

Search Solutions for the Enterprise

Incorporating search into your IT architecture

Search solutions for the enterprise can be a boon to organizational productivity and performance, provided they are deployed effectively and efficiently.

INTRODUCTION: THE IMPORTANCE OF ARCHITECTURE

Search solutions for the enterprise are emerging as powerful tools to help organizations achieve a universal objective — making better use of information.

Multi-tiered search solutions that comb the desktop, departments, the enterprise and the Internet for information promise to boost productivity and increase organizational effectiveness. These solutions offer a path to overcoming major issues impeding organizations today — redundant depositories of identical data, various retrieval methods and a lack of consistent semantic definitions across the enterprise. They also can help meet the growing expectations of workers that enterprise systems be as easy to search and as user-friendly as Internet software and services.

A key to achieving the benefits of organizationwide search is effectively integrating the search solution with other technology systems within the organization. BearingPoint has identified the key principles underlying organizationwide search, and from that has developed a recommended architecture for its successful deployment.

THE PRINCIPLES OF ORGANIZATIONWIDE SEARCH

Several general principles guide development of search solution architecture:

Simplicity. The operating principle of organizationwide search is to keep systems and solutions simple. Complexity inhibits productivity of software developers; makes products difficult to plan, build and test; introduces security challenges; and causes end-user and administrator frustration. Computing and communications technologies have evolved to the point that a services model is viable. Service-enhanced software and lightweight development of search solutions addresses user demand for a compelling, integrated experience that “just works.”

Functionality. The priorities of the search solution architecture are that it will be (in order of priority): usable, reliable, secure, scalable (in terms of users, messaging volumes and services), and capable of independent growth through addition of new services provided by the search solution framework.

IN THIS POINT OF VIEW:

INTRODUCTION: THE IMPORTANCE OF ARCHITECTURE	1
THE PRINCIPLES OF ORGANIZATIONWIDE SEARCH	1
OPERATING GUIDELINES FOR SOLUTION DEVELOPMENT	2
THE SEARCH SOLUTION ARCHITECTURE	3
Conceptual Architecture	4
Example Search Solution Architecture	4
THE SEARCH SOLUTION OPPORTUNITY	8

Dependability. The search solution platform needs to be a solid infrastructure. Consequently, it should avoid cutting-edge technologies. As a general rule, it will use technology in its core architecture that:

- Provides maximum transportability across any environment.
- Has been tried and trusted in comparable usage scenarios elsewhere.

Mission-criticality. A search solution for the enterprise should be viewed as a mission-critical enterprise component that is highly secure, highly scalable and available 24x7. The search solution backbone must be architected from the ground up with these attributes in mind.

Accessibility. To support wide usage, the search solution must be architected for pervasive access. This capacity must be designed into the architecture from the outset, because such functionality cannot be bolted on later.

Flexibility. The concept of flexible services is the key to a successful search solution. Not in the sense of an end-user service, but services as discrete units of functionality available on the search platform through defined interfaces.

OPERATING GUIDELINES FOR SOLUTION DEVELOPMENT

Based on the principles above, the search solution architecture should adhere to the following guidelines:

Open standards, open technology and common interfaces. The search solution should be based on open standards that maintain technology independence by leveraging standards such as J2EE, .NET and Web services. This allows the greatest flexibility, helps prevent the need for interface changes when another application interface is altered, and hides the complexity of application programming interface (API) connectivity between related systems and applications.

Modular architecture. The search solution is based on an n-tier, component-based architecture. This approach provides abstraction of each layer and for each component, creating component independence and logically divided functionality. The architecture should be loosely coupled and highly granular.

Interoperability. The use of extensible markup language (XML) and extensible style sheet language transformation (XSLT) creates interoperability that provides an abstraction separating the presentation domain from underneath logic. Dividing presentation into a presentation sublayer and a presentation-interface sublayer isolates presentation logic. The architecture relies on the independence of devices, platforms and technologies to establish interoperability.

