Content-driven Applications in the Federal Government

Table of Contents

Content-driven Applications in the Federal Government

Maximizing the Value of Content: Go Beyond Search and Database Technology

State of the art net-centric applications allow you to meet mission-critical and mission-specific needs for information including:

- Leveraging actionable intelligence in real time to compare and analyze information across all your content sources to find important trends, track policy changes, enforce the law, or help locate a terrorist.
- Getting inside the boundaries of documents, deep into the structure to leverage additional metadata or semantically marked up information for deeper insight. No matter the format, you can have greater access to more relevant information.
- Allowing you to quickly and easily load XML content from different information sources with widely varying structures so that you can exploit that information in a wide variety of methods.
- Accessing, sharing, and collaborating with experts from within and outside your organization. Comment on documents, tag them with ad-hoc metadata, and rate them, making the content richer and easier to find.
- Assembling virtual documents (e.g., custom battle books) by pulling relevant sections of text and media from within millions of pages across thousands of sources stored within hundreds of terabytes of information. Get exactly the content you need—in a matter of seconds.

Today, government organizations are meeting these needs head on with a platform that allows them to get precise pieces of information from within stored documents and deliver the most relevant information in any format required, from a simple web page or PDF to a mobile-device alert or document format.

What does it take to build these mission-critical and mission-specific applications?

You already have the documents—reports, contracts, policy documents, eLearning materials, emails, message traffic, compliance documents, and training manuals. You may have tried to re-purpose these documents using traditional tools like search engines or relational databases but found them inadequate for the task. Any search engine can tell you which of these particular documents contain your keyword or phrase most frequently. Relational databases which have been extended with limited XML handling capabilities struggle to load real content and often become bogged down when trying to manage differing structures.

The most important information is commonly buried within sections, subsections, paragraphs, and even sentences within these documents. Enterprise search engines do not provide this level of granularity; they simply return a long list of links to whole documents where your important information may be buried. Relational databases force you to try and guess your queries ahead of time which is a nearly impossible task.

The challenge is to get inside the boundaries of documents so that you can connect the dots across and within different pieces of information, relate that information with other stored information, and pull together very precise pieces of information into a custom view or report. In a word, you need answers, not links, and you need them fast.

Today, government organizations are meeting these needs head on with a platform that allows them to get precise pieces of information from within stored documents and deliver the most relevant information in any format required, from a simple web page or PDF to a mobile-device alert or document format.

Mark Logic enables government agencies to improve knowledge management and information sharing, to analyze vast amounts of content, to deliver information in new ways, and to make it easier to complete mission-critical assignments while lowering costs through powerful XML–based technology.

MarkLogic Server includes a unique set of capabilities to store, aggregate, enrich, search, navigate, and dynamically deliver content.

With MarkLogic Server:

- Users get answers, instead of links
- Users have the ability to access content within and about documents, including metadata—not simply the documents themselves
- Users get content in context—in a format meaningful to them
  - Intelligence analysts: analyzing and collaborating on content regarding links and relationships
  - Soldiers: receiving custom battle books, training manuals, tactical situational awareness, geospatial visualization
  - Policy analysts: current litigation information

MarkLogic Server is simply the best place to put XML, and below are real-world examples of how defense, intelligence, and civilian agencies are all using an XML content platform today to meet their unique mission goals, and deliver useful information hundreds of times faster, for greater mission success.

Intelligence

Challenge: The intelligence community utilizes large amounts of content from open sources, and they must efficiently search through the information to discover important relationships within the data. MarkLogic Server helps solve critical mission challenges with this content:

- Integration among Open Sources
- Integration and fusion with all-source intelligence
- Enrichment of content with in-context semantic extraction
- Search over hundreds of terabytes of text
- Analysis of search using temporal, spatial, and relationship visualization
- Delivery of analytical results as finished intelligence product

Scalability, speed, and accuracy are critical in successful Open Source Intelligence exploitation, for purposes such as counter-terrorism, force protection, and evaluation of threat capabilities and intentions. Reliable intelligence comes not just from collecting a vast amount of information, but by analyzing that information and establishing relationships—making connections where none were apparent before—and then presenting that information in the appropriate context. This is the goal of open-source intelligence.

Open-source includes a diverse set of content, but the largest component by far is content harvested from the Web. Spiders collect data that is highly variable—some complete, some incomplete, some malformed— from a vast array of sites, in different formats, in multiple languages, from all over the world.

