



eZ Find

Integrating Java and PHP to create a
powerful search application

Kåre Køhler Høvik (kk@ez.no)

Project manager @ eZ Systems

Background

- Provider of eZ Publish
- Existing search in eZ Publish s*ck
- \$\$\$ 3rd party solutions



Goals

- A state of the art search engine for eZ Publish and its environment
- Enhanced navigation, employing cutting edge information retrieval technologies
- A sustainable platform for future search applications
 - eZ Components
 - Stand-alone applications



Requirements

- User permissions
- Content object mapping
- Language priority
- Multiple eZ Publish installations
- Easy to use
- Easy to install
- PHP 4 + PHP 5



Create a new search engine

- The good
 - Perfect functionality
- The bad
 - Long time to market
 - Large investment
 - Uncertain result
- The ugly
 - Costs



Using existing engine

- The good
 - Low costs
 - Proven functionality
 - Short time to market
- The bad
 - Rigid functionality
- The ugly
 - Licensing



Choosing an engine

- Xapian
 - Small developer base
 - Unknown in the market
- Lucene
 - Large community
 - Proven functionality
- Solr
 - Lucene on steroids (in a good way)



Solr, out of the box



- Tunable for highest relevancy, adapts to structured (eZ Publish) and unstructured content (files, external web pages)
- Advanced searching and navigation tools
 - Facets (drill down interfaces, meta-data)
 - Highlighting
 - Automatic related content (heuristics based)



Implementation options

- Pure PHP based solution
- PHP-Java bridge
- Web service based



PHP based solution - Issues

- Performance
- Scalability
- Maturity
- System requirements
- Incomplete port (missing functionality)



PHP based solution - Implementation

- Using Zend Framework Search
 - Rewrite of Lucene search extension

```
$this->indexDir = new Zend_Search_Lucene_Storage_Directory_Fileystem(  
    $indexDirPath );
```

```
$this->searchLogDir = new Zend_Search_Lucene_Storage_Directory_Fileystem(  
    $searchLogDirPath );
```

```
$this->analyzer = new Zend_Search_Lucene_Analysis_Analyzer_Common_Utf8();
```



Java bridge - Issues

- Stability
- System requirements
 - <http://php-java-bridge.sourceforge.net>
- Performance
- Locking issues
- Licensing (Apache / GPL)



Java bridge - implementation

- Started as community extension
- Example code

```
$this->indexDir = new Java( 'java.io.File', $indexDirPath );  
  
$this->searchLogDir = new Java( 'java.io.File', $searchLogDirPath );  
  
$this->analyzer = new Java(  
    'org.apache.lucene.analysis.standard.StandardAnalyzer' );
```



Java bridge - Example site

- <http://nifab.no>
 - National centre for alternative treatment
 - Uses dl for loading java bridge module



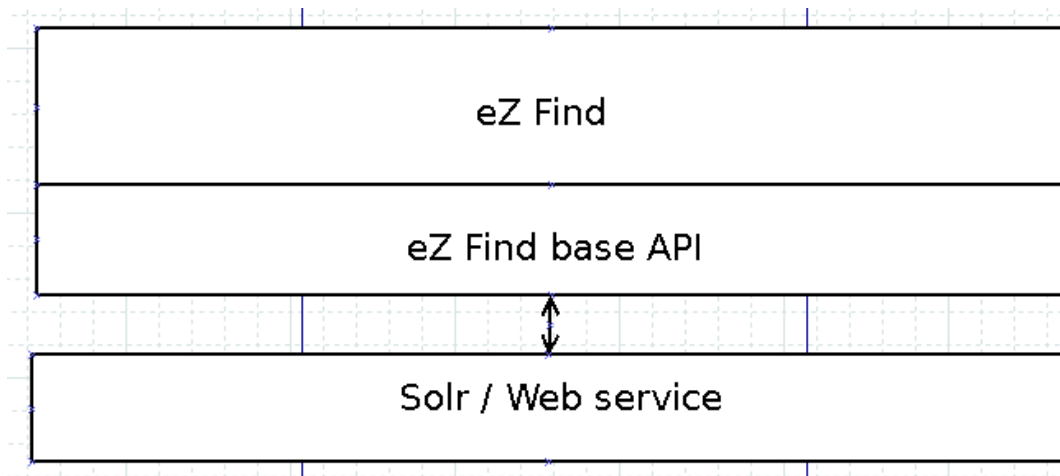
Web service - issues

- Stability
- System requirements
- Scalability
- Service interface / documentation