Metadata management. Metadata management addresses the creation, storage, access, aging and maintenance of metadata content, including governance and synchronization of shared metadata. Enterprise metadata management allows an organization to:

- Automate the assignment of metadata to existing content. This makes search more effective by finding the right content assets regardless of where they are.
- Develop capacity to associate metadata readily with content assets as close to the time of their creation as possible.
- Establish clear organizational stewardship for development and ongoing management of taxonomies and authority lists.
- Implement processes for keeping taxonomies current and auditing metadata associations.
- Obtain different views of the same content, such as by role, time and process.
- Integrate content sources—The search solution should have an open architecture to support integration of multiple content sources. Open standards are required to make the platform function smoothly, particularly with respect to interoperability of multiple content sources.

Multi-tiered search solutions offer a path to overcoming major issues impeding organizations today.

Security. A search solution for the enterprise should embrace and extend the security principles associated with the various sources of content within the enterprise. Principles of authentication, authorization, audit and identity management should be applied so that only appropriate users can have access to privileged information. Search solution security should comply with enterprise security policy and regulatory requirements. In some cases, document-level security will be required to make effective use of unstructured content.

Network infrastructure. Solution deployment must address the impact on network infrastructure and the organization the infrastructure supports. In the case of organizationwide search, considerations include where the data being searched resides, if outside people or organizations will leverage the search functionality, security and overall performance.

General network design considerations and guidelines include:

- How much traffic will the search generate on the enterprise network and associated segments? The number of data sources crawled, where the data sources reside and the type of network infrastructure deployed will drive this design consideration.
- Where does the user base for the search solution reside, in the same building, on a different campus or across a wide area network (WAN)?
- Has network traffic classification been performed? Traffic classification is an important consideration in an environment where storage area network (SAN) data replication, such as Symmetrix Remote Data Facility (SRDF), is taking place. The organization must consider the impact of the search on the network and determine proper network segmentation policies and quality of service (QoS) implementations.
- What type of access control is available within the network? In many implementations, the search solution will be used to access and serve proprietary or otherwise confidential data, making proper deployment crucial. Proper network segmentation, virtual local area network (VLAN) design, firewall implementations and network logging must be considered.
- What type of network resiliency is required? Because it serves the finance department, human resources and other management functions, a search solution for the enterprise is often considered a critical application requiring high availability. To achieve this, the search solution is often deployed in a clustered fashion with single- or multi-site failover options.

Disaster recovery and business continuity. Adhering to disaster recovery and business continuity principles requires hardware and software redundancy to minimize the single points of failure. The general design philosophy is to use redundant components to create redundant infrastructure. This increases resiliency to individual component failure and improves system availability to internal and external users and processes.

THE SEARCH SOLUTION ARCHITECTURE

The architectural concept for organizationwide search is a robust architecture that addresses the goals of the enterprise. The concept becomes reality as a modular, n-tier enterprise architecture that conforms to the standards and guiding principles outlined above. It can support a search solution that addresses current organizational needs, supports future growth, speeds time-to-market and lowers total cost of ownership (TCO).

Conceptual Architecture

Figure 1 describes the search solution conceptual architecture.

The conceptual architecture components are:

Content sources. Content sources include the enterprise intranet, Web sites, file servers, content management systems, such as Documentum and FileNet, and enterprise applications spanning the breadth of transactional systems, including customer relationship management (CRM), supply chain management (SCM) and enterprise resource planning (ERP). Fundamentally, this block represents the various systems within an enterprise that could serve as sources of content users would want to search.

Connectivity services. The connectivity services layer provides the basic adapters to the underlying sources of content. This layer will include two kinds of adapters: connectors and federated access. Connectors provide for build-out of various technology adapters required to connect to content sources. Federated access connectors provide connection to third-party content.

Metadata services. The metadata services layer provides additional value-added services to the content. It provides two types of services: metadata mapping and data dictionary mapping. Metadata mapping services help define the enterprise data across various systems. They also provide for single mapping between common elements via the build-out of the business data dictionary across multiple systems. The business data dictionary is represented by a set of metadata definitions and representations of data elements.

Federated access. The federated access layer provides for the build-out of query brokers that can obtain information from underlying transactional systems. Accessing the enterprise data requires building a universal index for related semantics.

Example Search Solution Architecture

Figure 2 describes an example search solution architecture derived from the conceptual architecture above.

The layers of the architecture are:

User interface

This is the user interaction layer that controls user interactions through a Web browser or even a handheld device. This layer provides the presentation of screens and associated data to search users.