The success of such a system depends on the comprehensiveness of the data being evaluated. Therefore, the ability to support repositories that scale up to 100's of terabytes of text, distilled from petabytes of data, is paramount. The underlying content continues to grow exponentially as the Web grows, more sources are mined, and more data is extracted.

A wide range of applications manipulate the data, from geo-tagging tools that can tag documents based on their origin or recipient location to entity extraction technology that identifies people, places, and organizations—and can then establish potential relationships based on the text—to visualization tools that can map these relationships in time, geography, and relationship.

Underneath these applications MarkLogic Server stores and manages the information, provides sophisticated search capabilities, and delivers appropriate elements of the content to the applications on demand. As content is enriched, additional metadata is associated with each document—such as when the document was created, where it originated, and who created it—so that it can be searched as well and thus provide even more targeted and fine-grained results.

For example, the system can start with a name from a known terrorist blog and allow the intelligence analyst to drill down on various locations mentioned in documents that come from a selected range of dates. Based on all the other information already stored in the system, the intelligence analyst can then identify other postings originating within a selected geographic range in the past week, and extract from them the names of persons of interest for further scrutiny.

Being able to access content within documents—going beyond simple search, and digging inside documents to retrieve the precise, most relevant piece of information—is key to effectively combating terrorism. With this level of granularity, intelligence analysts are much more effective in creating product and accomplishing their mission.

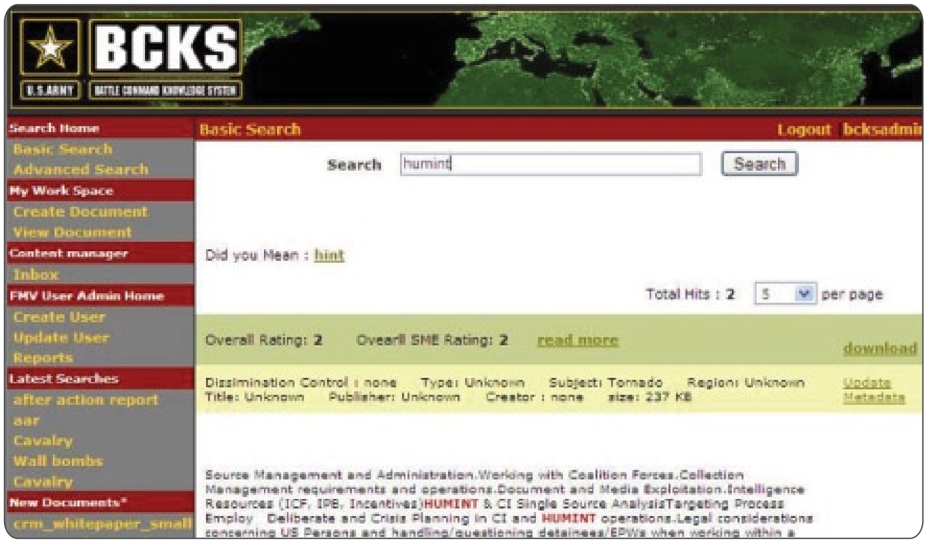
Department of Defense

Challenge: The U.S. Army needed a real-time, lessons-learned repository that allowed soldiers to share information with one another and access that information in real time from the theater of operation. MarkLogic Server allowed the DoD to address the following mission-critical problems:

- Knowledge management
- Content collaboration
- Search
- Custom publishing

The Army is highly regarded for its ability to collect and analyze the results of operations and training programs and then incorporate the lessons learned into doctrine and operating procedures. A knowledge base can be one of the most effective ways to share information. It can help create a community of users that are geographically dispersed—in different countries and in different time zones around the world. A knowledge base can also be one of the most effective ways to disseminate information quickly and securely.

In the case of the U.S. Army— and particularly troops in the field—getting immediate access to up-to-the-minute advice and information is essential. With the speed of the modern battlefield, however, the Army needed a more rapid and less structured way to share time-sensitive information as well as knowledge unique to a specific unit or area of operations.



Search humint

Did you Mean : hint

Total Hits : 2 5 per page

Overall Rating: 2 Ovearll SME Rating: 2

Dissimination Control : none Type: Unknown Subject: Tornado Region: Unknown
Title: Unknown Publisher: Unknown Creator : none size: 237 KB

Source Management and Administration.Working with Coalition Forces.Collection Management requirements and operations.Document and Media Exploitation.Intelligence Resources (ICF, IPB, Incentives)HUMINT & CI Single Source AnalysisTargeting Process Employ Deliberate and Crisis Planning in CI and HUMINT operations.Legal considerations concerning US Persons and handling/questioning detainees/EPWs when working within a

The U.S. Army owns and operates the Battle Command Knowledge System (BCKS). Here warfighters and mission planners can post combat derived information in a secure forum so soldiers in the field can help their comrades—providing advice, for example, on spotting and avoiding improvised explosive devices (IEDs). BCKS enables the people in our armed forces to leverage each others' and the Army Command's institutional knowledge and invaluable experience.

