MarkLogic™

"In order to better serve our customers and users, MarkLogic Server will be a core of all our major electronic products going forward."

— David Marques, Chief Technology Officer
Elsevier

Unlocking the Value of Content at Elsevier

Integrating content to enable the rapid development and delivery of new information products.

ELSEVIER

Industry Overview

The last thirty years have seen a major shift in the production, storage and retrieval of content. Electronic, rather than printed, content now powers a new class of products and services, with the Internet key to delivery. Driving critical activities within companies, it also provides many of their content-based services to customers. Yet this exploding volume of content, typically comprising 85% of an enterprise's information, is fundamentally different from the structured data that relational database technology was created to deal with. A cornerstone in most enterprises, relational databases function by structuring data and applications into tables. They cannot readily accommodate the unstructured content residing in vast reservoirs of Word files, lab reports, data sheets, operating manuals, messages, HTML documents, PDFs, PowerPoint slides, emails, etc. So organizations wishing to repurpose and realize more value from digital content are stymied, since the applications needed to deliver added value to customers are difficult and expensive to create.

These issues have had even more impact in the publishing industry, where content is a company's greatest asset. And for Elsevier, a leading publisher and information provider for medical, academic and health-related organizations, they were particularly challenging. Elsevier supports, and continues to enlarge, a digital content repository unsurpassed in its market. Yet despite Elsevier's significant investments in search technology, their users found it increasingly time-consuming to extract the information they needed from this mountain of data. Elsevier was unable to quickly create applications that would make extracting content faster and easier...until Mark Logic demonstrated a product that could rapidly ingest enormous volumes of content, and then execute complex, fine-grained queries against it with lightning speed.

Elsevier's Challenge

Reconciling the need for relevancy with the necessity for volume.

One of the world's leading publishers, Elsevier prides itself on supplying customers with the information they need to conduct research, perform experiments, aid patients, and achieve mission-critical objectives. To this end, Elsevier invested heavily in digitizing its content, amassing vast repositories of medical and scientific information, and making it available via a range of online database-driven solutions. However, as Chief Technology Officer David Marques points out, users often have little time to locate the data most relevant to their work. "If a doctor is at the point of care or a scientist is working in the lab on an experiment, they don't have time to go searching through 10 or 20 possible sources." Yet as more content amassed, its sheer volume meant customers were spending more time refining searches to winnow out the content most relevant to their needs. Elsevier's greatest asset was growing more difficult to deliver with the level of granularity required by users. And this, explains Marques, is precisely the kind of value- added service Elsevier wished to supply. "We wanted to help customers solve the problems they face in their particular setting...By enabling our customers to extract only the pieces of content that matter to them at that moment, and to flexibly combine them, Elsevier can provide maximum value per use of content."

About Elsevier

A world-leading publisher of scientific, technical, and medical information products and services, Elsevier is the science and medical publishing division of Reed Elsevier Group plc. With a staff of 7,000 employees in 74 locations worldwide, Elsevier publishes more than 20,000 products and services – and over 1,800 journals and 2,200 new books every year. Offering a wide range of journals, books, electronic products, services and databases, Elsevier's main customers include librarians, authors, editors, scientists, researchers, and medical practitioners.

To achieve its objective for increased customer satisfaction, Elsevier set two goals: quickly transform the content rigidly held in its many separate databases into a liquid asset easily tapped by users in any way they desired. And in so doing, establish a common platform for developing future products.

But facing Elsevier were four formidable challenges:

1. Lack of central repository. Each body of content existed in a separate database – either in a relational database format or a proprietary one – with several applications on each database.

2. Huge range of file formats. Normalizing content was extremely time-consuming. For one application project alone, there were 35 different document formats involved.

3. High cost. New functionality was time-consuming and expensive to build. The complex logic needed to deconstruct a document and analyze relationships between documents had to be built application-by-application. Moreover, from a performance perspective, forcing this logic into an application was inefficient, compared to leveraging a specialized content server that can efficiently retrieve large amounts of information.

4. Massive amounts of content. The final content repository was estimated to exceed 5 terabytes in size. Included: More than five million full-text journal articles across 1,800 journals; over 60 million citations and abstracts (separate from the articles); 20,000 in-print books; 9,000 out-of-print books; and thousands of informational pamphlets.

Defining requirements

Preparing the way.

In an increasingly aggressive industry, Elsevier required shorter delivery cycles for its competitive offerings. So as a pioneer in the digital marketplace, they defined the parameters for products that extracted content from authors fast and put it online even faster. Equally significant, Elsevier recognized that to give users exactly the information they wanted, any new solutions must have the power to dynamically assemble relevant information from across multiple sources. Recognizing the potential of tagged search elements, Elsevier started in the year 2000 to redesign products along Web services architecture.

