Semantic Technology Accelerates Document Search:

How LMI Implements Semantic Search with OpenPolicy™

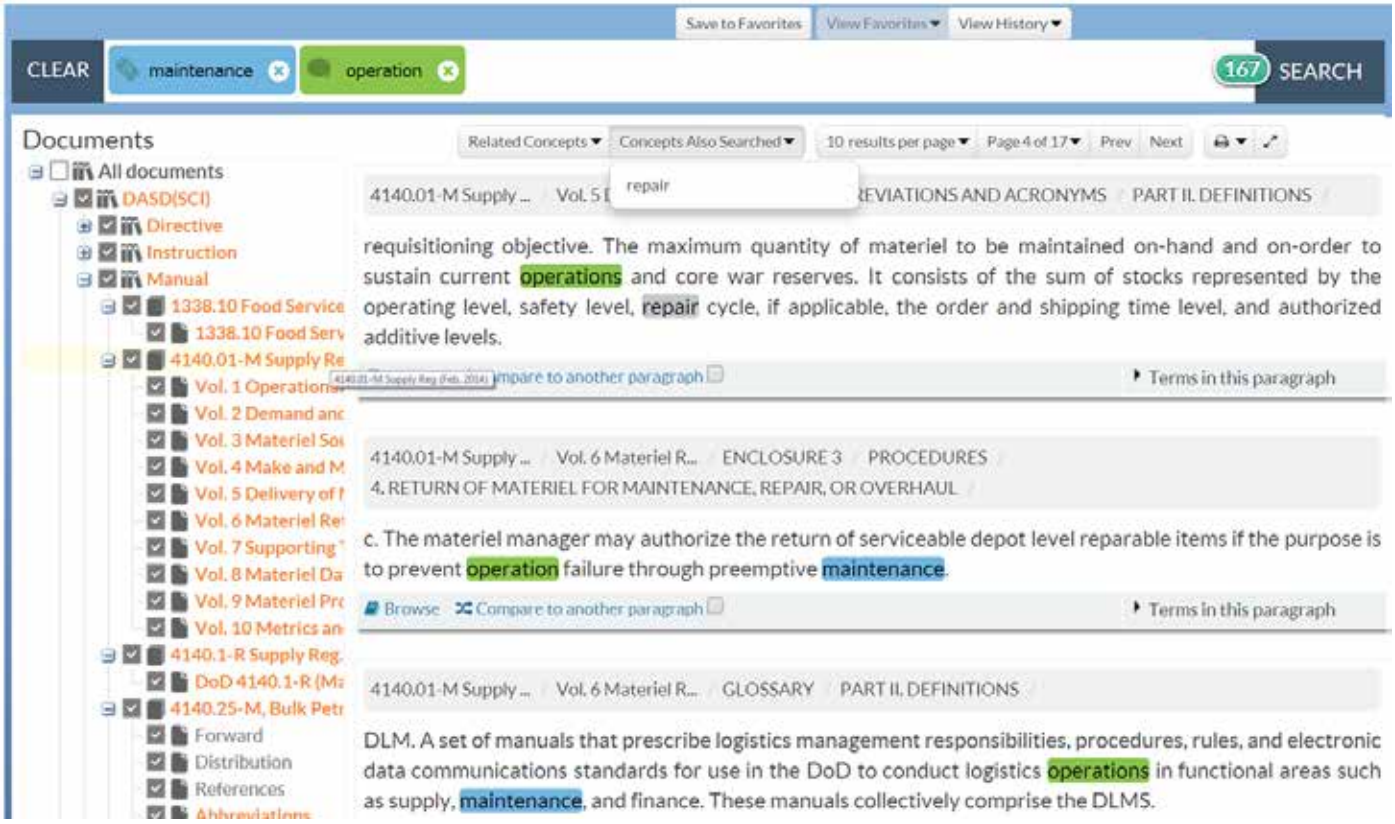
LMI

OpenPolicy™ Can Eliminate the "Find Next" Paradigm

Business and government enterprises have massive amounts of information and knowledge virtually locked away in electronic text files of written prose. Until now, researchers and analysts were required to fish through important online documents by downloading individual files, opening them, and using the "Find Next" search feature, looking for one key-word at a time. Imagine sifting through more than 20,000 pages of the U.S. tax code looking for key words!

LMI's OpenPolicy™ semantic search tool eliminates this time-consuming task. The tool can examine scores of documents simultaneously using subject-specific thesauri of key terms, synonyms, acronyms, and other related terms and phrases. As shown in Figure 1, this enhanced vocabulary-based semantic search pinpoints relevant paragraphs and highlights search terms inline. By avoiding the one-at-a-time document search paradigm, OpenPolicy can reduce the time spent searching across documents by more than 90 percent.

