The Center for Cyber Defenders

Expanding Computer Security Knowledge

Federated Search

Aaron Easter and Thomas Reese,
Missouri University of Science and Technology

SAND2012-5867C

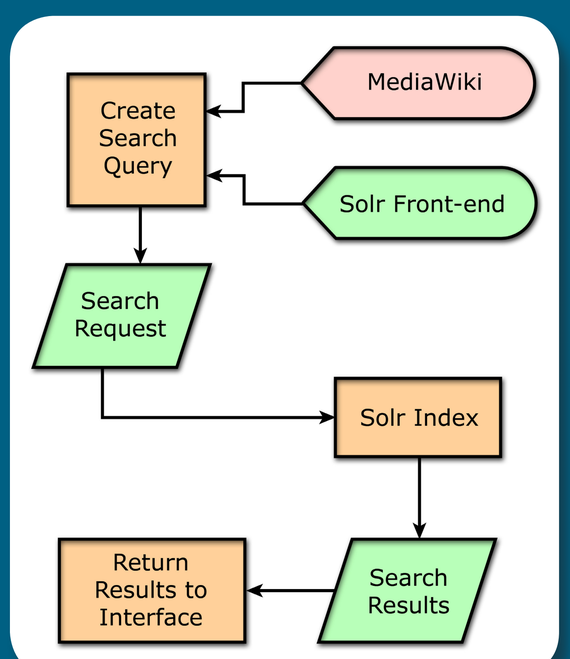
Project Mentor: Timothy Eriksson, 5563

Problem Statement:

Sandia maintains large quantities of documentation. However, these documents can be stored different file systems, various wikis, Confluence, or SharePoint. This can make it very difficult to locate a the desired documents, especially as team members move to different projects.

Objective:

One solution to this problem would be to implement a federated search utility. Federated search is a technology for retrieving information from multiple search locations. Rather than randomly searching for a document through a hierarchy, one can enter a title or keyword for a document and locate it without any knowledge of its location. The additional ability to conduct a faceted search through specific fields, such as author, date created/edited, content, and file type, as well as the ability to return highlighted text is also desirable.



Impact and Benefits:

This setup facilitates finding documents when no location data is known about it. It also provides the capability of a project to use a wiki for storing project information, and have the ability to search for any documents or other wiki pages in the same context.

Approach:

- Solr – Open source enterprise search platform from Apache. Indexes documents given to it, then can be queried to find those documents.
- Nutch – File and web crawler used to explore specified hierarchies to find new or updated files and given them to Solr for indexing.
- Tika – File parser used by Solr to extract content and metadata from many different files (including docx, xlsx, pdf, zip, etc.) for indexing.
- MediaWiki – Platform used for recording progress, as well as proof-of-concept integration with Solr.

Results:

- Solr was used to index several different types of documents, confirming its ability to search their content and metadata.
- MediaWiki was used to create an interface to Solr. The SolrStore extension was modified to bypass the default search engine, and query Solr instead, allowing the user to search for any document from inside MediaWiki.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND2012-XXXXP.