Beginning with Standard Generalized Markup Language (SGML), Elsevier moved forward, keeping pace with the evolution of descriptive signature technologies and ultimately investing in the benefits of XML (Extensible Markup Language). Of course, such advances helped deliver greater content relevancy to users. But enabling the highest degree of granularity meant the structural relationships of tagged content had to be leveraged in a way that allowed relevant information deep within documents to be parsed and reassembled into new content.

Accordingly, the absence of a centralized content repository had to be remedied, since an intelligent terminus would be required for all searches – enabling the deconstruction and synthesis of documents into context- specific results. This eliminated the deployment of traditional relational database systems, whose concepts and data models were conceived in an era of short, highly-structured records of data, and not the unpredictable and time-varying structure found in content.

Instead, the new solution would need to:

- Exploit the wide variety of unstructured content, rather than be constrained by it.
- Eliminate format and content-prejudiced conditions for standardization of information.
- Function without a single, standard, pre-defined schema, and indeed in the presence of many different and changing schemas
- Achieve performance without sacrificing relevance. Both the user interface and the returned results had to operate quickly and efficiently, and to deliver the right pieces of information at the right time.

Roadblocks to rapid product development

When developing new products, Elsevier had to confront:

- **Difficulty in leveraging and synthesizing information from documents held among a wide range of different databases and the applications on each.**
- **Lack of a "content common denominator" for normalizing information within applications or databases.**
- **Risk of poor ROI due to long and expensive implementation of new functionality and applications.**
- **A mountain of disparate content with no existing unifying solution.**

Putting Mark Logic to the test

Transforming a mountain of documents into a single, searchable contentbase.

By the year 2004, Elsevier had reengineered their products along the lines of web service architectures, creating an XML repository offering new efficiencies to their IT staff and higher functionality for users. But the apron strings of relational database technology still tied the company down to long, expensive product development cycles and less than optimal performance. To get reasonable content performance from their database management systems they still needed to pre-define schemas and access paths: time-consuming tasks that ultimately limit content ingestion and the power of resultant searches. And after intensifying their hunt for new ways to shorten time to market and add greater value to their content they found what looked like a perfect way to leverage their significant investment in XML: MarkLogic Server.

"Our promise was simple. Hand us any amount of data, as is, from your archives. We'll hand you back an entirely new application based on that content. Elsevier's team handed us an entire product line of 20 medical textbooks, each a thousand pages or more...and in about a week we came back to them with a fully functioning application."

— Paul Pedersen Co-founder and Chief Technologist
Mark Logic

"The system lets you reach across large content sets, extract exactly the information that you need, and then present it as a new document that was created automatically."

— Paul Pedersen Co-founder and Chief Technologist
Mark Logic

"We offered to show Elsevier how the MarkLogic Server could leverage their investment in XML to deliver on Elsevier's vision," recalls Mark Logic Co-founder and Chief Technologist Paul Pedersen. "Our promise was simple. Hand us any amount of data, as is, from your archives. We'll hand you back an entirely new application based on that content." And as Pedersen further described to Elsevier, "The system lets you reach across large content sets, extract exactly the information that you need, and then present it as a new document that was created automatically."

Intrigued by the prospect of being able to simply pour existing archives and content into MarkLogic Server and receive a fully functional application, Elsevier agreed to the test. And to see just how short a timeline Mark Logic could deliver a competitive product in, they made it a demanding one. According to Pedersen, "...Their team handed us an entire product line of 20 medical textbooks, each a thousand pages long or more. They didn't even provide the DTDs," he recalls. "They just said 'Go.' So we did, and in about a week we came back to them with a fully functional application."

Moreover, according to David Marques, the application Mark Logic delivered in just a few days was more flexible than anything Elsevier had online at the time. This accomplishment was all the more remarkable considering that the 0.5 terabytes of content loaded into MarkLogic Server was comprised of over 35 different formats – a flexibility matched only by the level of granularity provided by searches using the resulting application. Impressed, Elsevier engaged Mark Logic and is using MarkLogic Server to consolidate all of its archives, rapidly build new applications, and create value-added services from its repository. As Marques affirms, "MarkLogic Server will be a core of all our major electronic products going forward, since it allows us to even better serve our customers and users."

Benefits of the MarkLogic Server

Putting content in its place, *fast.*

From Mark Logic, Elsevier found an immediate solution to all the key challenges facing publishers who need to hasten the deployment of new, more competitive online products. They are now consolidating all of their content archives, rapidly bringing new applications to market and enhancing existing applications with value-added functionality that makes every last byte of content available to users in the most relevant way.

