their influence on business processes.

Technology Impact

Three drivers underlie the macro trend of Internet technology adoption. Each has its own unique impact, and together they represent a major change in the way businesses process information.

1. **Web protocols are taking center stage in the evolution of corporate computing.** Protocols include TCP/IP as the primary transport, HTTP as the main protocol for sending data, and XML/SOAP/Web Services as a way to provide data inter change. These protocols affect remote access to enterprise systems, application integration, and distributed application architectures.

   There are many reasons:

   Deployment flexibility — using browsers as the key interface and server-side application management, Web-based applications make deploying and managing software easier. They also enable better returns on investment as they provide a foundation for communication between new and existing legacy systems and the data that reside in both.

   "Easier" to integrate — providing open protocols means that new technology can access old systems without extensive custom coding. Additionally, Web-based components can document-enable other applications, giving broader context to the content. Additionally, Web applications are built on a standards-based, component-driven architecture that provides a snap-in framework for customizing, connecting, and integrating systems.

   Modular implementation — Web applications typically are designed out of "building blocks" of code that can be scaled to meet demand.

   Data consolidation — Web presentation standards allow the seamless representation of disparate data. Additionally, standard components and a similar user interface mean that integrators and users alike save the time they would have spent learning entirely new methods.

   Data distribution — providing a single point of access for data searching and retrieval means that distribution can be managed more easily. Paper distribution costs averaging six cents per page can be cut by 60 percent to 80 percent according to Gartner Group.

Separation of document and applications — In traditional systems using electronic documents, the content is inseparable from the way that content is presented as well as how the content should be processed. This often leads to an inflexible system that requires multiple processing of files and rewriting of applications when new functionality is required.

Emerging standards now allow documents to describe themselves, creating reusable and, most important, portable logic that is distinct from the application managing them.

2. **Browsers are becoming the norm for application interfaces.**
Most computers come with a browser, which provides ease of deployment. The simplicity of adding plug-ins to expand functionality effects design modularity. And because of uniformity in the basic interface, most people know how to use a browser at a basic level.

3. **Scope of the term document is expanding.**
Documents now include XML/HTML pages, enterprise application reports, MS Office–type documents, scanned paper, images, and emails.

Business Process Impact

Business processes also have been profoundly impacted by the Web. Three primary elements are driving this change:

The goal is to simplify administration by managing data in one place where it can be collectively searched and accessed.

1. **Companies are moving toward "corporate portal(s)" integration.**
Companies are continuing to aggregate all corporate information into common storage and delivery mechanisms based around Web portals. The goal is to simplify administration by managing data in one place where it can be collectively searched and accessed. The results are better user adoption, quicker access to the entire corporate knowledge base, and lower total costs for support.

2. **Remote or extended working environments are emerging.**
In order to be more cost effective, flexible, and growth oriented, companies are expanding beyond typical corporate borders — implementing telecommuting programs, occupying smaller branch offices, and opening global operations.

3. **Use of Web presentment for self-service applications is growing.**
To create differentiation or to lower costs, companies are placing entire business processes online. Examples include online stores, online customer service, electronic billing, and partner trading hubs.

Influence of the Web on Document Systems

As one might imagine, no technology solution, whether hardware or software, has been left untouched by the number of Web technology and business drivers. Among them are systems that manage the diverse types of business information stored in documents, both physical and electronic.

Web technologies are driving four major shifts in the methods companies use to manage their documents.

Transformation of Documents for Web Applications

Businesses are transforming the documents themselves into more accessible data types, rather than modifying their systems.

Businesses are transforming the documents themselves into more accessible data types, rather than modifying their systems.

Five common types of transformation currently exist:

- Conversion to PDF — formats a document in the standard PDF structure for easy portability and protects document integrity.
- Rasterization to GIF or JPEG — takes a snapshot of the content and presents it in the form of an image. The content typically is not alterable, nor can additional data processes be applied.
- Reconstitution of contents to HTML or XML – captures print stream or OCR and converts into HTML/XML. Actual content is now self-describing, but documents are still static in nature.
- Conversion to HTML or XML — captures print stream along with formatting data and translates that information into a formatted HTML/XML document. Images are large tables. This process is similar to reconstituting except that transformed documents retain the original's fidelity.
- Translation to highly formatted HTML or XML — uses Cascading Style Sheets; this type is the same as conversion, but it utilizes strict HTML/XML formatting to accurately render the document in the browser (Source: Bill McCalpin, MHE).

The purpose of this transformation is to enable applications to place archived data into use in a Web environment. Examples are electronic bill presentment, online report distribution, and customer self-services.

Increased standardization on PDF

There are many reasons why PDF continues to gain new ground as the de facto document standard for Web publishing. PDF provides not only the ability to capture original document fidelity, but it also enables extensibility to transform these documents to match other types of media using XML and tags. The combination of XML and PDF will accelerate the convergence of the print stream to the Web data stream. A number of tools to create or transform PDF files for Web-based applications are currently available.

