Why you should be considering open source search

You could already be using an enterprise search engine. There are many available, from large and small vendors, and all of them promise to help your users wade through the digital morass of information.

Maybe you're a user of the FAST ESP product, recognised to be a very powerful and flexible solution and used by many large, global companies. You will have noted how Microsoft bought FAST back in 2008, and then announced[1] how it would stop development on any platform other than Windows – a worrying sign for anyone with a FAST installation. The entire product has been rewritten in .Net and added to the Sharepoint platform – great news for those already on Sharepoint (assuming you can afford it[2]) but not necessarily for those unwilling to adopt it. Luckily there are suitable migration routes to open source search engines[3], and with open source you won't suffer from vendor decisions or the vagaries of acquisitions and mergers – you will always have access to the code that runs your search.

You could also be in the position of having to scale further your existing commercial product. License fees for closed-source search technologies are often based on the number of documents to be searched, number of collections of these documents or the number of users. Since (as we all know) digital information only ever gets larger, you may run up against either a limitation of the technology itself or into a entirely new (and unaffordable) realm of pricing – some license fees can run into six or seven figures. Of course, enterprise search has always been expensive (in our view, for no good reason – after all, the basic ideas – inverted indexes, relevancy ranking – have been around since the 1970s[4]). It's easy to sell search as a universal panacea to the ongoing problem of employees or customers wasting time and money searching for information, and many companies have exploited this fear of waste in their marketing.

Up until a few years ago there wasn't any serious alternative to the commercial search vendors – but various open source projects have since 'come of age' and have also received substantial commercial backing. The most well known software library is Apache Lucene[5] but alternatives such as Xapian[6] and Sphinx[7] are also worth considering. The missing links for most open source projects are support on a predictable basis and a number of vendors provide this including the US-based Lucid Imagination[8], who recently recieved $10m in funding to build solid commercial backing including training, documentation, custom applications and support on a SLA basis *(disclaimer: we're one of their UK partners).* Lucid also provide a complete search platform built on Lucene/Solr, LucidWorks Enterprise[9], which offers features previously only available in closed-source software such as automatic translation of a long list of document formats and a detailed management interface.

The projects driven by these open source engines are impressive – Twitter supports a billion queries a day using Lucene[10] and IBM's Watson system which recently competed on the TV quiz Jeopardy also contained a Lucene-powered searchable index. Xapian powers search for the UK's Newspaper Licensing Agency[11] (currently at 20 million documents from 140 newspapers) and the clippings service from the Financial Times[12]. Solr, a 'search server' built on Lucene is used by the Hathi Trust to search over 1.7 trillion words in 1.2 billion pages[13]. Since not everyone using open source search will go public about it you can be sure there many more examples – some commercial search vendors even use open source code as the foundation for their products, licence permitting.

It's important not to consider search technology just as a 'bolt-on' to an existing website, intranet or application, but to realise that it can be the foundation of new developments. The Guardian newspaper, at the cutting edge of the new 'digital journalism', use an Oracle database to store their articles but power their Open Platform API[14] with Lucene/Solr. The search engine effectively flattens the cost curve of complex queries on the database. Durrants, the UK's leading media monitoring company, uses Xapian to power thousands of complex searches on every single news article they monitor[15] - over 1.5 million a month. These applications and others like them are far removed from the traditional search box.

In these rapidly changing times we don't know what we will need to search tomorrow – so it's important to be adaptable, flexible and able to cope with data volumes that may not scale linearly. Maintaining control over the future of your search software is also key. Open source search has come of age and every modern business should be aware of its advantages.

Article written by Charlie Hull
Managing Director, Flax
2011

About Flax

Flax is highly active in the information retrieval market with international clients from sectors including academia, public relations, e-commerce, government and private businesses. Flax's clients include The Financial Times, Accenture, the Newspaper Licensing Authority (NLA), The University of Cambridge and Mydeco. Flax delivers a cutting-edge enterprise search solution, using the power of open source software to drive down costs and provide world beating search performance with no software licence fees. Flax is an authorized partner of Lucid Imagination, the commercial company behind Lucene & Solr.

[1] http://www.channelregister.co.uk/2010/02/08/fast_microsoft_linux_unix/
[2] http://www.realstorygroup.com/Blog/1969-SharePoints-FAST-2010-is-not-as-cheap-as-you-might-think
[3] http://sesat.no/moving-from-fast-to-solr-review.html
[4] http://en.wikipedia.org/wiki/Tf%E2%80%93idf
[5] http://lucene.apache.org/
[6] http://www.xapian.org/
[7] http://sphinxsearch.com/
[8] http://www.lucidimagination.com/
[9] http://www.lucidimagination.com/enterprise-search-solutions
[10] http://engineering.twitter.com/2010/10/twitters-new-search-architecture.html
[11] http://www.nla.co.uk/default.aspx?tabid=43
[12] http://clippings.ft.com/
[13] http://www.hathitrust.org/large_scale_search
[14] http://www.guardian.co.uk/open-platform
[15] http://www.flax.co.uk/our_clients#durrants