
If nothing else, production teams could use better ways of finding the content they need — the age-old search and retrieval problem.

Contending with content

Spurred by new technologies, Corporate America is rethinking how best to create, store, manage, and “repurpose” its documents

In the last issue of this letter, we explored some emerging technologies that promise to help get new products to market with maximum impact: as fast as possible and with just the right features. After decades of using automation to reengineer and squeeze cost from back-office, front-office, and production processes, growing numbers of manufacturers are turning to rapid turnaround in new products as a key way to boost profits. And to do so, as we saw, they’re flocking to software and services designed to help manage virtually every aspect of the so-called product life cycle.

This week, we’ll look at an array of emerging tools and techniques that purport to help with an equally important and more or less parallel life cycle that most companies, government agencies, and other organizations are now struggling with, the one that deals with what’s commonly, if somewhat awkwardly, called content. Whether it’s a PC disk drive, windshield wiper, or Shrek 2 lunch box, virtually every new product gets accompanied through its life - in a virtual sense, anyway - by an ever-expanding heap of content: Word documents, PowerPoint slides, lab reports, data sheets, operating manuals, JPEGs for Web pages and .BMP files for box tops. Accurate and timely content is critical to developing, designing, making, marketing, using, and maintaining virtually all products and services. And so, organizations are reevaluating the often sloppy and ad hoc processes by which they generate, store, index, share, retrieve, rework, and reuse all that stuff in all its myriad forms.

Islands of opportunity

This rethinking of processes immediately raises sticky technical issues that we’ve encountered elsewhere in business computing. There’s the babel of fragmented and incompatible information systems. Since almost by definition none of the “unstructured” data types that make up the category of content fit particularly well into the pigeon-holes of a relational database, all sorts of specialized and isolated systems have been set up to handle that data. Department by department, workflow by workflow, desktop by desktop, these are quite disparate systems whose numerous pools and streams of content are now begging for integration.

If nothing else, production teams could use better ways of finding the content they need - the age-old search and retrieval problem. It’s a problem that, in miniature, is plaguing lone users of desktop PCs. Steve Jobs last week described the next version of Apple Computer’s OS X operating system as providing a facility for searching the content of all kinds of files on a Macintosh, no matter which applications they’re associated with. Microsoft is thought to be working on something similar for a future version of Windows, code-named Longhorn.

Go with the flow

There are other problems demanding attention from entrepreneurs. A great number of the production workflows now in place were originally engineered to produce print-based products almost exclusively. As a result, the systems and worksteps required to deliver electronic formats were often afterthoughts and therefore not as efficient and flexible as they might be. Web content was managed separately from other, more traditional forms, for instance. But

CONTENT MANAGEMENT

There's the push by many content providers to segment markets into ever-finer categories, each requiring its own variation on a shared set of content.

now, as enterprises see that future content delivery will be primarily electronic in nature, they need to rethink and rebuild their workflows and supporting information systems in quite radical ways. What's more, as more business partners are enlisted to design, produce, and market new products, these workflows must stretch across corporate boundaries yet remain robust and manageable - another major technical challenge.

Shattering experience

As our friends at The Seybold Report point out, many factors are fueling change in content production and management within corporations and in the publishing and electronic media industries, too. And this flux is creating much opportunity, we believe, for early-stage technology companies. There is, for instance, the ever-increasing variety of end-user devices and storage formats to take account of: umpteen brands of cellphone, a handful of popular personal digital assistants (PDAs), Windows, Mac, and Linux-based PCs, CD-ROMs, DVD-ROMs, and the Web, to name a few. There's the push by many content providers to segment markets into ever-finer categories, each requiring its own variation on a shared set of content. Content is increasingly being produced for global use, too, created centrally and then translated into any number of target languages. What's more, corporations are exploring all sort of new business models for the sales and delivery of content. They're syndicating it and aggregating it from multiple sources, they're bundling it with software-based services, and they're producing everything from college textbooks to retail catalogs in individually personalized, on-demand print runs. Last but not least, techniques for producing and managing content are being buoyed, too, by on-going improvements in the cost and performance of computing hardware, data storage, and networks.

The story repeats?