Figure 1 – Instant results for "maintenance/repair" and "operation" across 10,000 pages of defense policy



Why semantic search?

Users are able to pinpoint the exact paragraph they need out of tens of thousands of paragraphs.

Thousands of documents can be searched side-by-side.

A semantic database drives the search…They get highly relevant results.

Semantic search improves the productivity of the workforce.

A "Knowledgeable" Way to Search Documents

Combining the features of a highly intuitive user interface with web-based service components, OpenPolicy enables users to upload thesauri, upload documents, browse documents, and semantically search their content. Eliminating the "Find Next" button to page through mountains of individual documents lets users apply knowledge-based thesauri to instantaneously locate related content across large libraries. This approach offers huge benefits to an enterprise knowledge management practice:

- ***Captures knowledge of subject matter experts (SME) in a thesaurus.*** As the federal workforce is reduced, senior knowledge is lost. OpenPolicy captures knowledge in a digital thesaurus, which is reusable among all government customers. The thesaurus enables the search of keywords, their synonyms, and other variations.

- ***Enables search of large document libraries.*** Because relevant documents are grouped together in an OpenPolicy library, simultaneous search offers highly meaningful results. Users can compare multiple documents in a single search, which is especially useful for identifying conflicting policies or regulations.

- ***Dramatically reduces search time.*** OpenPolicy's use of a thesaurus takes the guesswork out of identifying keywords. And because it eliminates downloading and opening each document, and it lists search results as entire paragraphs, users cut out scores of keystrokes and mouse clicks from their search routine. Speedier searches magnify worker productivity.

Intelligent Thesaurus Speeds Searches

What makes OpenPolicy so fast and smart? The service relies on a combination of Worldwide Web Consortium (W3C) semantic web standards like the Resource Description Framework (RDF), Simple Knowledge Organization System (SKOS), and others combined with the industrial strength triple-store OWLIM™ database from Ontotext. Indexing the RDF triple-store database against a SKOS thesaurus delivers search results in seconds or less.

Use of these robust standards enables OpenPolicy to index content in an intelligent way. The tool's power lies in the use of a customized thesaurus that includes key terms and phrases about your subject area. The focused vocabulary in the thesaurus enables OpenPolicy to extract semantically relevant paragraphs from the content and present them in an easy-to-browse interface display. LMI creates this vocabulary for clients. We then use a multistep process to curate content. Combined, the thesaurus and content become the knowledgebase, the brains behind the highly targeted search.

- ***Build the Thesaurus Terms and Phrases.*** First, clients provide LMI their documents—policies, regulations, directives, training materials—any text-based file. LMI runs the entire library through a word/phrase analytics process. This process extracts numerous potential thesaurus terms from the library. A domain subject matter expert whittles this extraction down to a smaller collection of terms that are relevant to the subject area. Next, the subject matter expert use SKOS to create relationships among terms in the thesaurus. The expert creates synonym, acronym, and other relationships that OpenPolicy links together for simultaneous search (see Figure 2). These relationships enable OpenPolicy to build the intelligent index that guides the user to smarter results.

