IBM®

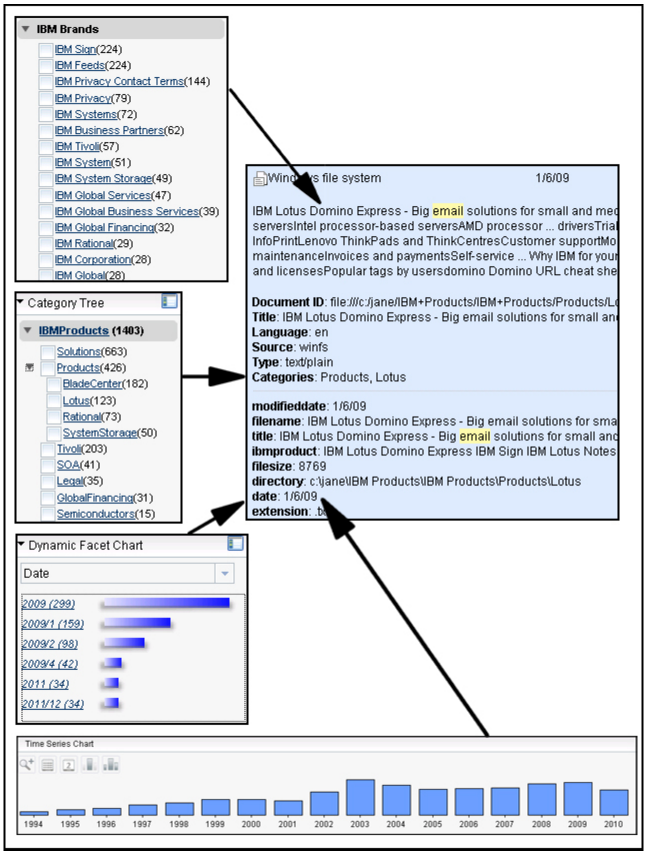
Redbooks®

Building Enterprise Search Solutions using IBM Content Analytics with Enterprise Search

IBM Redbooks Solution Guide

IBM® Content Analytics with Enterprise Search (ICAwES) addresses two categories of use cases: content analytics and enterprise search. Content analytics focuses on the analysis of a set of content to find patterns, trends, and anomalies in that content. Enterprise search focuses on the discovery and retrieval of documents by using various query and visual navigation techniques. The ICAwES enterprise search solutions can integrate fields from multiple content repositories to create a single, integrated user search experience. In addition, the enterprise search solutions can use fields and facets in various ways to create diverse views of your search result set, thus helping you identify the hidden meaning of your unstructured content. This IBM Redbooks® Solution Guide explains, from a high level, how to build enterprise search solutions with ICAwES.

The following figure shows how enterprise search solutions help you identify the hidden meaning of your content.



Windows file system 1/6/09

IBM Lotus Domino Express - Big email solutions for small and med
serversIntel processor-based serversAMD processor ... driversTrial
InfoPrintLenovo ThinkPads and ThinkCentresCustomer supportMo
maintenanceInvoices and paymentsSelf-service ... Why IBM for your
and licensesPopular tags by usersdomino Domino URL cheat she

Document ID: file:///c:/jane/IBM+Products/IBM+Products/Products/Lo
Title: IBM Lotus Domino Express - Big email solutions for small an
Language: en
Source: winfs
Type: text/plain
Categories: Products, Lotus

modifieddate: 1/6/09
filename: IBM Lotus Domino Express - Big email solutions for sma
title: IBM Lotus Domino Express - Big email solutions for small and
ibmproduct: IBM Lotus Domino Express IBM Sign IBM Lotus Notes
filesize: 8769
directory: c:\jane\IBM Products\IBM Products\Products\Lotus
date: 1/6/09
extension: .tx

Figure 1. Enterprise Search solutions help you identify the hidden meaning of your content

Note: For ease of reference, we use "enterprise search solutions" to refer to the enterprise search solutions that are built on top of ICAwES.

Did you know?

- Enterprise search solutions can automatically identify a wide range of date formats and provide a single interface for viewing, sorting, and retrieving documents by a document's creation or modification date, or any date that appears in your documents.
- You can set up your own synonym dictionary to expand search queries to include any number of word variations.
- You can write custom annotators to extract concepts and add them to your search.
- You can integrate different types of repositories in a single, integrated search.

Business value

Enterprise search solutions that are created with ICAwES add value performing the following tasks:

- Creating a single interface when there is a need to create logical connections over multiple repositories.
- Extracting discrete entities (such as personal names, telephone numbers, or addresses) from unstructured content. This is also important for linking non-structured content with structured data.
- Overcoming repository inconsistencies and lack of organization, such as disk drives with unplanned folder structures, and increasing accessibility to these resources.

Enterprise search solutions provide a unified interface to diverse structured and unstructured sources by creating a single index. Users perform searches in a fully integrated search environment. Such an approach reduces the investment in restructuring, cleaning, and maintaining multiple repositories, which might result in systems that no longer function optimally. Enterprise search can link fields and concepts from different sources to a single search field and filter inappropriate content and outdated records, all without any changes to the original repositories.

