

Case Study



“Content Interaction Server will be a core of all our major electronic products going forward.”

David Marques
CTO, Elsevier

Elsevier

Elsevier is a leading, multiple-media publisher of scientific, technical and health information products and services, and is the scientific and medical division of Reed Elsevier Group plc. Elsevier publishes more than 20,000 products and services, including journals, books, electronic products, services, databases and portals. Their primary customers are authors, editors, scientists, researchers, faculty, students, professionals, medical practitioners and librarians.

Project Goal

In the publishing industry, content is more than a by-product of doing business; it is *the* product. Content is a publisher’s greatest asset, and its value is on the rise as documents like reference books, journal articles, and research reports are being made available online by subscription. Search and classification technology companies have enabled many people to search and browse these large archives of content. Increasingly, though, customers who purchase online content want more than simple search and retrieval. They want to go directly to a precise piece of content, and analyze it within the context of other pieces of relevant content.

This is the kind of value-added service that Elsevier wanted to offer its customers. David Marques, Elsevier’s Chief Technology Officer, explains that the company’s mission is to provide user-centered and activity-centered content: “Elsevier wants to help customers solve the problems they face in their particular setting,” he stated. “Content becomes much more valuable in the pursuit of these solutions, when it is optimized for a specific use and framed within the context of the user’s goal. By enabling our customers to extract only the

pieces of content that matter to them at that moment, and flexibly combine them, Elsevier can provide maximum value per use of content.”

Their goal was to develop these capabilities and get them to the market quickly and affordably, while establishing a common platform for future product development.

Content Challenge

Elsevier needed a technology that could overcome four primary challenges:

First, there was no centralized repository. Each body of content existed in a separate database, either in a leading, relational database format or a proprietary format. There were only a few applications on each database.

Second, normalizing content was extremely time-consuming. For one application project alone, there were 35 different document formats involved.

Third, new functionality was expensive to build. The complex logic needed to take apart 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, versus building it into a database that can efficiently retrieve large amounts of information.

Fourth, Elsevier had massive amounts of content. They estimated that the final content repository would exceed several terabytes in size. This includes more than five million full-text journal articles across a variety of journals; more than 60 million citations and

abstracts (separate from the articles); thousands of books; as well as a few thousand informational pamphlets.

Why Mark Logic

Elsevier selected Mark Logic's comprehensive, content database platform because it allows them to consolidate all of their content archives, rapidly build a variety of new online applications, and enhance existing content applications with new, value-added functionality.

High-Performance XQuery Implementation

Mark Logic Content Interaction Server incorporates a full implementation of XQuery, and is the only one that delivers high performance against multi-terabyte datasets, without a fixed schema.

Rapid Application Development

By building a set of rich, value-added services into the repository, the application development task becomes easier and faster. Rather than having to build on a basic set of search and retrieval primitives, a "smart" repository which can deliver summaries, highlighting, link analysis, and other key services, allows developers to support activities in line with Elsevier's strategic objectives.

Element Level Granularity

Content Interaction Server reaches inside content to extract sub-document level elements, e.g. chapters, sections, paragraphs, sentences. This means Elsevier can create new content by summarizing information across different bodies of content, grouping information across content categories, and building in other types of analysis.

Extreme Flexibility

Content Interaction Server accepts content "as is" from many sources, enabling Elsevier to virtually eliminate the lengthy process of preparing content. Rather than having to plan ahead for every possible use of the content, Elsevier can rely upon the flexibility of the technology to evolve applications over time.

Applications

Using Mark Logic's content database, Elsevier built a unified, company-wide repository to serve as a platform for a variety of content

applications. Rich, robust applications leveraging this common foundation can be brought to market quickly and cheaply, helping Elsevier to better serve its customers.

They completed the pilot version of the repository in six months and anticipate launching the first application on it by the end of 2004. To date, they have about ½ terabyte of data loaded into the repository. They are also preparing to migrate a second, major online service to Content Interaction Server.

Results Achieved and Benefits

Content Interaction Server will enable Elsevier to accelerate the design, prototyping and development of new content services and dramatically simplify loading content into the database – a critical bottleneck in getting applications to market quickly.

In addition, Elsevier customers can now do more than search for and browse whole documents. Applications built on Content Interaction Server let them extract targeted, precise pieces of content, and view them within their larger contexts.

"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," Marques said. "The products we build with Mark Logic Content Interaction Server allow physicians to quickly pull out only the relevant passages from across a range of different books, in order to reach an informed diagnosis."

"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

Mark Logic Corporation
2000 Alameda de las Pulgas
Suite 100
San Mateo, CA 94403
650 655 2300 Phone
650 655 2310 Fax
www.marklogic.com