The user interface is an XML-based interface dynamically created on the fly by the application layer. Depending on the querying device, an XSLT file delivers the appropriate presentation. This separates presentation from content and clearly divides responsibilities. This layer also ties in closely with the security components of authorization and authentication to make sure that search results are displayed to those people with appropriate authorization to view them.

Key considerations for enhancing the usability of the search solution include role-based presentation, view/edit and contextual help modes, and paging for search results.

Figure 1. Search Solution Conceptual Architecture

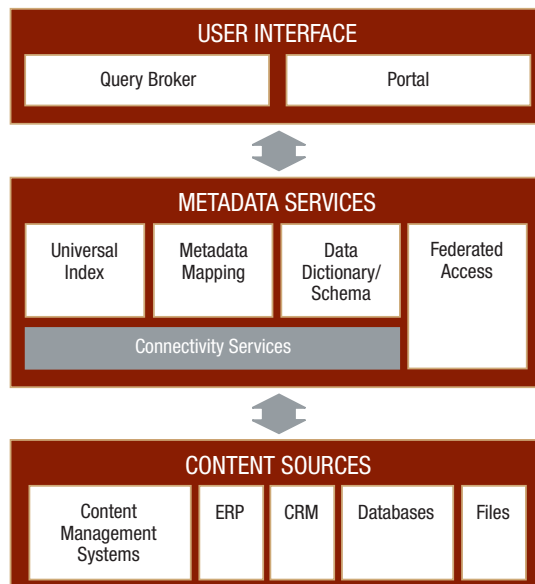
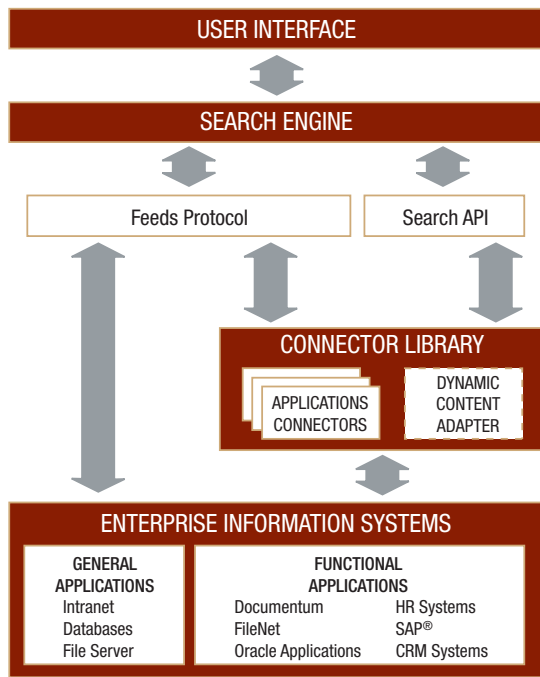


Figure 2. Example Search Solution Architecture



A typical enterprise environment has numerous document and Web-content stores within the confines of the intranet. Within these stores are various technologies leveraged to perform the actual hosting of content. Hosting includes inherent security mechanisms, individual index catalogs and back-end stores of the content itself. The search solution architecture should provide capabilities within three domains:

- **Query preprocessor.** The query preprocessor is a central service that can be called from a variety of places. Its job is to break the query down, validate spelling, translate to alternate languages, identify synonyms, identify concepts and identify the most appropriate query type for the associated content store; for example, keyword, concept, Boolean and full-text.
- **Catalog adapter.** Catalog adapters provide the integration contract between the initial query and the back-end content index catalogs. A single index capability across all content stores will not provide

the security characteristics required by the security information agents. Instead, the recommended implementation method is to have adapters that can broker the query to the base content catalog on behalf of the individual submitting the query.

- **Results aggregator.** This domain collates the results from various catalogs that match the query and presents them back to the user. Results should include the ability to provide feedback on any of the results returned, inferences to alternate search queries, and provisioning to apply a faceted or categorized search back to the query engine.