Distribute-then-print

We're all familiar with the idea of taking reports or other documents, sending them to the printer, and then distributing them. The Web is taking over from the old print-then-distribute routine and turning it around. Now users can easily route documents to any number of recipients. Depending on the type of document it is, it may never need to be printed. This not only reduces costs associated with printing, but it increases productivity, as electronic distribution is instantaneous.

Records integration

When companies consolidate, one of the first realizations is that now twice as many disparate data silos exist. The Web makes consolidation easier by providing a common interface in the form of a Web browser through advances in data aggregation and common search applications.

Advantages of Web Archive and Retrieval Systems

When businesses decide to distribute data in a Web environment, they may not realize the wide range of solutions available to them. Solutions range from sophisticated enterprise application portals to basic static Web sites. Meeting somewhere in the middle in terms of capability is the Web archive. Essentially, Web archive solutions provide a streamlined process of capturing a wide variety of business data in the form of documents and publishing them to the Web. This type of solution is precisely what many companies want and need — unfortunately, they often end up with an overly complex system or one that has too little capability — either way, the value of the system is limited. Web archive systems, on the other hand, are simple yet powerful, they are cost effective, and they protect existing IT systems.

Essentially, Web archive solutions provide a streamlined process of capturing a wide variety of business data in the form of documents and publishing them to the Web.

Simplicity

Simplicity is one of the primary attractions of Web archive solutions. Most Web archival and retrieval systems are relatively easy to implement, require minimal, if any, integration, and are quickly up and running, all due to the overall simplicity and specialized design of the system. This is in direct contrast to Web content management or enterprise portals, which often require business analysis, systems integration, or dedicated system administration personnel.

Businesses typically will not get personalization, dynamic content, or sophisticated business process management functions in a Web archive solution. These capabilities, however, are usually superfluous to the actual business need: publishing business data online. They will receive a dramatically lower total cost of ownership and satisfaction from a solution that delivers on its promise.

Cost Effectiveness

Companies often mistakenly believe that a complex Web publishing system or enterprise portal is necessary. These high-cost products exceed the needs of most companies.

Imagine the types of data most businesses would want to share with partners or customers over the Web. Invoices, purchase orders, statements, and business documents are the most likely candidates for publishing to the Web. These types of data are usually stored in file cabinets, employees' desktops, or mainframe systems. Web publishing systems and enterprise portals, though capable of handling some of these data, are suited to more Web-centric types such as HTML and do not easily allow the capture and aggregation of these disparate types of data.

Web archive systems provide a very effective solution, enabling businesses to simplify the capture of these common types of business data and virtually any other type of document, whether through scanning, facsimile, print stream output, email, or simple desktop drag and drop. Additionally, these systems can easily present the data in a Web environment, including the ability to perform searches on the entire archive. In most cases, a Web archive can be set up to require little or no integration with existing systems and to be in use within days.

Protection of Existing IT Investments

After the heady days of the late 1990s, many companies are suffering from IT indigestion. Hardware and software was purchased, sometimes indiscriminately, in hopes of a competitive advantage that often failed to materialize. Now companies are spending smarter and insisting on solutions that deliver what they promise.

Following in this new age of common sense, CIOs are most interested in solutions that enable them to protect and extend their existing IT infrastructure investments. Solutions that require replacing or significantly altering existing hardware and software infrastructure are not welcomed.

This factor rules out many solutions that require adding a new layer between systems, thus necessitating custom integration, or render existing systems useless.

Web archive and retrieval does not require companies to invest in new infrastructure — rather, these solutions utilize what's already there to accommodate new uses.

Web archive and retrieval does not require companies to invest in new infrastructure — rather, these solutions utilize what's already there to accommodate new uses. They accomplish this task by capturing output, regardless of form, preparing it, storing it, and then making it accessible via the Web. No costly integration and no lost investment.

Uses of Web Archive and Retrieval

With specific regard to document archival, Web solutions can meet the requirements of companies that need to:

- Provide distributed access to partners
- Allow extended enterprises with remote offices to centralize archived data
- Provide self-service applications using archived customer records.

Distributed Access

Integrated partner access to documents was first tackled by Electronic Data Interchange (EDI) and value-added networks services in the late 1960s. Using these technologies, companies were able to share invoices and purchase orders with little effort. However, the costs associated with integration and maintenance were, and still are, prohibitively expensive for many companies.

The result is limited adoption by any but Fortune 500 companies. Businesses looking for the opportunity to tap into efficiencies brought about by partner integration can now use Web standards to deploy Web-based applications. These applications not only are much simpler to deploy than EDI-integrated applications, but the switch to Web-access means that private leased lines and networks and their associated costs are eliminated entirely.

Additionally, common browser-based Web interfaces practically eliminate the hassles of ensuring that partners have the right applications, configured in the proper manner and that users will easily adopt and use the system.