Clearly, content has enjoyed a great deal of

previous attention from entrepreneurs. Technologies such as desktop publishing, Web content management, and document management were each a mini-revolution that not only greatly improved the scope and capacity of what users could accomplish but yielded a fair number of startup success stories. Among them are Adobe Systems, Frame Technology, Interleaf, Vignette, Interwoven, Documentum, and Verity. Today's content-focused startups, it appears, are arriving somewhat late to the party, their intended users already relying on lots of systems and related processes. What's more, the content management landscape has seen more than its share of mergers and acquisitions in recent years as established players such as Filenet, IBM, Microsoft, and OpenText extend their product lines and jockey for position. Even EMC, a leading provider of high-end disk storage, has entered the market in a big way, acquiring Documentum last year.

Change of view

On the other hand, we believe that the difficulties that users are encountering as they try to adapt their current hodgepodge of systems to new markets means that they're quite open to innovation. They're recognizing and treating content as a corporate asset that's more valuable than ever and as a critical success factor in many new product and service strategies. Much as previous waves of startups have attacked the problem of integrating disparate enterprise applications and databases, there are now several early-stage firms aiming to integrate content. They include **Context Media**, **Vamosa**, **Venetica**, and **XyVision Enterprise Solutions** (XyEnterprise), which have come up with schemes that pave over technical differences between various content systems and repositories and, in essence, give applications developers a single, virtualized interface for using them. The aim is to free these independent systems from the specific business processes they were originally intended to support and enable the construction of entirely new and

In terms of the visual presentation of data, XML goes well beyond HTML, which is one reason it's creating such excitement among publishers and others working with content.

improved processes. There is a move, in short, away from the document-centric view of content of the past to a more database-centric view, with any combination of content elements available for instant retrieval, modification, and reuse in new, unanticipated ways.

Not surprisingly, a key technology here is XML, or extended markup language, which provides a means of tagging not only documents as a whole but each of their many components. Like HTML, a close cousin, XML is derived from a markup scheme called SGML (standard generalized markup language) that was created about 30 years ago to help publishers make their digitized content easier to adapt to new, even unanticipated applications and devices. SGML never caught on as widely as it might have, largely due to its complexity, but it directly inspired the creation of both HTML and XML. Each is designed to extract structure from content and therefore make the latter more malleable and reusable. XML tags can provide contextual information about a document or any element of data therein, indicating the topic of a report and the name of its author, locating its individual headlines, subheads, and paragraphs and, like HTML, suggesting type fonts for the display of each element.

On display

In terms of the visual presentation of data, however, XML goes well beyond HTML, which is one reason it's creating such excitement among publishers and others working with content. The client whose job is to display a set of XML-tagged information can be given tremendous freedom in how to interpret and act on those tags. This client can be equipped with a potentially infinite range of rules to guide its interpretation of XML tags. RenderX, a company specializing in such technology, has created a demo that illustrates this quite well. Its software can be set up to translate standard shorthand notation - e2-e4 - describing the successive moves of a chess game into a series of chess-board diagrams that illustrate a game move by move. Likewise, RenderX's code helps utility companies to produce monthly bills whose layouts automatically accommodate any amount or type of information that may be needed. There's no fear of billing data spilling onto an

adjacent graphic and thereby becoming illegible. The same technology is helping Goldman Sachs, Charles Schwab, and Dun & Bradstreet print statements, loan documents, and other documents that merge information pulled from databases with graphics and other elements.

Joining hands

Equally important, XML-based metadata can help with the searching of multiple repositories of content. And XML can be used to help guide documents or individual assets such as digital images through complex, computer-managed workflows. Finally, tagging the individual elements of a document - its sections, sub-sections, individual paragraphs, and graphics - can make those elements much easier to organize, retrieve, and recombine into entirely new documents or to present in radically different formats.

Judging by what the startups are doing, the most popular approach to so-called enterprise content integration (ECI) appears to be a form of federation. By creating a layer of middleware and applying heavy doses of XML, it's possible to weave together a set of incompatible content-oriented systems and create a platform on which to build new applications. These apps may run the gamut, from organizing the movement of document images from desk to desk within an insurance company to streamlining the production of continually revised technical manuals.

Controlling personality

Context Media has concentrated its energies on helping corporations manage their marketing content, such as brand-related images. When launched in 1999, the company set out to develop software for managing so-called digital assets: photos, graphics, video clips, chunks of text, and other items that a multimedia publishing outfit like Simon & Schuster might keep on hand for use in creating textbooks, workbooks, teachers' materials, CD-ROMS, and TV programs. Such companies have long faced the challenge of not only keeping track of where such assets reside in their many computers but also the specific rights they may have to use each asset - how much each use would cost. Some may be licensed for use only in the U.S., for instance, while others may come with global permission.

