Inverting the Corporate Information Pyramid:

User Driven Information Discovery

By Bjørn Olstad, CTO, and
Silvija Seres, VP Strategic Market Development, FAST

The information chain is a pyramid

In past decades, information-intensive businesses profited from dropping storage costs. Computer systems offered simple and inexpensive data retention—in databases, document repositories and email archives. Within the scope of their design, these systems succeeded, creating massive amounts of digital information typically stored in application "silos."

The competitive landscape has changed. Return on intellectual capital drives success at least as much as transactional growth does, and infrastructures that served information storage well enough in the past fail miserably at the task of exploiting the enterprise's information. Companies also face new risks, such as compliance laws demanding higher levels of visibility. Information access has become strategic: In the knowledge-based economy, companies win by providing effective access to information.

Consider the traditional enterprise information infrastructure: an information value chain with production at the low end, consumption at the high end. The levels in the chain create a pyramid that widens from top to bottom, its width at any point corresponding to investment and use at that level. In most enterprises, expensive RDBMS storage and legacy architectures make this pyramid bottom-heavy. We believe there is an innovative way to turn the cost structure of this pyramid on its head.

Inverting the Pyramid

To quote the CIO of a large, multinational enterprise: "We are world champions at producing information. Finding what we need is an entirely different matter."

At the bottom of the pyramid are data sources, which in this enterprise include corporate, regional and local documents, in five languages and 220 formats. To manage this information and its associated applications, the enterprise has large teams of database architects, programmers and managers, plenty of software licenses and numerous high-end servers.

Further up are tools and people who organize and refine, people who provide access for knowledge workers. Librarians, application managers and user support spend man-year upon man-year building taxonomies, applying metadata, controlling access and developing new information services.

At the top of the chain are the information consumers: employees, analysts, management and customers looking for product information.

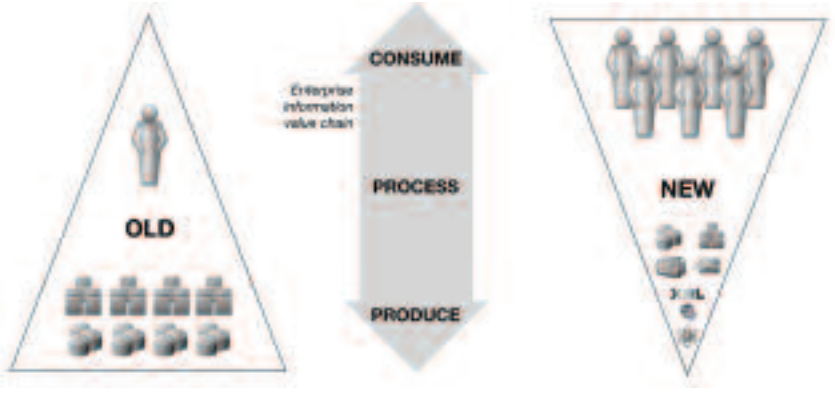
Information access in RDBMS and legacy systems is inflexible, costly and slow. Metadata maintenance is complex, expensive and imprecise; new information services are major undertakings; the system architecture is scattered and expensive to maintain; coordination of work is non-trivial. There is duplication of work; there is duplication of content; and still, important information gets lost. The information is often untimely.

If we introduce an information architecture based on a modern search platform, our bottom-heavy pyramid changes dramatically. A good search platform turns it on its head, and thereby improves information access, reduces total cost of ownership and increases overall enterprise performance.

Immediate Information Discovery

The traditional pyramid grew out of a focus on transactions. There was less data, and what there was of it was mostly structured. Times have changed: Information consumers understand the value of accurate, timely information and know how to use it. Better service affects their performance directly. Better information access is now a strategic requirement. This new "immediate information discovery" model, with intuitive search front ends, relevant results, sub-second performance and advanced tools to filter and refine the answers, drives many new competitive business models.

