
The Enterprise Search and *Toros Celebres*

Large companies are embracing Lucene/Solr (search), Hadoop (distributed processing), Alfresco (content management), and Linux (operating system). Each of these products is open source. “Open source” means that no one company owns the software. Instead of a single firm controlling innovations and fixes, volunteers provide the technical heavy lifting.

Lucene is open source software code that one can use to build programs that perform key word indexing. Lucene delivers the type of search provided if a user entered key words into Bing.com or an Autonomy system. Solr takes the Lucene software code and puts it in a server, ready to run. Lucene/Solr provide an alternative to the traditional systems provided by dozens of commercial software vendors, just without the limitations and costs of the proprietary solutions. According to David Fishman, vice president, Lucid Imagination, an open source search vendor in Silicon Valley, “Lucene and Solr deliver the same functionality that one would expect to get from a commercial enterprise search vendor’s system.”

The rising interest in open source software is due to several factors. The need to reduce costs is one factor. Open source search software, for example, costs nothing to download. There are costs associated with installing, upgrading, and customizing the system. But with commercial search software licenses there are annual fees. Eliminating these delivers an immediate cost saving. Many organizations are interested in a Web 2.0 approach to information. Standards and interoperability are essential to the newer approaches to delivering systems that “play well with others”, as the child psychologists say. Another key factor is the desire of many organizations to take control of their own destiny, not leaving it in the hands of a vendor of proprietary software who wants to limit a customer’s freedom. Vendors want to create lock in to help ensure that a customer remains a customer, often for many years.

Most executives remember the Hemingway classic *Death in the Afternoon*. One statement may apply to the celebration now unfolding between search vendors who offer proprietary enterprise search systems and a new breed of search vendor. The “new breed” refers to providing open source search software without a fee. The vendors’ business model depends on those using the open source search system paying for training, technical support, and engineering services. As Hemingway said about another conflict: “I remember that after we searched quite thoroughly for the complete dead we collected fragments.”

Can open source search break the grip proprietary search solutions have on the enterprise market? Will Lucene/Solr—two popular open source search systems—turn giants like Autonomy Ltd. (Cambridge, UK), Endeca (Cambridge, Massachusetts), Google (Mountain View, California), and dozens of other search-and-retrieval companies into a *filete a la parrilla*? Will open source search leave some of those involved in the rough-and-tumble enterprise information retrieval market broken or dead?

Lucid Imagination, a Silicon Valley start up, has received venture backing totaling \$16 million. The company's business model is to make available a build of Lucene/Solr that is ready to install. The cost? Nothing. Lucid's business model generates revenue from services. The company offers training, runs conferences, and provides engineering and technical support. According to the firm's vice president of marketing David Fishman, "Lucid is growing. Open source search is becoming increasingly disruptive. Adopters range from companies like Macy's to high-profile Web services like Salesforce.com."

I don't know whether proprietary search, open source search, or blended search will emerge victorious. I do know that a digital bull fight is underway. The unceasing increase in digital information exacerbates "findability" for employees, management, and stakeholders. In the hot summer morning before the economic downturn, some search vendors promised government agencies, commercial enterprises, and not-for-profit entities systems that would make an organization's electronic information findable. Adjectives like "fast," "reliable," "scalable", and "stable" were used to make enterprise search appear a "no brainer", a phrase one vendor used as recently as April 2010.

As the sun sets over many organizations, the shadows are lengthening for many proprietary search and retrieval systems. Customers now know that making in-house information findable is a difficult undertaking. Dealing with different types of information poses a number of obstacles. These range from connecting the search system to other enterprise systems, ensuring that the information in the search system is timely, and the results are useful to those who are looking for a specific item. To add difficulty to the obstacle course are such novel content types as short text messages, audio and video data, and archived information from legacy systems. Most organizations find that the information target is one that is moving, often in unpredictable ways. The "open" in open source means that an organization using an open source search system can, with some caveats, customize the application, write extensions, and tailor the functions to meet the specific needs of an organization. A proprietary system places constraints on what the licensee can do. One reason open source search may be getting hard looks from the enterprise is the ability to make changes as the need arises. Instead of modifying a licensing agreement or buying additional functionality, the organization can make the change. Flexibility and agility may vie with free as factors influencing the decision to embrace open source search solutions.

Content outside the organization is becoming more important to many enterprises. Social content flowing through the public Internet and specialized services like Facebook and Twitter may have a direct, immediate impact on a commercial enterprise. Southwest Airlines learned about social media when a passenger was forced off a flight because of his physical size. Via Twitter, the aggrieved passenger ignited a firestorm. Most commercial search vendors do little to counter the growing body of research data that suggests user satisfaction with enterprise search systems is low. Based on the research Martin White and I conducted for *Successful Enterprise Search Management* (Galatea, 2010) revealed that as many as two-thirds of the users of a search system are dissatisfied with that system. Not surprisingly, enterprise search systems have to be rejigged to handle certain specialized tasks such as indexing and making findable social content. Ideally, the search systems will function more like early warning systems than old-fashioned, library research systems.