Combining the power of database-style queries against content, with the speed and scalability of search engines, MarkLogic Server repurposes content on-the-fly, combining information into new content for users seeking answers to different questions involving the same subject matter – literally creating new content from old and adding value in the process. Massively scalable in both storage and performance, it can manage millions of documents and terabytes of content – with no degradation in executing queries and updates.

A boon to publishers and their customers, Mark Logic revolutionizes search technology by enabling:

• **Consolidation of content archives.** MarkLogic integrates content from many sources into a single repository, then creates new content by summarizing information across various content categories.

• **High-performance XQuery implementation.** A complete XQuery implementation delivers high performance against multi-terabyte datasets, thanks to MarkLogic's search-engine-style indexing mechanisms.

• **Rapid application development, no fixed schemas.** MarkLogic does not require schemas or document type definitions (DTDs). MarkLogic loads content, as is, and allows you to instantly start building applications that leverage it.

• **Element-level granularity.** Using XQuery, MarkLogic Server fulfills searches by reaching deep inside documents to identify, analyze, combine, and extract pieces of content exactly relevant to the task of the user, precisely within the context in which they're working.

• **Extreme flexibility.** MarkLogic accepts content "as is" from many sources, eliminating the lengthy process of preparing content. Rather than having to plan ahead for every possible use of the content, publishers can rely upon the flexibility of the technology to evolve applications over time.

A single solution to many bottlenecks

With Mark Logic, Elsevier has achieved:

- Consolidation of all content archives into one centralized repository.
- A high performance platform for multi-terabyte contentbases.
- Higher efficiency through centralized storage of content and indexing.
- Element-level search granularity for users.
- Preparation-free content loading.
- Speedy application development thanks to the power of XQuery and the elimination of extensive content preparation.
- Just-in-time delivery of information that is precisely tailored to users' needs, within the context they're working in and in the form they need.

Benefits to Elsevier's customers

The results.

The power of a database, the speed and flexibility of word and phrase search functionality, the ability to deliver it all in a fraction of the time previously necessary… With MarkLogic, Elsevier not only speeds the delivery of new, more competitive products, but enables users to get exactly the data they need to complete their tasks 5 to 9 times faster than before.

For example, Elsevier no longer needs to normalize content to transform it into their repository. Now they can build directly on the inherent variability of different types of content – slashing time to availability by two-thirds. And, says David Marques, the ultimate benefit to users is fantastic granularity: "…When a user has a question, we want to reduce the number of search results from 10 possible documents down to two precise sections or paragraphs so we deliver just the right bit of content the user needs."

But for the Elsevier team, the gratification of providing this kind of added value to users goes beyond feelings of pride in their technical accomplishment. As Marques explains, by enabling researchers and medical professionals to find fast answers to urgent questions, they also help improve treatments and outcomes for patients: "Medical reference books are invaluable resources for making a diagnosis, but laboriously searching and cross-referencing a number of different books is an inefficient way to do this. The products we build with MarkLogic allow physicians to quickly pull out only the relevant passages from across a range of different books, in order to reach an informed diagnosis."

And the future for Elsevier and its customers holds even more promise. MarkLogic has dramatically accelerated the deployment of products and services, while greatly reducing the costs of content loading and design – translating into even faster research cycles and clinical diagnoses, thanks to a new generation of solutions for helping professionals find exactly the information they need, when they need it most.

"The majority of our time on a project is consumed by deciding exactly how the content will be used and preparing it for the database. With Mark Logic, we've now cut that time in half."

— David Marques CTO Elsevier

About Mark Logic

Mark Logic Corporation is a leading provider of Information Access and Delivery solutions used by government, publishers, agencies, and other large enterprises to accelerate the creation of content applications. The company's flagship product, MarkLogic Server, is an XML content platform and includes a unique set of capabilities to store, aggregate, enrich, search, navigate, and dynamically deliver content. Designed for high performance and scalability, MarkLogic Server can deliver millisecond response times against multi-terabyte contentbases. Mark Logic holds two patents on its innovative technology for managing XML content using the W3C standard XQuery language. Mark Logic is privately held and backed by Sequoia Capital and Lehman Brothers. For more information, or to download a free trial copy of MarkLogic Server, go to www.marklogic.com or visit the Mark Logic CEO blog at marklogic.blogspot.com.

Mark Logic Corporation
www.marklogic.com
Headquarters
999 Skyway Road, Suite 200
San Carlos, CA 94070
+ 1 650 655 2300
New York
+1 646 378 2104
United Kingdom
+44 (0) 207 643 1712

Version 2
April 2008