Web service - implementation

- Generic document base API
- Separation of web service and API



Web service - result

- Easy to use API
- Example web service request

/select

```
hl.fragsize=100&hl.requireFieldMatch=true&facet=true&indent=on&qf=attr_name_t+attr_first_name_t+attr_subject_t+attr_title_t+attr_description_t+attr_last_name_t+attr_caption_t+attr_short_description_t+attr_message_t+attr_short_title_t+attr_author_t+attr_short_name_t+attr_index_title_t+attr_body_t+attr_product_number_t+attr_user_account_t+attr_text_t+attr_multioption2_t+attr_tags_t+attr_left_column_t+attr_image_url_t+attr_calendars_t+attr_image_t+attr_category_t+attr_signature_t+attr_intro_t+attr_center_column_t+attr_right_column_t+attr_bottom_column_t+attr_additional_options_t+meta_name_t^2.0+meta_owner_name_t^1.5&hl.simple.pre=<b>&hl.fl=attr_name_t&hl.fl=attr_first_name_t&hl.fl=attr_subject_t&hl.fl=attr_title_t&hl.fl=attr_description_t&hl.fl=attr_last_name_t&hl.fl=attr_caption_t&hl.fl=attr_short_description_t&hl.fl=attr_message_t&hl.fl=attr_short_title_t&hl.fl=attr_author_t&hl.fl=attr_short_name_t&hl.fl=attr_index_title_t&hl.fl=attr_body_t&hl.fl=attr_product_number_t&hl.fl=attr_user_account_t&hl.fl=attr_text_t&hl.fl=attr_multioption2_t&hl.fl=attr_tags_t&hl.fl=attr_left_column_t&hl.fl=attr_image_url_t&hl.fl=attr_calendars_t&hl.fl=attr_image_t&hl.fl=attr_category_t&hl.fl=attr_signature_t&hl.fl=attr_intro_t&hl.fl=attr_center_column_t&hl.fl=attr_right_column_t&hl.fl=attr_bottom_column_t&hl.fl=attr_additional_options_t&wt=php&hl=true&version=2.2&rows=10&fl=meta_guid_s+meta_installation_id_s+meta_main_url_alias_s+meta_installation_url_s+meta_id_si+meta_main_node_id_si+meta_language_code_s+meta_name_t+score+meta_published_dt&bq=meta_installation_id_s:f4dc4be2fb8c8e8954ab83b500d447b1^1.5+meta_language_code_s:eng-GB^1.2&hl.snippets=2&start=0&q=grenland&hl.simple.post=</b>&qt=ezipublish&fq=(+(+(meta_installation_id_s:f4dc4be2fb8c8e8954ab83b500d447b1+AND+(+(meta_section_id_si:1+))OR+(+(meta_contentclass_id_si:29+OR+meta_contentclass_id_si:30+OR+meta_contentclass_id_si:31+OR+meta_contentclass_id_si:32+OR+meta_contentclass_id_si:33+OR+meta_contentclass_id_si:40+))AND+(+meta_section_id_si:3+)))+)+OR+meta_anon_access_b:true+)+AND+(+meta_language_code_s:eng-GB+)+)
```