The immediate information discovery model addresses needs throughout the value chain. In our example enterprise, the management wants to grow the business, so new revenue streams and low TCO are important. IT managers like the innovation potential it affords their services; product managers love the increased user traffic. Existing users relish new ways to access and use information: Contextually sensitive search and navigation, user-specific relevancy ranking, spell checking, search for similar documents...all combine to create new ways to tap company information. New customers are attracted by new search-based services: vertically focused subject search, topic maps, update alerts and search for experts. These capabilities open new revenue channels, increase the



The enterprise information value chain can be seen as a pyramid, with its width representing levels of investment and use. In most enterprises, expensive legacy systems make this pyramid bottom-heavy. With search, a large number of end users, and the enterprise on the whole, benefit from a more flexible and effective access to information.

number of visitors, enhance user satisfaction and improve business performance.

New services include digital reports and books, video and audio content search, enhanced topic search and reference research, all combined with the traditional strengths of the enterprise: subject matter expertise and proprietary content management. And the cost aspect: Through this search-based model, the example enterprise has seen 60%-80% savings in HW/SW acquisition costs through fewer licenses and the use of commodity servers, and 10% savings in operation and scaling costs through faster installations, easier scalability, simplified feature extension and fewer people.

Semantic Index

Good enterprise search engines turn the information pyramid around. They enhance the whole information chain, placing the bulk of functionality where it is most needed: at the information consumption point. Some search engines, however, go beyond information access enablement. They transform the whole information storage, processing and consumption model by enabling true semantic—meaning-based—information handling.

This model extends the "semantic web" idea to all content, structured or unstructured. It automates the recognition of information context and meaning, at several levels, and replaces costly and slow specialist data mining tools with a single user-friendly system that enables *end-users* to create sophisticated queries and get accurate results at lightning speed. The semantic model maintains enterprise control over business and security paradigms, but moves the power balance in information access from producers to consumers, where it belongs.

This contextually aware information model is based on an embedded part of the search engine called the *semantic index*. The heart of a contextual search engine, this index stores files, database content and multimedia with no loss of structural information. It retains HTML structures and database records; it recognizes a practically infinite number of entity names; it knows the granularity of sentences and paragraphs. It translates the original context into searchable elements, and thereby brings an unprecedented level of "understanding" to information providers and consumers. It is grounded in years of research and development, and is the core of next-generation search engines.

The semantic index preserves contextual information in the data, and provides users with unprecedented analytic freedom and speed.

This index can be explored on two levels: the mechanistic representation level, which affects the richness with which information can be encoded in the search engine, and the intelligent information analysis level, which affects the user's power to explore context-rich information.

"There is an innovative way to turn the cost structure of this pyramid on its head."

Full contextual representation of data in the semantic index has two main consequences for end users. One is that the index contains enough information to support any XML-based query, and thus opens the way for full integration of structured and unstructured data. It accepts complex queries based on X-Path or X-Query and solves them in sub-second time. Any native content can be made searchable with its structure preserved, *without* a priori knowledge of data structures or schemas. This is true schema flexibility: schema independence, but with full schema exploitation.

Contextual analysis lets users change the *granularity of search* at will: they can choose the atomic unit of search dynamically. This is a fundamental change in the way search engines work. Traditionally, the unit of indexing (e.g., a Web document or a PDF file) was also the unit of retrieval. Now the unit of recognition and retrieval can be moved to a sub-document level. This means improved search precision and less work for users—no more wading though long documents in search of the relevant sentence, or listening through an entire newscast in search of the

SEARCH REVOLUTION *continues on page 24*

SEARCH REVOLUTION *continues from page 5*

relevant snippet. Plus, contextually aware analysis allows *on-the-fly text mining*, where users can pose queries previously reserved for expert data miners with too much time on their hands. Such queries combine structural data context and schemas with recognized entities, categories and other implicit information, then use sophisticated statistical matching to answer questions like: "Which persons appear in sentences that contain any company names and the word 'scandal'?"

The semantic index cannot be implemented as an incremental change to existing search platforms. It is not simply a new feature, a new capability or grafted-on handling of a new content type. It is a search revolution.

Budding Personal Pyramids

We have explored how an enterprise can invert its information pyramid to provide its users with the information they need—simply, quickly and cheaply. We have explored how an enterprise can use the contextual power of the semantic index to make the pyramid intelligent. What more can the enterprise do for its employees and customers?