Figure 2 – LMI uses SKOS to create thesaurus relationships

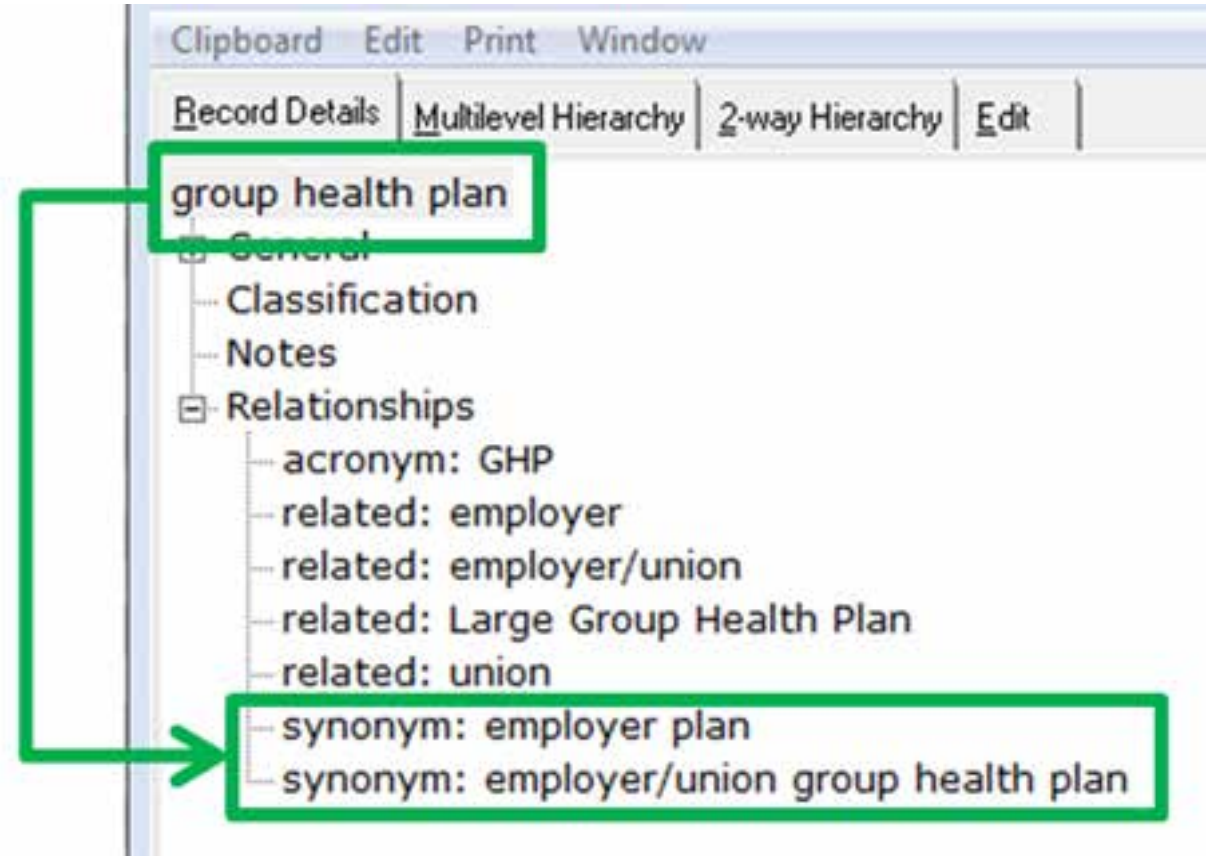
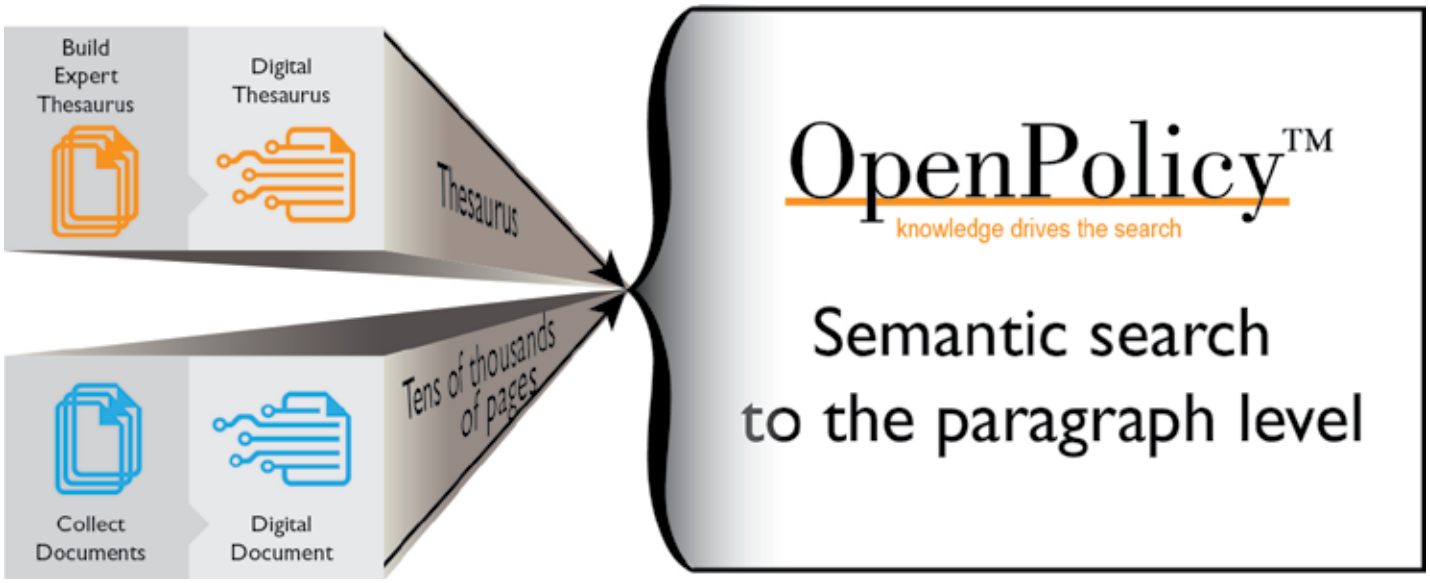


