

## Q&A Session for Improve your PHP Application's Search Capabilities with Lucene

Date: November 7, 2007

---

Zach Conrad - 9:08 am

Q: Will Zend\_Search\_Lucene search in multiple languages? Does Zend\_Search\_Lucene support partial word matches?

Alexander Veremyev - 9:29 am

A: Yes. Lucene stores info in UTF-8 internally. You must specify input encoding correctly while indexing and searching. Some details can be found here:  
<http://framework.zend.com/manual/en/zend.search.lucene.best-practice.html#zend.search.lucene.best-practice.encoding>

Alexander Veremyev - 9:33 am

A: But you may need a custom text analyzer for this  
([http://framework.zend.com/manual/en/zend.search.lucene.charset.html#zend.search.lucene.charset.utf\\_analyzer](http://framework.zend.com/manual/en/zend.search.lucene.charset.html#zend.search.lucene.charset.utf_analyzer)).

Alexander Veremyev - 9:34 am

A: Partial word matches are available with wildcard queries  
(<http://framework.zend.com/manual/en/zend.search.lucene.query-language.html#zend.search.lucene.query-language.wildcard>) or may be done at the analyzer level (Stemming, for example)

---

Zach Conrad - 9:23 am

Q: Does Zend\_Search\_Lucene index words stored in flash files? (These are often included in our scripts...what is the best practice for that?)

Alexander Veremyev - 9:40 am

A: Zend\_Search\_Lucene doesn't have support of any special file format (except HTML). So it's your program's responsibility to parse data (ex. Flash or MS Word documents) and prepare documents (Zend\_Search\_Lucene\_Document objects) for indexing.

---

Zach Conrad - 9:29 am

Q: Does the getLinks() method work with document root and site root URLs? (../something.html vs. /dir/something.html)

Alexander Veremyev - 9:45 am

A: URLs are returned "as is". The HTML parsing implementation is based on the loadHTML() method of the DOMDocument class (<http://www.php.net/manual/en/function.dom-domdocument-loadhtml.php>) and has the same behavior.

---

Geraint Howell - 9:32 am

Q: Is it possible (or easy) to index a document ignoring any content that is not within a specified div element?

Alexander Veremyev - 9:54 am

A: I would recommend that you take a look at `Zend_Search_Lucene_Document_Html` class as an example. The `DOMDocument->loadHTML()` method transforms the given HTML to a DOM document and it's easily searchable using XPath, so any special `<div>` element can be very easily retrieved.

---

Zach Conrad - 9:42 am

Q: Will Lucene search subdirectories automatically from the `START_URI`?

The `START_URI` was defined as part of our crawling example. You can implement a crawler starting from any page and following any links according to your needs.

---

Zach Conrad - 9:43 am

Q: is there a way to limit queries by roles/privileges?

Alexander Veremyev - 10:01 am

A: The Lucene index file format itself doesn't support privileges, but this can be emulated with an additional field for privileges info. These limitations can be applied at search time by mixing the user's query with a result set limitation subquery (through the Query API).

---

Zach Conrad - 9:47 am

Q: multi-term queries would be done like "+force strong -dark" ?

Alexander Veremyev - 10:02 am

A: Yes. Exactly.

---

Zach Conrad - 9:54 am

Q: are results ordered by score by default?

Alexander Veremyev - 10:03 am

A: Yes.

---

Vlad Fratila - 9:54 am

Q: Would pagination (something similar to the mysql Limit syntax) be faster on a large result set?

Alexander Veremyev - 10:05 am

A: Yes. You can see the details here:<http://framework.zend.com/manual/en/zend.search.lucene.searching.html#zend.search.lucene.searching.results-limiting>

.

---

Brendon Kozlowski - 10:02 am

Q: Will these slides be available for download?

A: Yes. <http://www.zend.com/webinar>.

---

Zach Conrad - 10:09 am

Q: Is there anything else 3rd party or otherwise, that allows for searching PDF, doc, of?

A: You simply need to extract the relevant text from these documents. There are many such extractors available implemented in different languages; one example of PDF text-extraction library is <http://www.pdfbox.org/> .