From a technology perspective, the BCKS is a repository of content with a wide range of information from documents to manuals to information feeds to user-created input. An important factor in maximizing the value of the information is that the content can be easily updated and/or augmented by field personnel in just a few hours to reflect experience in a day's battles.

For the BCKS to be effective, warfighters must be able to query this information by any means available, whether computer or laptop, PDA or RSS feed. The warfighter must receive information back—not just links, but complete content—in a way that they can read and process it. And, this information must come back fast.

Based on a content application built on MarkLogic Server, BCKS lets warfighters perform full-text and metadata searches, post and receive information in the field, and interact with the information.

Today, BCKS provides 90,000 soldiers and civilians with the ability to discuss ideas, share knowledge, and rapidly solve problems. The system, which supports both deployed units and training programs, has accelerated the transfer of expertise and experience, improved decision making, and ensured knowledge capture and sharing.

Civilian

Challenge: Within the U.S. Patent and Trademark Office (PTO), patent examiners needed to be able to quickly search across a constantly changing repository and produce custom reports from the search results. MarkLogic Server helped them solve the following business problems:

- Search
- Content mining and analytics
- Content delivery
- Custom publishing

The MPEP—Manual of Patent Examining Procedure—is an extremely important document for anyone working as part of the patent process. Published by the U.S. Patent and Trademark Office, this book is used regularly by patent examiners, patent lawyers, patent holders, and those requesting patents. It is constantly updated and continually evolving, as new patents are added and existing patents change and evolve.

Until recently, the PTO provided access to the MPEP through a client-based application for PTO employees. In other words, the MPEP was manually loaded on each client machine. For users outside the PTO, the MPEP was distributed as a PDF document. This process was difficult to manage and nearly impossible to support outside the PTO itself.

After an extensive RFP process, the PTO implemented a content delivery and custom publishing application based on MarkLogic Server, whereby the information within the MPEP is stored centrally and made web-accessible.

The system, however, is not a simple as a relational database and search engine combination. With the new system, users both inside and outside the PTO can search for and access information within sections, paragraphs, and chapters of documents—receiving content, not simply links—at a level of granularity not possible with enterprise search engines.

The content-delivery application is similar to a book-viewer, where users can navigate through links embedded within the MPEP, search on specific topics, bookmark topics, and even provide annotations within the MPEP itself. Users can also review a range of analytics based on information mined from the documents, such as related patents, citations, and other inventions by the same applicant.

For the PTO, the system has provided vastly greater usability, functionality, and efficiency. And, because updates to the MPEP take effect instantly, all parties have access to the most accurate and up-to-minute information at all times.

Mark Logic – Providing the Core Technology

The defense, intelligence, and civilian real-world case studies are all based on MarkLogic Server—the industry's leading XML content platform.

With MarkLogic Server, these agencies are able to load, query, manipulate, and render repositories as large as 100's of terabytes of content, quickly and easily. With the ability to go well beyond the capabilities of ordinary search engines, MarkLogic Server gives warfighters, intelligence analysts, and civilian staffers precisely the information they need to be more effective in their mission. They are going beyond what has been available with traditional tools to solve complex integration problems like handling all source/open-source information as well as perform geo-spatially enabled queries against very large contentbases.

With MarkLogic Server, users can also access and share vital information directly with each other, regardless of the format or location of that information. The technology combines the best of relational database management systems and enterprise search capabilities, running highly complex queries to produce the most accurate, targeted results.

MarkLogic Sever was created to address the complex nature of XML content. Some specific assumptions were involved at the outset of these efforts, assumptions very different from those of 30 years ago when the relational database was invented:

- The schema is unknown
- Multiple schemas may coexist
- The system should be ACID compliant
- Documents may be small (1KB) and large (1GB)
- Collections of documents may be small (thousands) and large (billions)
- XQuery is the query language
- Indexes should be automatic and universal
- The system should be robust on commodity hardware

The result of these efforts is the industry's leading enterprise class XML content platform. There are a number of functional areas that address the needs of government organizations looking to build information and knowledge management, search and discovery, and information delivery applications:

Native XML Persistence

Documents are stored as documents; they are not deconstructed into more primitive data types. XML is parsed by the server, indexed, and stored in a proprietary compressed DOM. In MarkLogic, the XML is parsed once—when it is loaded—and never again. All interactions with the document occur with this optimized, compressed form. Documents in a MarkLogic are organized by a variety of means, including directories, collections, security, and metadata. Interacting with documents in MarkLogic is similar to interacting with a file system—take any WebDAV client, such as Windows Explorer, and access documents and directories in the contentbase to read or edit in a familiar fashion.