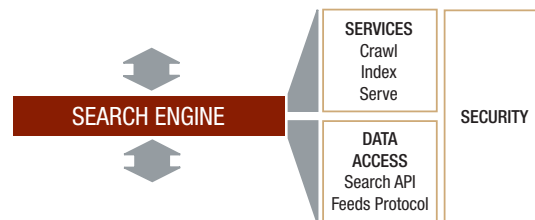
Search engine

The search engine is a hardware and/or software product that provides the features of organizationwide search in a box. It is expected to search hundreds of different file formats in any language and can index millions of documents. It must have security features to help ensure that users see only documents to which they have proper authorization.

The main functions the search engine should provide include (see Figure 3):

- **Serving.** Serving provides a standard search interface that can be hosted from the search engine by default. This interface should be customized to modify the underlying XSLT style sheet. Additional search features such as KeyMatch, synonyms and filters help promote specific Web pages as part of search results. Synonyms suggest alternate words or phrases for search queries. Additional features create logical information buckets called “collections,” which help meet specific user search needs.

Figure 3. Search Engine Functionality



- **Crawling.** Crawlers are agents that request and retrieve documents from Web servers for automatic indexing. Data crawling can be performed using two methods: data discovery and data recovery. A configuration interface allows crawling from various data sources. All content is aggregated to create a master index and the index is updated on re-crawl.

Crawling can be considered a three-step process: initial discovery (new URLs to crawl), indexing and continuous crawling. Each new document encountered by the crawler is scanned for links. The links are either traversed immediately or scheduled for later retrieval. The crawler is adept at dealing with secured content and handles secure hypertext transfer protocol (S-HTTP) communications. The search engine crawler can negotiate basic authentication, NT LAN Manager (NTLM) authentication, and custom cookie and form-based access. The engine should crawl content from databases, including Oracle, SQL Server, MySQL, IBM DB2 and Sybase. A data type the crawler cannot access can be fed directly to the search engine in an XML format.

- **Indexing.** Several methods can be used to index the data upfront:
 - A feed to the search system using standard or custom adapters.
 - Accessing from the search system directly (only the standard adaptor supplied with the system can access data directly).
 - HTTP or non-HTTP data feeds.
- **Feeds.** Some types of documents can best be pushed to the search engine using feeds instead of being found through links on crawled Web pages. These include:
 - Documents that cannot be fetched using the crawler. For example, records in a database or files on a system that is not Web-enabled.
 - Documents that can be crawled but are best re-crawled at different times from those set by the automatic crawl scheduler running on the engine.

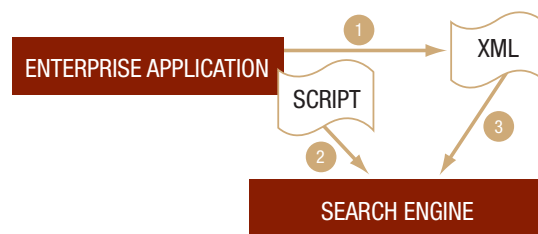
- Documents that can be crawled but are without links on the Web site that allow the crawler to discover them during a new crawl.
- Documents that can be crawled but are more quickly uploaded using feeds because of Web server or network problems.

Possible feed types include:

- **Content feeds**—Content feeds contain both URLs and their contents. They may also contain some metadata such as the last-modified date of the record. Any data source name can be specified for a content feed.
- **Web feeds**—In a Web feed, each of the records contains URLs but not contents. These URLs are crawled as normal. For such feeds, the term “Web” is used as the name of the data source. All URLs must contain a fully qualified domain name (FQDN) in the host part of the URL.
- **Database feeds**—Custom connectors can be written to push records from a database for creation of an index.

The setup process for search engine feeds is shown in Figure 4.

Figure 4. Search Content Feeder API



- 1 Script exports data into XML format to generate the feed file
- 2 Script posts the XML file to the search engine
- 3 Search engine imports the XML file and indexes the content

- **Security in the search engine.** The search engine should augment the enterprise security system. It must run all services behind a firewall, opening only a few posts to allow communication with the engine through the firewall. The search engine crawls and indexes both public and confidential documents. An optional security package enforces the organization's document-level security policies.

Delivering the correct set of search results to any user is based on filtering the results of the searched index. The content is divided into two categories: public content and all content. Users can search over public content only or over both public and secure documents in the index, as specified by the index administrator. By default, content remains completely secure so that users without access will not see any content that they are not authorized to view.