Well, people have personal data. Laptops and PDAs contain a wealth of information in files and emails, chats and blogs. People want access to this personal information in the same way that they have access to corporate information or information from the Web. So they install desktop search tools, from random providers. These tools have known security issues. What they don't have is finely tuned access to corporate information. What they don't have is all the advanced linguistic and analysis tools described above. It is like giving people a Ferrari to explore corporate information, but a Matchbox toy car to play with their personal information.

This toy car can do damage. Security leaks are one thing. Traffic leaks, leading users away from the portal and to other search engines, is a strategic loss for the enterprise.

The agile enterprise can provide protection of its content and its users in the face of demands for personal searching. It can provide its user base with its own personal search platform (PSP), based on the same principles as the enterprise search platform, and compatible with it. This has many advantages. First, there is reuse of information between the platforms, with the same capabilities of contextual insight, semantic indexing, and advanced processing and mining tools. Second, there is significant reuse of skills, processes and resources for deployment and maintenance of the two systems. They share functionality in the control systems. Third, the feature set is greatly expanded, compared with freely available desktop search products. It includes navigation, content preview, highlighting, similarity searching, taxonomy usage and native content editing. Finally, where desktop search tools tend to be monolithic, PSP is modular. The main modules of PSP—the indexing engine, the content source APIs, the federation engine, the front-end APIs and the control engine—are all flexible and separately programmable, allowing the fine-grained granularity in control and deployment enterprises need to make this PSP their own—to simultaneously support their users' needs and their business models.

Our example enterprise uses PSP to connect its employees' personal content with the company's intranet content, and to enable offline searching of some of its central resource management tools. Another example, a large national portal, allows its customers to download their own copies of PSP. Since customers can search both the Web and their own content from their personal PSP installations, this portal has effectively locked in its traffic in a user-friendly way, even when people "only" want to search their own files.

PSP provides enterprises with tools that enable safe access to personal information and increase functionality far beyond that of standard desktop platforms. PSP protects company brands by avoiding traffic leaks to other search engines.

Pyramid To Go?

Our example enterprise now has a consumer-focused, intelligent enterprise search that allows its users safe access to personal content. But there is more—mobile intranet, for example. Users like their mobile phones and their PDAs, and the enterprise is considering the third pillar of search in an enterprise setting: the mobile search platform (MSP). MSP allows users to search corporate or personal content from handheld devices—to find a customer address while on the road, or browse digital products from preferred content partners. The enterprise can also use MSP to create its own crawl of the mobile Web, gathering only high-quality documents relevant to its user base. MSP provides the advantages of the other two platforms—modularity, contextual insight, high-speed searching in a user-friendly context—and exploits still more synergies between personal search, mobile search and enterprise search. Shared content, shared features, economies of scale. MSP is secure and tunable. Our example enterprise is now fully searchable, and the true value of its information is seen in its users' ability to use it, not just their ability to produce it.

Traditional information systems optimize information production and storage, not information consumption, and too often leave would-be information consumers unsupported and uninspired. To stay competitive, the enterprise needs to enable and entice its employees and customers to make the best possible use of what may be its most valuable asset: information. Companies that apply information, instead of just creating and storing it, have a huge and sustainable strategic advantage. ▮

"Traditional information systems optimize information production and storage, not information consumption."


Bjørn Olstad

Bjørn Olstad, Ph.D., serves as the chief technology officer at Fast Search & Transfer (FAST). Before joining the company, Dr. Olstad held key positions within General Electric Medical Systems, including director of research and development for Cardiac Ultrasound. He has served as a professor in computer science at the Norwegian University of Science and Technology (NTNU), where he was awarded the youngest professorship ever.


Silvija Seres

Silvija Seres works in Norway as a vice president of strategic market development with FAST. She holds an MBA from INSEAD in France and has extensive scientific background in algorithm design and optimization, with a Ph.D. and Prize Fellowship from Oxford University. She has held several visiting posts at leading research institutions and has developed international educational programs.