The ICAwES enterprise search capability can extract elements according to the type of data source (as shown below) and map the results to a single index:

- HTML metadata
- XML elements, according to the entity name, attribute name, and value
- Relative database fields (IBM DB2® and any database with a JDBC connection)
- IBM Content Manager attributes
- IBM FileNet® Content Manager (FileNet P8) properties

Enterprise search can also crawl Microsoft Exchange Server, IBM Case Manager, IBM Connection, IBM Quickr® for Domino®, and IBM WebSphere® Portal. Each content source has its own properties or metadata that can be added to the index.

Choosing which extracted elements to index, and how the index is used in search, gives you control over the solution functions.

ICAwES includes a component for writing custom annotators: IBM Content Analytics Studio (ICA Studio). Using ICA Studio, you can build custom annotators for extracting personal names, product IDs, serial numbers, and so on, from your free-text documents. ICA Studio is particularly useful in building custom annotators to extract domain-specific information.

Solution overview

An enterprise search solution is configured by using the Content Analytics Administration Console. There are three stages to setting up an enterprise search solution:

1. Crawler: Chooses which repositories to crawl.
2. Parse and Index: Chooses which crawled entities to map to index fields. Configures different annotation stages to extract facets and elements from the free-text document.
3. Search: Customizes the enterprise search experience by expanding users' queries by using synonyms existing in the content, through dictionaries and rules.

The enterprise search solution interface provides components for displaying different aspects of the search results. These components provide not only a meaningful overview of the content, but they are also interactive components for drilling down into the result set, refining the results according to the properties that are chosen for the user. Each component lists the result document count per element and allows you to add the elements to the drill-down search.

The search results are displayed with the following components:

- Facet Dialog: Provides a list of elements that are discovered by content analytics annotators.
- Category Tree: Displays categories that are defined by configurable rules.
- Dynamic Facet Chart: Displays date ranges.

Solution architecture

An enterprise search solution consists of two servers: an controller server and a search server. The controller server collects information from the crawled sources and builds an index for the search server. The search server carries out the actual search based on user queries.

The following data sources can be crawled and collected by the controller server:

- Content management sources
- Database systems
- IBM Case Manager sources
- Email, file system, and web sources
- WebSphere Portal sources
- IBM Connections sources
- IBM Lotus® Domino sources

The following figure shows the enterprise search solution architecture.

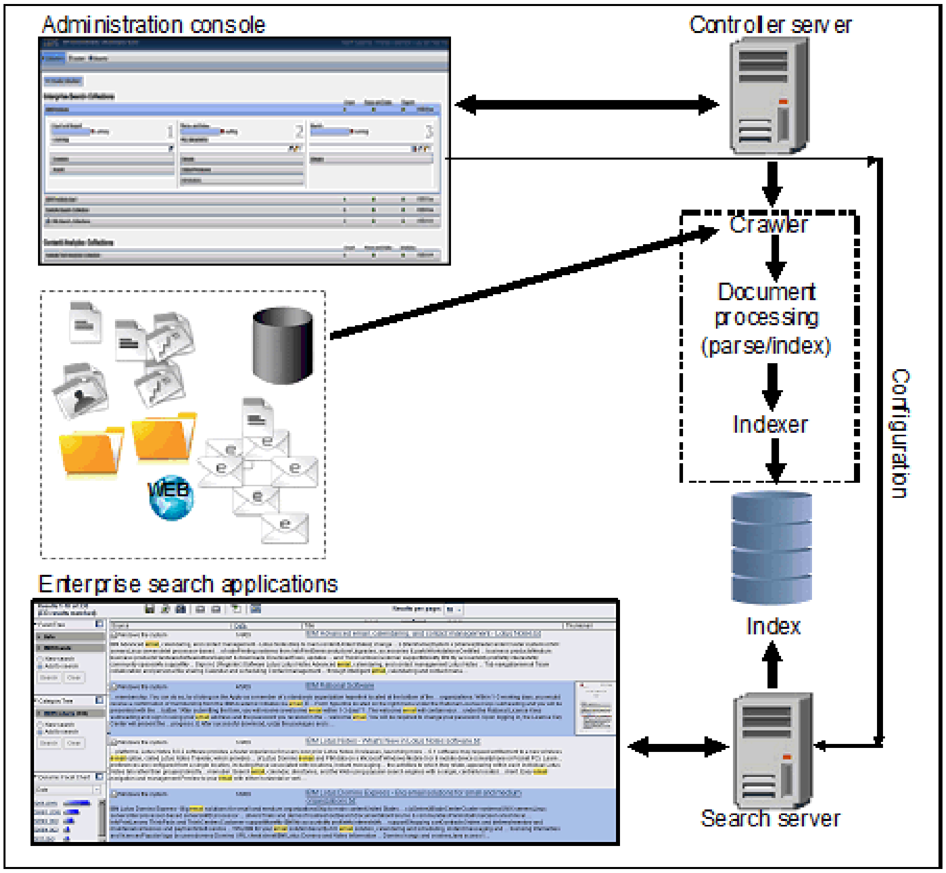


Figure 2. The enterprise search solution architecture

The administration console is responsible for configuring both servers. The search application can also define many search parameters, especially the query expansion and search results document ranking rules.