Extended Enterprise Support

Gartner Group named "place independent workplaces" as one of the top ten forces that will change the workforce. Consequently, supporting an enterprise that has widely distributed operations will become an integral part of the business process.

Businesses cannot afford to have branch locations cut off from central systems. The result would be a tangle of business processes using outdated data. Harnessing simple access to Web applications ensures that employees in Manhattan, Kansas, are accessing the same data as those in Manhattan, New York, without the high traditional client/server costs of deploying applications to several locations and the associated high learning curves of those same applications. Businesses can continue to centrally manage data, but the actual data use can occur at the farthest edges of the enterprise.

Customer Self-Service

Increasingly, businesses are opening up their data to allow access by customers so they can realize the significant savings to be found within self-service applications...

Increasingly, businesses are opening up their data to allow access by customers so they can realize the significant savings to be found within self-service applications, take advantage of reduced support costs by not having to provide around-the-clock CSR availability, and then take the excess capacity and offer more personalized "high-touch" services aimed at enhancing customer profitability.

The Web obviously has sped up this process by making it easier for customers to access this information. Whereas customers dislike talking to customer service representatives for small issues, they quickly find value in accessing the answers themselves. Also, customers can get the answers they want, when they want them, because computers don't sleep.

What Is the Future of Web Archival?

The advances and seemingly endless announcements of new standards and technologies beg the question, "Which developments should my company invest in?" In response to that question, three trends appear to have a lot of momentum.

1. **Increased use of Web services**
   Web services represent a profound change in the way people and businesses use and experience software. Web services, simply put, are Web-based protocols that enable applications of any type to access data stored in, or functions of, other applications. Yet, companies should not expect them to become the one answer to all business application questions. Though Web Services will make deploying applications much easier, we will probably see specific capabilities tied to vendors. This means that Microsoft services will work better on Microsoft, Sun on Sun, and so forth. We will see industry consortiums like AIIM (The Association for Image and Information Management) lead the effort to create standard document services that companies can deploy within or without their enterprise. This eventually will lead to true document portability across the business landscape.

   With document portability, companies will be able to reuse documents in a variety of applications, enabling an even wider access and use of corporate information without requiring extensive integration efforts or rewriting of applications. This translates into greater productivity with less effort.

Businesses need to actively investigate available solutions in order to leverage all of this new information into their everyday business processes, or they will suffer a competitive disadvantage from broken processes, insufficient access to vital data, and the ultimate inability to react to market changes.

2. Return of the living archive—reference data

Though archives are made of seemingly static content, the data that reside in those records can be used for many purposes beyond their initial intent. With the advent of XML-based, self-describing record structures, archive data can be searched for, identified, and extracted. Archive data does not have to be defined by the original document. Data now can be separated from presentation. The recent moves by Microsoft, which announced its XDocs strategy, and Adobe's moves to transform the static nature of PDF documents to more dynamic containers validate this direction.

This development means that the resulting data extractions, from one or many records, can be formed to create entirely new views of the archive data. An example would be a time series of customer purchase orders that are stored within the archive. Although the documents are static in nature, data within these documents can be actively extracted and processed to reveal new insights into purchasing behavior.

3. Web bureaus

Companies that currently provide outsourced data extraction and document imaging services will actually be able to sell their processes and output as "integratable" — on-demand Web-services that can be accessed by other network-enabled applications. In the future, a customer may click a button to view a record at their bank Web site and may actually be serviced in real time by that bank's service provider. Other companies will be able to build applications and subscribe to a bureau's services through a basic Web connection.

The rationale for Web bureaus is that services can easily be located, deployed, and integrated with minimal hassle. This enhancement leads to greater flexibility of use and upgrade, lower cost installation, reduction of personnel on the payroll, and the ability to add and drop services as business demands change.

Summary

The rapid transformation of typically "technology-free" business processes to incorporate a more complex mix of Web and mainframe-based applications means businesses are generating ever greater quantities of information that must be managed effectively. Businesses need to actively investigate available solutions in order to leverage all of this new information into their everyday business processes, or they will suffer a competitive disadvantage from broken processes, insufficient access to vital data, and the ultimate inability to react to market changes.

While the prevalence of Web content seems to be taking center stage, it is only one fraction of the total business information generated. Scores of other types of information trapped in the form of invoices, purchase orders, email communication, and planning documents all need equal attention in order to provide a competitive advantage through lower costs and the ability to more quickly respond to market pressures.

Web-based document archive solutions are the logical starting choice. They provide flexible information-capture capabilities, robust long-term management of data, and easy-to-deploy Web search-and-retrieve tools, all within a low-cost, rapid-deployment solution. Once implemented, these solutions can support a wide variety of new and existing business processes, driving up efficiency up while driving down costs. Additionally, while the data Web archive solutions store is mostly static, the solutions themselves are continuing to evolve to meet tomorrow's business needs. Choosing a Web archive solution today means that businesses can realize dramatic results now and in the future.