Figure 3 – LMI subject matter experts build the thesaurus, prepare the documents, and load them to the OpenPolicy service



- ***Prepare the Document Library.*** LMI then gathers documents into a content management system to prepare them for ingestion and track their progress. The domain expert examines a sample of the documents to determine the complexity and variety of their structure. Document curators convert files into one of three formats: Microsoft Word, XML, and XHTML. Both XML and XHTML are easily ingested. After preparing a document, the document curator applies up to 40 metadata fields to it. These data fields enable tuning of the OpenPolicy document navigation filters.

Launching the OpenPolicy Search Service

To launch OpenPolicy, LMI loads the thesaurus and document library to a prepared computing infrastructure and activates user accounts.

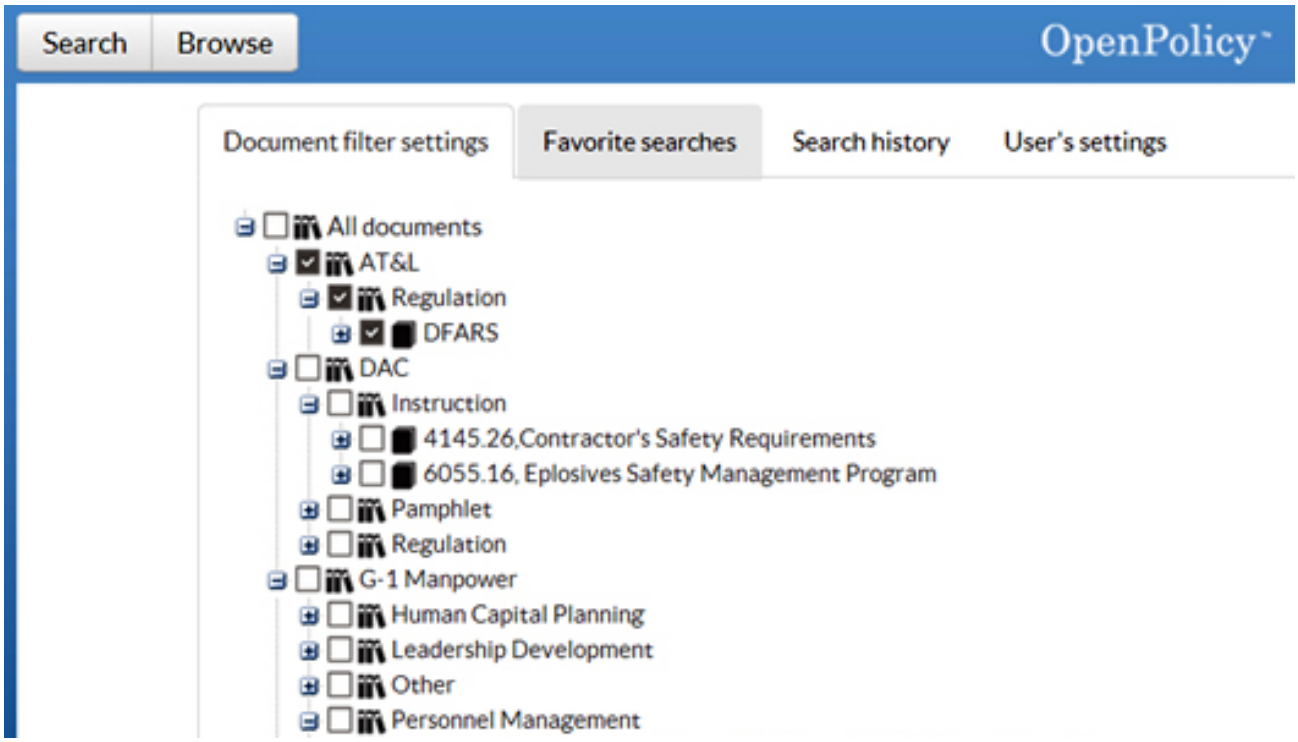
- ***Load a Thesaurus.*** An administrator points OpenPolicy at a thesaurus file and activates the upload feature. It is the knowledgebase for semantic search terms and phrases along with their synonyms and other relationships. OpenPolicy processes the thesaurus and, if a document library is already present, will automatically index the document library. Loading the thesaurus enables the search line drop-down menus that suggest expert tuned search terms and phrases.