The search engine should support basic and NTLM (NT LAN Manager) authentication and form-based authentication. Each of these authentication protocols is handled differently by the search engine and requires a different setup by the administrator. Each method can crawl and index the protected documents on intranet sites and can authenticate on a search of those documents.

A search authorization API allows a Web service to translate between the search engine authorization API and the enterprise server that provides access control services, referred to as the access connector (AC). The AC provides a layer between the search engine and the organization's access-control system.

Connector library

This software layer forms the messaging and/or the access layer to various data sources in an enterprise, ranging from structured data sources to enterprise applications. Typically, an integration platform or custom connectors or adapters are built to integrate the data sources.

Custom connectors can also be built following these steps:

- Convert the data into XML in the format specified in the search engine feed.

- Upload the XML to the engine using the HTTP protocol.

This layer forms the underlying backbone for exposing information across an enterprise regardless of source or storage format. This layer thus allows loose coupling between services and integration adapters built from various data sources.

The metadata defined in this layer could also provide the infrastructure necessary for consolidating information across disparate sources in an enterprise. Typically, this also enables quick discovery and retrieval of items that match semantically. Thus, a unified meta-model establishes consistent and contextually correct metadata definitions for consolidation of disparate information. It also enhances search results with information from context-driven links to structured data and enterprise applications.

Information sources and types

Information exists everywhere in enterprises today. It is available in structured or unstructured formats, in proprietary or non-proprietary repositories (see Table 1). Search engines can search information that can be indexed so retrieval is easier and search results are relevant. It therefore is imperative to classify data sources likewise and then use appropriate retrieval mechanisms.

Table 1. Information Sources and Types

	PROPRIETARY	NON-PROPRIETARY
Structured	Database management systems (Oracle, SQL Server, Sybase, Informix) Document management systems (Documentum, FileNet, etc.) Enterprise applications such as ERP and CRM	Database management systems (mySQL)
Unstructured	Network file systems/file systems Microsoft Office documents Image, video and audio files	Open office documents File systems/file server Text files E-mail

Structured versus unstructured. The major contrast between unstructured and structured information is a predefined information model. The easiest analogy for structured information is something that could be put in a predefined format like HTML or something that falls into a table, such as with database technology. Another way to look at structured information is in terms of metadata—information about information. Unstructured data, on the other hand, does not have a predefined format and thus exists in free-flowing form. Examples of such data include text files and PDF documents. In the unstructured world, the type of search, in terms of the tools intended as a solution, becomes an information retrieval problem.

Proprietary versus non-proprietary repositories. Enterprise information is available today in ERP, CRM and knowledge management systems, as well as in static and dynamic portals. For information retrieval, some of these systems provide proprietary formats for publishing information. Some employ non-proprietary formats such as XML, which is steadily gaining traction. Also, although text today remains dominant in many applications, the relevance of other media types, such as image, audio and video, increases steadily. To this end, an efficient combination of automatic text retrieval, retrieval in metadata (usually created manually) and content-based retrieval of multimedia data is needed.

THE SEARCH SOLUTION OPPORTUNITY

A search solution for the enterprise can be a catalyst to greater employee productivity and more informed decision making. However, any search solution must be brought into the enterprise environment with a clear understanding of its architectural requirements and capabilities. Careful architecture planning and deployment are essential to realizing the powerful benefits a search solution offers.

To learn more about how our solutions can empower your company, [Let's Talk](#).

GLOBAL MANAGEMENT AND TECHNOLOGY CONSULTING FOR TODAY'S BUSINESS ENVIRONMENT

BearingPoint is a leading global management and technology consulting company that serves the Global 2000 and many of the world's largest public services organizations. Our experienced professionals help organizations around the world set direction to reach their goals and create enterprise value. By aligning their business processes and information systems, we help our clients gain competitive leadership advantage—delivering results in an accelerated time frame. To learn more, contact us at 1.866.661.FIND (+1.603.589.4089 from outside the United States and Canada) or visit our Web site at www.bearingpoint.com.

BearingPoint provides strategic consulting, application services, technology solutions and managed services to Global 2000 companies and government organizations.

BearingPoint

1676 International Drive
McLean, VA 22102
www.bearingpoint.com