As product cycles continue to shorten, there will be a premium on streamlining the process of selecting and pulling together the right marketing materials and crafting new content in print and electronic forms.

After addressing that problem, Context Media has beefed up its software's ability to manage access to multiple repositories of assets in a unified, centrally controlled way. Its program creates a master index of all assets and enables that index to be searched directly from within most popular desktop apps. When someone wishes to retrieve an item, it's delivered by the system where it's stored, but Context Media's software keeps track of each such transaction so that the use and modification of each item can be audited in the future. It also blocks the viewing of items that a particular person or workgroup may not have permission to see. Sony has used this setup to help its business units share content: To promote a new plasma TV screen, for instance, Sony's electronics arm was recently able to find just the right images of Spiderman controlled by a sister motion pictures unit.

Transborder data flow

That kind of cross-promotion will become more popular in the future, we imagine, as will other forms of cross-selling and joint marketing within and between corporations. As product cycles continue to shorten, there will be a premium on streamlining the process of selecting and pulling together the right marketing materials and crafting new content in print and electronic forms. Even the text and graphics used on packaging can be crucially important content when launching new products. Increasingly, companies are wooing customers online and off with highly personalized product brochures, but their on-the-fly creation may be practical only with sophisticated content-management systems. And many marketing companies now work on content with scores of outside partners, making central control of brand-related assets critical. One of Context Media's customers, the firm tells us, is a large food company that works with some 250 ad agencies, each one permitted to see only a selected portion of the firm's marketing and advertising assets.

Pulling together

Venetica is out to integrate a different set of content and related systems, namely those that corporations rely on to execute their core business processes. Insurance under-

writers, for instance, might depend on access to digitized photos and images of hand-written reports submitted by adjusters in the field. Problems arise, though, when a company wishes to pull selected document images and other pieces of content from multiple systems and use or present them in a new way. This situation can arise when two companies are merged, each running its own systems to manage content and oversee automated workflows. Or, to comply with new regulations such as Sarbanes-Oxley, a company may strive to tie together several repositories to gain a unified, corporate-wide view of important content. Or, a self-service Web site may have to extract content from multiple systems.

Venetica's VeniceBridge software is designed to handle these tasks by virtualizing any number of content repositories and the applications that use them. As a piece of transactional middleware, VeniceBridge provides two-way connections between apps, enabling them to exchange pieces of content with each other in real-time and enabling the creation of new workflows that can span multiple apps. Venetica provides preprogrammed, full-function interfaces into more than 20 popular enterprise apps and a toolkit with which users can build their own interfaces. VeniceBridge also translates between different content formats.

Platform play

The company tells us its software appeals particularly to companies sophisticated enough to develop new systems and workflows on top of legacy apps. In many cases, it can cost these companies too much to yank out and rebuild from scratch their home-brewed imaging and other systems. By providing virtual, consolidated views of those systems' content and the tasks and events they manage, VeniceBridge hides the apps' peculiarities and extends their useful lives. Venetica has worked hard to provide the kinds of interfaces and software components that developers can use to build new systems. One major integration opportunity Venetica looks forward to tackling: enabling customer relationship management (CRM) systems, which are now based on relational database systems, to serve up unstructured content like customer correspondence, purchase orders, and faxes.

Venetica has earned a major vote of confidence from IBM, which is bundling Venice-Bridge with its strong-selling DB2 Information Integration suite.

Manual dexterity

XyEnterprise describes its software as helping to produce and manage complex technical content such as user manuals, data sheets, and legal tomes. Customers such as the U.S. Navy, U.S. Postal Service, plumbing supplier Koehler, and Lexis/Nexis use the products to maintain large repositories of XML-tagged content elements and rapidly pull just the right pieces together to create customized content whenever it's needed.

Ship-shape

Thus, when one of its ships comes back to port after months at sea, the Navy can quickly update all of that vessel's documentation in one go. Instead of simply loading a complete new copy of the documentation, however, the Navy can use XyEnterprise's code to revise just those pages and paragraphs that require changing. Many suppliers are involved - several providing electronics, say, while others supply the ship's engine - but XyEnterprise's software, hosted by the Navy for remote use by its suppliers, provides a central store of all their documentation. As long as they tag their contributions with the Navy's XML vocabulary, the software can summon up any combination of elements as quickly as needed and produce hard-copy, CD-ROMs, or any other required format in a flash. Likewise, Lexis/Nexis can produce textbooks that are customized to the needs of individual law schools, pulling selected passages from a XyEnterprise-based repository. The post office uses XyEnterprise to produce unique documentation for each configuration of machines that it assembles for individual sorting facilities.

