

**SOLR**

*Your one stop search shop*

**Chris Rimmer - We7**

Search

**NETFLIX**

**digg**

**github**  
SOCIAL CODING

Etsy

**theguardian**



**“But I'm only  
building a  
small system”**

**"But I don't  
want to write  
any Java"**



103

YAD

question  
everything







For further info:

<http://lucene.apache.org/solr/>

<http://wiki.apache.org/solr/>

Contact me:

[chrisr@we7.com](mailto:chrisr@we7.com)

[@nespera](#)



Credits:

Shop Photo by Eric P: [http://www.flickr.com/photos/ep\\_jhu](http://www.flickr.com/photos/ep_jhu)

Teapot Photo by Matthew Burling: <http://www.flickr.com/photos/bonehead>

Graffiti Photo by Duncan C: <http://www.flickr.com/photos/duncan>

Cupcakes Photo by Clare and Dave: <http://www.flickr.com/photos/clarebatemanking>



I work for We7. We are a music streaming company here in Oxford. We're a bit like Last FM or that company whose name rhymes with "Swat a fly".

I'm here to talk about....

The word "Search" is displayed in a large, sans-serif font. Each letter is a different color: 'S' is blue, 'e' is red, 'a' is yellow, 'r' is blue, 'c' is green, and 'h' is red.

Search.

It's really important.

So important that I hear rumours that people have built a billion dollar business on the back of it.

Folks like Google and Amazon do a really good job and that's one of the reasons they've done so well.

Unfortunately, users expect all sites to be that good. Damn! They're making the rest of us look bad.

So what to do?



Well Apache has a Java-based project called Solr that powers search on lots of big, high-profile sites like Netflix, The Guardian, TicketMaster, Digg, Etsy, even WhiteHouse.gov....

Could that help us to keep up with the big boys?

You might say...

**“But I'm only  
building a  
small system”**

You might think that something used to build these big sites would be overkill.

It's true that Solr will scale up to many tens of boxes. But to get going all you need is a servlet container like Jetty or Tomcat.

Install Jetty, copy Solr into place, set a property to point it to a config file (examples are in the download), start Jetty and you are done.

So it's simple, but you still might say...

**“But I don't  
want to write  
any Java”**

“I'm a PHP/.Net/Python programmer. I don't want to write any Java”.

You don't need to. At We7 we used it to reimplement our music search and even though our site is mostly coded in Java we didn't write any custom code in Java on the Solr side at all.

We configured it how we wanted and then only accessed it over HTTP. So it could have been written in COBOL for all it mattered. Most major languages have client libraries.

“OK, you've convinced me. So how do I use it?”

Right, here are the practical steps...



Assuming you have Solr running on a box somewhere, first you need to fill you index with data.

This can be done using a REST-like API by POSTing XML, CSV or JSON to it.

You can also get Solr itself to load data directly from a database. This can be set up to do an incremental load.

There are ways to load other documents like PDFs, Word Documents etc

Obviously you'll need to incrementally do this going forward, assuming your data is not fixed.



Now it's time to query your index.  
Again, it's very simple, you just issue HTTP GETs.  
It'll return data in XML, CSV, JSON or whatever

It has query handlers that will take user input with  
very little pre-processing.

Did I mention that it is very quick? We do search-as-  
you-type over the meta data of our 10 million tracks,  
so we need sub-second response times.

“So what about all those fancy bells and whistles that  
Google have?”





Solr has lots of goodies built in:

Synonyms, i.e. 6 vs six

Phonetic Matching, i.e. Socks vs Sox

Stemming, i.e. Walks vs Walking vs Walked

Spell checking i.e. "Did you mean?"

Query auto-completion

Result Highlighting

Result Faceting i.e. Splitting results into categories

Spacial Searching

....

Also lots of caching, plus replication, sharding and so on, so that when your site goes massive you can keep on scaling up.

For further info:

<http://lucene.apache.org/solr/>

<http://wiki.apache.org/solr/>

Contact me:

[chrisr@we7.com](mailto:chrisr@we7.com)

[@nespera](#)



Credits:

Shop Photo by Eric P: [http://www.flickr.com/photos/ep\\_jhu](http://www.flickr.com/photos/ep_jhu)

Teapot Photo by Matthew Burling: <http://www.flickr.com/photos/bonehead>

Graffiti Photo by Duncan C: <http://www.flickr.com/photos/duncan>

Cupcakes Photo by Clare and Dave: <http://www.flickr.com/photos/clarebatemanking>

So in summary, if you're doing search, I suggest you take a look at Solr