Keep in mind that commercial software has some advantages. When a problem crops up, the vendor often fixes the problem. The licensee may not be allowed to use the software in a

certain way or be permitted to make some types of modifications. The constraints may make the commercial software appear more stable. But these advantages come with a price. There is a license fee and often maintenance and engineering fees. Nevertheless some organizations have a policy that stipulates that commercial software must be used.

The downside of commercial software is that some vendors do not update the product quickly. Certain enterprise search systems perform some functions particularly well. Other functions such as indexing social content or handling speech-to-text functions for telephone calls or podcasts. Some search and retrieval systems may be inflexible, slow to add such features as entity extraction, automated report generation, or video indexing. As a result, some of the organizations Martin White and I interviewed for our 2010 study may buy specialized systems. We found that it is a rare organization that has a single enterprise search systems. Most had five or more. There were systems to search legal information. Systems to search databases. Systems to search Web content. Systems to search data in the research lab. And systems to search email. Employees had to look multiple places for information and then stitch that information together by hand, often on deadline. Not surprisingly, information was often stale, incomplete, or fragmented. No wonder some organizations cannot seize opportunities. In today's world, if one cannot locate needed digital information, work becomes difficult if not impossible. Even worse, a decision made based on incomplete information may be little better than allowing decision makers to just guess.

What are the benefits of using community-supported, free search software? According to Charlie Hull, Managing Director of FLAX, an open source search vendor in the UK, "Freedom is the key one - freedom to choose how your search project is built, how it works and its future. Flexibility is important, as every project will need some amount of customization. The lack of ongoing license fees is an important economic consideration, although open source shouldn't be seen as a 'cheap and basic' solution - these are solid, scalable and high performing technologies based on decades of experience. They're mature and ready for action as well - we have implemented complete search solutions for our customers, scaling to millions of documents, in a matter of days." Open source search solutions are available without a fee and are not closed systems like many proprietary commercial information retrieval products.

Among the most visible open source search systems are Lucene and Solr. Both are open source products and freely available from repositories such as SourceForge.net or from commercial firms specializing in open source search. You can download a certified package of Lucene/Solr from Lucid Imagination, a start up based in Silicon Valley. Lucid has received \$16 million in venture funding. The firm makes Lucene/Solr software available e

IBM may be an unlikely adherent of open source search technology, but IBM is not alone. Cisco Systems, the US network giant, is embracing open source for its search applications. These companies have joined LinkedIn.com, eHarmony.com, Netflix.com, and hundreds of other organizations to rely on open source search software, not mainstream commercial search and content processing solutions. Cisco revealed in an interview appearing in my Web log Beyond Search in August 2010 that the company has both open source and commercial search solutions. The caution is admirable, but one cannot overlook the fact that three years ago, open source search was a fringe solution.

Open source refers to software that is available without charge under one of an alphabet soup of license agreements. The code is available to anyone from public repositories such as SourceForge.org or from commercial companies who package together essential components and offer a bundled download. Either way, the search and content processing software is available without charge. Fixes and enhancements are provided by what open source specialists call “the community.” The idea is that those using the software will fix problems and make available additional software shims and widgets to improve the product.

Fishman added, “You can navigate to Lucid Imagination’s Web site and download Lucene/Solr. Without a couple of hours, you can have the system up and running.” Lucid Imagination provides a comprehensive build of Lucene/Solr so, added Fishman, “you can have a commercial experience without the commercial pain.”

Lucene/Solr offers a number of advanced features, including assisted navigation. Fishman noted, “The commercial vendors have started indexing a document or other content object with category tags. If you run a query on a site like Borders.com, you will see a number of links that provide one-click access to related information. Solr performs the same type of indexing just without the commercial vendors’ license fees.”

Lucid Imagination is one of several open source search vendors. Tesuji.net in Budapest and Lemur Consulting in Cambridge, England, also offer open source search solutions. The business model for these companies is to provide the open source search system as a free download. The company’s engineers provide technical support and training on a fee basis.

FLAX’s Hull said: “Integrating search engines is what we do at Flax day-to-day. Since we’ve chosen highly flexible and adaptable open source technology, we can do this in a fraction of the time and cost. We don’t dictate to our customers how their systems will have to adapt to our search solution. We make our technology work for them. Whatever platform, programming language or framework you’re using, we can work with it.”

The open source search systems may be poised to disrupt the commercial enterprise search vendors’ business model. Lucid’s David Fishman said: “We are coming at search from a different angle. We like to say that we have a 90 degree play because we think that building on open source software is the right angle. There is also a sure of interest in open source software, and not just search. Google and IBM have been touting open source technology. But RedHat showed that a hundred million dollar business could be built on open source code. And open source software, in our opinion, is just better. There’s no lock in, no delays for fixes, and literally hundreds of developers creating and extending the system every day.”

According to the Otis Gospodnetic, author of Lucene in Action, “I get the feeling they [organizations] like the open-source/commercial blend. Plus, there is precedent – commercial support for open-source software has been around for many years now: MySQL, Red Hat have been doing this for years. Not only is this not confusing, it is welcomed.”

Stephen E Arnold, September 5, 2010

Stephen E Arnold is a consultant. His Web site is www.arnoldit.com. His blog “Beyond Search” is at <http://www.arnoldit.com/wordpress>.