Happy trails

XyEnterprise is not alone in addressing the technical documentation market. **Astoria Software**, formerly known as LightSpeed, has set out to help the aerospace, nuclear power plant, technical publishing, automotive, and other industries reengineer how they produce complex manuals and other documentation. The firm has developed an

object-oriented database setup that's designed to store XML-tagged content and enable customers to manage updates and quickly pull together any selected pieces. Astoria's target customer is grappling with five criteria, the company tells us: high costs for authoring and producing content; a frequent revision cycle; high risk and cost from any errors that creep into the content; high production costs, perhaps due to having to supply content in many different formats; and the need to comply with regulations, which may call for audit trails describing exactly who altered each piece of content and when, for instance.

Up, up and away

Astoria's technology, originally spun out of Xerox, had languished for several years but now, the firm has raised new venture financing, hired a new CEO, and set its sights on several vertical industries. It signed a significant deal with Boeing last year which could pay off in a big way as the airplane maker influences its suppliers to change how they produce technical content. As part of the deal, Astoria acquired a technical team from within Boeing. The focus of the new activity is to streamline Boeing's process for producing documentation for its 12,000 aircraft. Each airplane requires its own version of documentation, based on the specific features and equipment it has installed. What's more, airlines generally produce their own documentation, which must be merged with Boeing's. Astoria is also working with a Boeing unit called Jeppison, which supplies something called electronic flight bags for individual flights. These are collections of flight-specific information that pilots must have on board and refer to before and during each flight.

Two other startups that attack the content integration problem in their own way are Vignette and Interwoven (now a unit of Broadvision). So far, these two have not made much progress, evidently because their software was designed to manage solely Web content: all the graphics and text used on dynamic e-commerce Web sites, for instance. What they're missing, we understand, is an ability to harness multiple repositories of content and manage workflows that draw from any or all of them.

One of the most intriguing new technologies we see in the XML content market right now is a database developed by a startup called Mark Logic.

Mark Logic has an opportunity, we believe, to prove that there is a strong alternative to the federated database approach that other content integration startups are pursuing.

Post-structuralism

One of the most intriguing new technologies we see in the XML content market right now is a database developed by a startup called **Mark Logic**. Founded as Cerisent, Mark Logic recently emerged from stealth to show a product that stores content in a powerful new way. Taking advantage of XML and a related open query language called X-query, Mark Logic is out to create what it calls a content hub around which it hopes to foster an ecosystem of supporting and related content-oriented apps. This ecosystem, the company points out, has been distinctly missing in the document management market that Documentum pioneered and the document search arena where Verity is so strong. Because their products relied on a relational database and proprietary query languages, would-be providers of add-on apps have been put off by the slim profit margins they'd be forced to manage with.

Easy does it

Mark Logic describes its database in comparison to a standard relational product. Just as the Oracle database, for instance, enables joins across tables as a way to create new tables and answer queries, Mark Logic's product can do joins across documents and create new documents. It enables parametric searching of content, too - find all documents authored by Mark Twain, for example - and searching by keywords, a la Google. X-query is much like SQL in that it requires no knowledge by the requester about the data's internal organization. Because of all this, the Mark Logic software can make things much easier for developers trying to create new content-oriented apps.

For instance, with help from software supplied by startups such as **ClearForest**, **Olive Software**, and **Itemfield**, a collection of documents, even entire books, could be automatically parsed into their many constituent elements - sections, subsections, paragraphs, headlines, and so forth - and then all tagged in XML. Then, those tagged elements could be fed into Mark Logic's software and stored there under its control. Now, developers can write query-based apps that would find and pull together appropriate elements on the fly and create new documents or web pages as needed.

Creation story

As a demo, Mark Logic shows its software's ability to find the most relevant and useful information a doctor might need from within a large set of highly-technical and quite hefty medical books. Ask the system for information about "tennis elbow," for instance, and it uses a mix of database and search techniques to present an impressively well organized collection of paragraphs and illustrations about that topic. Further queries can quickly refine the information. Were the same mass of text searched based only on keywords, the list of "hits" would be much less likely to highlight the most important sections of text. Altogether missing would be the kind of internal links that Google takes advantage of to identify the most important and useful hits. Mark Logic, on the other hand, takes advantage of any clues provided by the internal structure of documents, all encoded in XML.