Universal Indexing

Because one of the design assumptions of MarkLogic was that the schema is unknown, documents of multiple schemas may coexist in the same contentbase. When a document is loaded or updated, values of all elements and attributes are indexed, the qualified name of all elements and attributes, as well as the hierarchical relationship of each element in the context of the document. This combination of values and structure is encoded in our patented index called a universal index. By combining what are typically separate indexes into a single universal index, it becomes possible to perform rapid evaluation of arbitrary XQuery expressions against billions of variant structured XML documents.

The universal index is completely automatic. It is unnecessary to explicitly declare which elements and paths should be indexed by the server: all elements in all paths are indexed by default. When new elements appear in documents, they are transactionally added to the index with no intervention.

High Volume Transaction Processing

Excellent and predictable performance is a key characteristic of MarkLogic. Many features come together to enable high performance in a robust system. Rapid in-memory ingestion, asynchronous index optimization, sequential disk IO, the universal index, 64 bit architecture, multiple layers of distributed caches, and other features all collaborate to provide millisecond response time against multi-TB collections of XML content. While there are no standardized benchmarks for XML content, Mark Logic has conducted numerous exercises with customers that include:

- Over 1 billion documents
- Hundreds of simultaneous schemas
- Hundreds of queries per second
- Over 4MB/second per CPU sustained load rate
- Over 1.5MB/second per CPU sustained load rate with simultaneous read/update/load operations

Mark Logic has reference customers in production with multi-TB systems.

Conformance to XML Standards

Mark Logic is a key participant in the W3C committees for XQuery, XPath 2.0, XML Schema, and other relevant standards. MarkLogic Server provides the most complete implementation of the XQuery standard, as well as hundreds of extensions to the standard for features such as update, xml search, try/catch, security, triggers, and many other types of semantics.

Ease of Integration

Because of this conformance to XML standards and universal indexing techniques MarkLogic Server can easily integrate with any 3rd party product that supports XML. MarkLogic has been implemented as the "hub" of a hub and spoke model of application integration where MarkLogic speaks all of the different flavors of XML spoken by the various systems. This speeds and eases integration of disparate systems in a large system of system architectures.

ACID Compliance

MarkLogic Server provides a transactional system that adheres to the ACID[1] model. Transactions are handled in a fully non-blocking manner. This type of architecture ensures users will never have to wait on the read/write operations of others. In order to provide transactional capabilities MarkLogic Server implements a temporal database. Updates are not processed in place, thereby eliminating the complex and expensive overhead of disk management incurred by most relational database management systems. In MarkLogic, all transactions are evaluated at a system timestamp, and documents are valid at a specific timestamp. All updates are processed in memory and written to disk in an optimized, asynchronous, serialized disk IO. Optimization of indexes is automatic, asynchronous, and processed as serialized disk IO.

High Availability

In a MarkLogic deployment high availability is delivered through several mechanisms including journaling, clustering, sub-second auto-reboot, automatic failover, and hot backup/restore. The journaling system used in MarkLogic ensures data integrity during update operations. A clustered architecture spreads the storage and query processing load across multiple servers over commodity hardware and gigabit Ethernet. Sub-second auto-reboot reduces system downtime for individual node failure, while automatic hot failover ensures transactional integrity through node recovery. Hot backup and restore capabilities allow for system administration and maintenance without impacting system performance or uptime. These features, taken together with the ease of installation, deployment, and administration—that are characteristic of MarkLogic Server, provide an exemplary platform for portfolio reconciliation.

All this sophisticated technology is maximized for scalability. MarkLogic Server provides millisecond response times against multi-terabyte content bases.

In the above examples, these defense, intelligence, and civilian agencies needed a way to maximize the content they already had, and find information they could not find with traditional search engines and relational databases. If you have the same requirements, Mark Logic can help your agency become more responsive, more agile, and better equipped to exceed mission expectations.

About Mark Logic Corporation

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.

[1] Atomicity, Consistency, Isolation, Durability. These are four goals that databases must achieve. It states that only complete, valid transactions are written to the database, and prevents any reads from returning intermediate data from one transaction.

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
Sep 2008