Performance

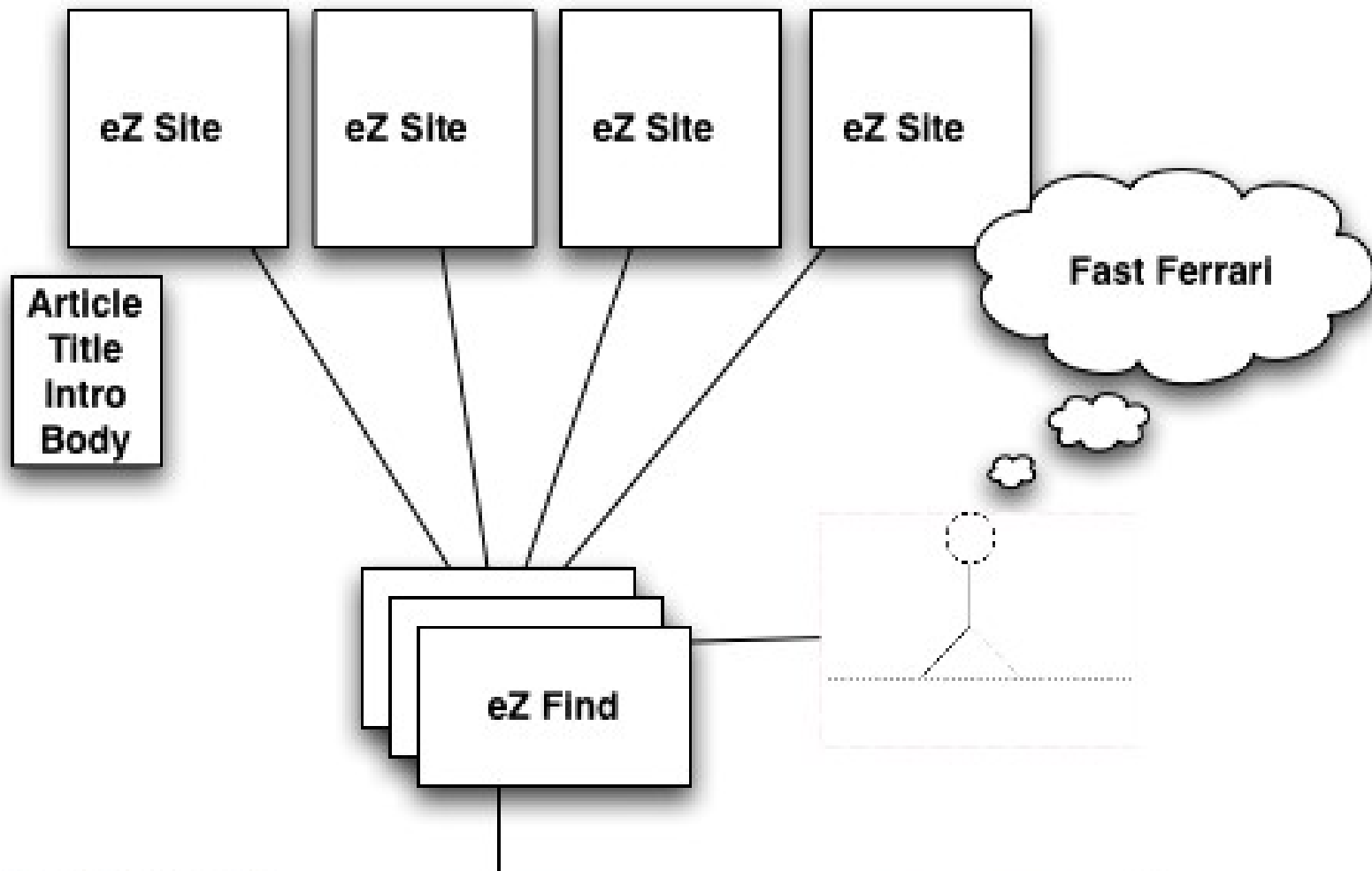
- eZ Publish installation with ca 8000 articles
- Index time (1.7GHz Pentium M, PHP 5.24, JRE 1.6)
 - Zend lucene port : 19 min
 - eZ Find (Solr) : 10 min



Conclusion

- Extracting data from eZ Publish : 8 min
 - Zend lucene port : 11 min indexing
 - eZ Find: 2 min indexing
-





- 73% WPS.2: Internal training 12m on KM system. Attendance at second annual project **workshop**. Preparation for, and facilitation of, review Annual **workshops** (internal training ... Holding second NF-Pro **Workshop** Programme of first annual (internal) **workshop**
- 66% UWC 24m WPS.2: Internal training The second **workshop** was organised and held in Cardiff in October 2005. ... on KM system. Attendance at second annual project **workshop**. Preparation for, and facilitation of, review Annual **workshops** (internal training ... The second **workshop** was organised and held in Cardiff in October 2005. The format of the **workshop** for the **workshop**, with six being offered financial assistance. All of the presentations and a number of written ... A third **workshop** is planned for the autumn of 2006. Details of the format and content



Future plans

- Provide eZ Component
- Enhanced pre / post processing
- SSO support
- and much more



Questions

- I'm available at the eZ Systems booth