Opportunity knocks

Mark Logic sees opportunity for its software in the publishing industry, particularly where complex, valuable, and highly structured legal, scientific, or medical content is involved. The defense industry is another target, and the company is actively working to recruit independent software vendors (ISVs) that have expertise in special content areas. So far, Mark Logic has a handful of customers but we expect that its marketing efforts, including a new developer site offering trial downloads of its software, should garner it a good amount of attention. There doesn't appear to be any other product on the market that's capable of doing what Mark Logic's can do - though an XML database developed by **X-Hive** in Rotterdam may come close - and we wouldn't be surprised.

Formalities

Mark Logic has an opportunity, we believe, to prove that there is a strong alternative to the federated database approach that other content integration startups are pursuing. By centralizing the storage and indexing of content, versus centralizing only the content's metadata, the number of moving parts, so to speak, can be greatly reduced. Issues of synchronization and integration, which add lots of complexity, are pretty much eliminated. Of course, Mark Logic must overcome potential customers' reluctance to bet on its technology,

which is so far relatively unproven and alone in the market. As yet, there doesn't appear to be a Ted Codd - the man who's considered to be the father of relational database technology - to authenticate this centralized approach. Mark Logic assures us, though, that its code is rooted firmly in rigorous mathematics, just not the same algebra that underpins the relational model.

Long bet

It's still a long way off, but Microsoft seems to be pursuing a vision similar to Mark Logic's. Its Longhorn operating system, officially slated to start shipping next year at the earliest, is expected to provide a comprehensive database facility that would replace the several independent and somewhat redundant database products that Microsoft currently supports, including Access, Exchange, and SQL Server. Details are sketchy at this point, but it would appear that Microsoft is set on building a fair amount of new XML functionality into its SQL Server product and making it useful for storing not only relational data but also mail, Excel spreadsheets, Word documents, and all those other documents known collectively as unstructured content. Already, the company has started laying the groundwork for this by adding significant XML facilities to Word and its other desktop apps. Mark Logic assures us it's not particularly worried about Microsoft because that firm's products and strategy will almost certainly be Microsoft-centric. Mark Logic reckons it can prevail with its markedly more open and brand-agnostic product.

No matter what happens in the database arena, content will call for further technologization. One of the fertile areas for new development is coming up with ways to weave content into other enterprise apps. Here, Web services will be of great help, we expect. A startup called **MetaCarta**, for instance, has come up with a way to automatically link geographic spots on digitized maps to points within documents where those places are mentioned directly or indirectly. This can be helpful not only to intelligence analysts working with maps, satellite photos, and field reports from agents, but to oil companies trying to manage exploration operations in far-flung locations. The software

relies on a specialized natural-language processing algorithm that MetaCarta has developed, which can identify geographic references in text. Not only can it recognize explicit addresses, such as 29 East First St., it can interpret phrases such as "26.4 miles north of the house where Joe lives." The result may be a map showing precisely located hyperlinks that point to specific documents or even passages within those documents. The firm tells us its program can analyze 4 million documents per day, tagging their geographic references by latitude and longitude. The firm sees lots of opportunity ahead as more consumer-oriented applications come on-line, such as those designed to help drivers navigate by car.

Another problem is simply helping users find what they need in the growing number of electronic manuals they have on hand. An outfit called **Softlib** has come up with a scheme that creates personalized portals, or virtual bookshelves, where selected IT product manuals can be searched by topic. This is in response to the many manuals, usually supplied on CD-ROMs, that IT shops have to cope with, especially when crisis hits and information and advice must be located as quickly as possible. Softlib's search mechanism uses only keywords now but in the future it may add concept-based searching, too.

Discontent with current content strategies and technologies should fuel continued innovation for many years to come, we believe. The prospects for this arena look good, even if most startups will likely get acquired by larger players. That scenario may simply be a fact of life for all enterprise applications firms. What's special about the content market is that the information it deals with is open to many more kinds of interpretation, analysis, manipulation, and processing than simple transactional records. Even without entering the field of rich-media content, such as audio and video, there are still lots of problems to solve and applications left to conceive of and build.