- ***Load a Document Library.*** An administrator points OpenPolicy at a document library and activates the upload feature. OpenPolicy processes the document library files and, if a thesaurus is present, will automatically index the document library. The administrative features of the tool enable an authorized user to manage the documents, adding, reprocessing, and deleting them (see Figure 4).

Figure 4 – Onboard system administration

- ***Begin Using OpenPolicy.*** Once the thesaurus and document library are loaded, authorized users can begin their searches. Filters allow users to search all or just a portion of the library. The web interface also allows users to view recent searches, save favorite searches, and tailor user settings to best meet their needs (see Figure 5).

Figure 5 – Manage document filters, search history, and user



Sustaining OpenPolicy Keeps Your Knowledge Current

To sustain OpenPolicy, a team of functional experts maintains the thesaurus and document library, and engineers maintain the software and computing infrastructure.

- ***Keep Thesaurus and Documents Relevant.*** An administrator may add a new or different thesaurus at any time, enabling users to slice across documents using different subject vocabularies. A document administrator may also add new documents or delete outdated files to keep the library current. Documents may be added or deleted individually or in groups. OpenPolicy's use of W3C semantic web standards also means that administrators can port their documents and thesaurus from the OpenPolicy service to other semantic web applications, if desired.

- ***Maintain the Computing Infrastructure.*** OpenPolicy requires care and feeding by highly skilled LINUX® engineers. In addition to round-the-clock monitoring of infrastructure performance, the LINUX engineer needs to keep software components, such as OWLIM, Java™, and RedHat® versions, current.

You Don't Buy, You Subscribe

LMI offers OpenPolicy as a service. We can also scale for open web access by supporting a service level of concurrent users.

- ***By User Seats.*** Currently, LMI manages OpenPolicy subscriptions in blocks of 250 users. User accounts are keyed on user e-mail addresses. Once users have obtained authorization, they can manage their own passwords.

- ***Open to the Web.*** OpenPolicy can scale to accommodate hundreds to thousands of concurrent users. No usernames and passwords are required. Sessions are controlled by a time-out feature and by closing browsers.

OpenPolicy: Under the Hood—*Cloud-Based Hosting Ensures Security and Scalability*

OpenPolicy is composed of several software components. These run in a secure cloud server environment, which also houses the document libraries and related thesauri.

Software. The OpenPolicy™ user interface is a lightweight JavaScript-based webpage that uses AJAX and JSON to enable the browser to exchange data with back-end services. The back-end service components consist of a framework of application code, a concept extraction service, a NoSQL content store, and Ontotext's commercial OWLIM™ semantic Resource Description Framework (RDF) triple-store database. All services run in a RedHat® LINUX® environment. The components run in independent webserver containers that interact via application programming interface (API) connectors. The system is secured behind a light-weight directory access protocol (LDAP)–enabled, single sign-on authentication system.

Computing Infrastructure. OpenPolicy currently runs in the LMI cloud infrastructure. The service runs in a virtual environment on a minimum 8-core server with 16 GB of memory and a large network attached storage array. It is further protected by firewalls, secure sockets layer (SSL) encryption, and load balancers. OpenPolicy can be deployed in a third-party cloud.

Scalability. The OpenPolicy architecture enables a horizontally scalable capability. As a result, it can load, store, and search billions of records shared among two or more physical servers.

Contact LMI for Additional Details

LMI's OpenPolicy semantic search tool is a fantastically powerful way to research written prose documents. This flexible tool enables customers to efficiently search massive knowledge storehouses at hyper speeds. Its use of focused vocabularies speeds queries and produces more meaningful results. Additional benefits include capturing SME knowledge in a reusable thesaurus, facilitating simultaneous document searches, and magnifying worker productivity.

Locations

Headquarters
Washington, DC

Regional Offices
Huntsville, AL
St. Louis, MO
San Antonio, TX

Satellite Offices
Aberdeen, MD
Baltimore, MD
Mechanicsburg, PA
Newport News, VA
Petersburg, VA

Contact Us

Gus Creedon,
Program Manager
703.917.7272
gcreedon@lmi.org

2000 Corporate Ridge
McLean, VA 22102-7805
703.917.9800

LMI.org @LMI_org Follow us on LinkedIn

LMI

Complex Problems. Practical Solutions.

www.lmi.org