

**AS SEEN IN...**

FEBRUARY 2006

# BOOK

---

## BUSINESS

## Search and Sell

Oxford University Press implements a new system to launch new products quickly, reducing 3 weeks of work to a single search command.



**Alex Humphreys, director of online engineering at Oxford University Press, U.S. division.**

### BY JOHN GARTNER

After more than 500 years of publishing experience, Oxford University Press (OUP) has mastered the techniques required to put words efficiently on a page. But, like most publishers, it has had to learn new techniques and systems to bring online its vast resources of scholarly research and reference publications.

“Our first online forays were big projects like the *Oxford English Dictionary* and *Oxford Reference Online*. These sites were large enough to deserve their own

platforms,” says Alex Humphreys, the director of online engineering at OUP’s U.S. division. “We built each new product from scratch, using unique resources and systems,” he says, adding that it took approximately 10 to 12 months to complete each destination.

These online resources are subscription services available to schools and libraries that generate a stream of income for OUP. But “after all the large watermelons had been picked,” says Humphreys, OUP was faced with a problem: Its new products could not generate enough revenue to sustain its own complicated infrastructure.

So OUP decided to develop a standard technology for its future, smaller projects so that the company could lower costs and reduce the development cycle, according to Humphreys. OUP decided to get its platform underway with the African American Studies Center (AASC), an online resource combining materials from its reference books and primary source material, along with specially commissioned materials. Five large A-Z reference encyclopedias will form AASC’s core content: *Encyclopedia Africana*, *Black Women in America*, *African American National Biography*, the *Encyclopedia of African American History*, and the *Concise Oxford Companion to African American Literature*.

In addition, AASC and the online resources to follow will feature a diverse collection of content types, including chapter-based monographs, primary sources, timelines, dictionary entries, tables, charts, maps and images.

The company selected MarkLogic Server as the basis of its publishing platform because the software “provided the greatest flexibility” in describing, organizing and searching content elements, according to Humphreys.

He explains that OUP uses XML (eXtensible Markup Language) to semantically tag its content. Publishing systems categorize each element class (such as a biographical profile or timeline) using an XML document type definition (DTD) that provides a structured syntax for each class. Each of the different content types mentioned above uses a different DTD to describe it.

Publishing systems that rely on flat, relational databases require normalizing every schema that was used to describe elements so that they can be queried, according to Mark Logic’s director of sales Bill Veiga. He says publishers’ attempts to replace all of their existing formats with a single standard results in “huge projects that never get finished.”

Oxford’s new publishing system allows for multiple XML DTDs and enables querying against the combined assets regardless of DTD. This frees publishers from requiring divisions to standardize on a single classification system for all elements.

For example, in its biographical profiles, a literary publishing division of the company includes authors’ major works, university degrees, and years of birth and death, while the reference division includes the dates of birth and death, spousal information, and middle and maiden names. Different divisions and content have varying business and semantic needs, so requiring all divisions to standardize on a single format would be untenable. Embracing multiple DTDs enables all of the information to be saved and staff to choose how the information is presented.

## **SORTING CONTENT AT THE CLICK OF A BUTTON**

Other publishers using MarkLogic Server include John Wiley & Sons and Elsevier, according to Veiga. The software enables staff at these and other companies to discover new methods of sorting and linking content by using XQuery, a search technology (a querying language for structured content, such as XML) that is expected to become a World Wide Web Consortium standard in early 2006, as it provides far greater flexibility than previous options like SQL or verity querying language, explains Humphreys.

Using XQuery, which is also supported by database vendors including IBM, Microsoft and Oracle, OUP has new capabilities for breaking content into its component parts and easily reassembling it, according to Humphreys. For example, editors could create a "Harlem Renaissance" collection by searching its materials for references to works created by African Americans in New York City in the 1920s and 1930s.

Humphreys says the software reduced a search function that previously took three weeks to develop to a single XQuery command. He says his team can easily customize the search results displayed on the Web sites by manipulating the software's text-querying algorithm. "It allows for much more iterative development of the search engine and of every aspect of the products," he says.

During the process, Humphreys learned that before starting to build interactive content from print resources, it's a good idea to dedicate additional time to content analysis and mapping the content against the needs of the product and market. "Even if you think you know your content, spending additional time with the content up front will be invaluable," he says.

## **JUST TAG AND GO**

Veiga says publishers such as OUP can add XML tags to categorize their content to form relationships between elements, enabling publishers to derive new revenue from "microproducts" that target specific audiences. According to Veiga, the key to helping OUP unlock the hidden

potential of its data was to separate the content, the business logic and the presentation into three distinct components.

Just by focusing exclusively on what was inside the content, the staff was free to explore it in new ways and developed additional business uses, Veiga says. In addition, with the new technology platform, OUP's online resources will not be tied to only one output, such as HTML. Instead, with just a little effort, such as enhancing it on the fly with style sheets, the content can be delivered via RSS (Really Simple Syndication) feeds or PDF files.

OUP hired interactive agency iFactory of Boston to assist in the integration of the MarkLogic Server with OUP's existing technology, as well as with the creation of the platform and the African American Studies Center. Humphreys expects the online resource to be completed this spring, which would be less than a year after work on the platform began.

One of the additional benefits of the approach OUP took to establishing the new system is that in addition to not being tied to a particular DTD, they will not be tied to a particular vendor, notes Humphreys. In fact, one of the primary challenges that OUP faced in creating a publishing platform was vendor management; there were many platforms available that would meet their current product requirements, but which would force the company to work with only one company on all future projects. OUP's approach was to have iFactory build an Application Program Interface (API) for the platform, so that for future projects, vendors could interconnect with MarkLogic using that API. Because of the inherent flexibility of XQuery and MarkLogic, vendors will be able to leverage each others' work while not forcing OUP into an overly rigid product or content structure.

## **BENEFITS IN DETAIL**

Humphreys says that one feature of the AASC highlights the benefits that XML, MarkLogic and XQuery provide: "At A Glance Pages." Using the tagging, linking and dynamic nature of the content and platform, these pages will bring together all the content on a particular topic.

For example, researchers studying Thelonious Monk will be able to view extracted biographical data, moments pulled from timelines relating to Monk, lists of his contemporaries, and excerpts and meta-information of each article on him in the AASC. "People will be able to

find and navigate through the content in ways not before possible," Humphreys says.

OUP expects to start another five to 10 projects within the next year and to complete each in less than six months, Humphreys says. In addition to being able to more rapidly prototype new online resources and reduce development costs, he says OUP's new platform positions the company to better handle a changing marketplace. He says, "XML,

MarkLogic and XQuery provide us with the flexibility to respond to the changing needs of an increasingly dynamic market. Because of that flexibility and our intimate knowledge of our content," he adds, OUP is better equipped to "compete in a Google world."

---

*John Gartner is a Portland, Ore.-based freelance writer and consultant. He is a regular contributor to Wired News and Technology Review. He can be reached at [JOHNGARTNER@COMCAST.NET](mailto:JOHNGARTNER@COMCAST.NET).*

---



**Bill Veiga, director of sales, MarkLogic.**



Mark Logic Corporation  
2000 Alameda de las Pulgas  
San Mateo, CA 94403  
[www.marklogic.com](http://www.marklogic.com)  
650-655-2300