Usage scenarios

This section lists some possible usage scenarios for ICAwES.

Scenario 1: Searching across multiple content repositories and content repository types

In the first scenario, a financial institution needs an enterprise search solution that enables its workers to efficiently and easily perform search across the following items:

- Multiple repositories
- Different repository types
- Over geographically different locations

The ICAwES enterprise search capability enables the institution to create such a solution.

The financial institution used to have uniformed content management systems with only one type of content repository. Over the years, it has acquired many other financial institutions, and each institution has its own set of content and its own set of business rules and operations. To consolidate and unify business rules, operations, and content repositories among all these acquired institutions will take many years of planning, designing, and implementation. For now, the financial institution wants to keep the same operation for many of the acquired companies, yet be able to search content over these various content repositories. In addition, the financial institution is interested in searching over the external websites that contain a vast number of consumer comments and feedback that might help the company to gain business insights.

In this scenario, ICAwES enterprise search is used to integrate multiple repositories that, despite being physically separate and of different repository types, are linked according to the business need. ICAwES enterprise search allows the company to access and link the existing systems and newly acquired repository systems without needing to convert the data among the repositories. Some systems present special challenges, such as inconsistent data format, inconsistent ways of naming fields, and inconsistent ways of using words in unstructured content.

The ICAwES enterprise search can help solve problems of data inconsistency by placing constraints at both the indexing stage and at run time (search) and by providing additional access that was not available in existing systems. For example, when moving from one relational database design to another one, a single index field can access both sources, each with its own table and field design.

Scenario 2: Domain specific searching

In another scenario, a medical insurance company wants an efficient search of medical records for patient care and business analysis. The insurance company wants to capture relevant elements in the medical records during searches, including diseases, medications, medical procedures that are performed, the outcome results, and so on.

Using ICAwES enterprise search, the insurance company can build custom dictionaries, parsing rules, and create medical domain-specific custom annotators to identify relevant metadata, text, and extract their values. Such information can then be retrieved by search, along with the structured database query, for a particular patient.

Integration

ICAwES enterprise search provides a REST interface for creating customer applications or for integrating with existing applications.

You can also enhance enterprise search and content analytics solutions by integrating ICAwES with one of the following products:

- IBM Content Classification
- IBM Cognos® Business Intelligence
- IBM InfoSphere® BigInsights™ Enterprise Edition

The following data sources can be crawled and collected into an enterprise search solutions:

- Content management sources
- Database systems
- IBM Case Manager sources
- Email, file system, and web sources
- WebSphere Portal sources
- IBM Connections sources
- Lotus Domino sources

Supported Content management data sources include the following ones:

- IBM Content Manager
- EMC/Documentum
- IBM FileNet Content Manager
- Hummingbird DM
- Microsoft SharePoint
- Open Text Livelink Enterprise

For specific versions of the supported products or solutions for integration, see the system requirements link in "Supported platforms".

Supported platforms

ICAwES enterprise search supports multiple operating systems, including IBM AIX®, Linux, Linux on IBM System z®, and Windows.

For the latest information about supported operating systems, see the IBM Content Analytics with Enterprise Search Version 3.0 system requirements at the following website: http://www.ibm.com/support/docview.wss?uid=swg27023676

Ordering information

ICAwES enterprise search is available with the full purchase of the IBM Content Analytics with Enterprise Search Version 3.0 product.

Ordering information is show in the following table.

Table 1. Ordering part numbers and feature codes

Program name	Program number
IBM Content Analytics with Enterprise Search Version 3.0	5724-Z21

Related information

For more information, see the following documents:

IBM Content Analytics: Discovering Actionable Insight from Your Content, SG24-7877
http://www.redbooks.ibm.com/abstracts/sg247877.html

IBM Content Analytics with Enterprise Search product page
http://www.ibm.com/software/products/en/contentanalyticssearch

IBM Content Analytics with Enterprise Search Version 3.0 Information Center
http://pic.dhe.ibm.com/infocenter/analytic/v3r0m0/index.jsp

IBM Offering Information page (announcement letters and sales manuals):
http://www.ibm.com/common/ssi/index.wss?request_locale=en

On this page, enter "Content Analytics with Enterprise Search", select the information type, and then click **Search**. On the next page, narrow your search results by geography and language.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service. IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law : INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you. This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document was created or updated on March 28, 2014.

Send us your comments in one of the following ways:

- Use the online **Contact us** review form found at:
  ibm.com/redbooks
- Send your comments in an e-mail to:
  redbook@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

This document is available online at http://www.ibm.com/redbooks/abstracts/tips1147.html .

Trademarks