Drupal and Apache Solr Search Go Together Like Pizza and Beer for Your Site

Peter M. Wolanin, Ph.D.
Momentum Specialist (principal engineer), Acquia, Inc.
Drupal contributor drupal.org/user/49851
coop-maintainer of the Drupal Apache Solr Search Integration module
Pizza Without Beer?
Pizza Without Beer?

- Ok, Drupal alone is great, but we can make it even more appealing and satisfying.
- Are you wondering how hard it is to actually integrate Apache Solr with Drupal?
- Do you like things that are easy yet powerful?
Drupal + Solr Provides Immediate Access to Rich Search Features

- Dynamic content requires dynamic navigation - which is provided by an effective search.
- Search facets mean no dead ends.
- Solr provides better keyword relevancy in results.
- Much faster searches for sites with lots of content.
- By avoiding database queries, Drupal with Solr scales better.
Solr Integration Challenges Are Already Solved for You

- The most important - content indexing.
- Facets, sorting, and highlighting of results.
- Immediate integration with the More Like This and spell-check handlers.
- Included sub-module integrates content access permissions by indexing to and filtering Solr results based on the current user.
What are the key Solr concepts you need to understand to get the most out of the Apache Solr Search Integration module?

How is the module admin UI organized?

How do I configure facets, search pages, and content recommendation blocks?

How can I index file attachments?
Drupal sends data to Solr as XML documents
POST XML to /update to add or delete.
Search via GET requests.
If something is not working as expected, you can try searching directly in Solr via URL
Solr also includes admin and analysis interfaces (you need to lock this down for production).
Enable the Modules

<table>
<thead>
<tr>
<th>ENABLED</th>
<th>NAME</th>
<th>VERSION</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Apache Solr Access</td>
<td>7.x-1.1</td>
<td>Integrates node access and other permissions with Solr.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Requires: Apache Solr framework (enabled)</td>
</tr>
<tr>
<td></td>
<td>Apache Solr framework</td>
<td>7.x-1.1</td>
<td>Framework for searching with Solr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Required by: Apache Solr search (enabled), Acquia search (disabled), Solr external index (disabled)</td>
</tr>
<tr>
<td>✔️</td>
<td>Apache Solr search</td>
<td>7.x-1.1</td>
<td>Search with Solr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Requires: Search (enabled), Apache Solr framework (enabled), Acquia search (disabled), Apache Solr external index (disabled)</td>
</tr>
<tr>
<td>✔️</td>
<td>Current Search Blocks</td>
<td>7.x-1.2</td>
<td>Provides an interface for creating blocks containing search results.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Requires: Facet API (enabled), Chaos tools (enabled)</td>
</tr>
<tr>
<td>✔️</td>
<td>Facet API</td>
<td>7.x-1.2</td>
<td>An abstracted facet API that can be used by various components.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Requires: Chaos tools (enabled)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Required by: Apache Solr external index (disabled), Chaos tools (enabled)</td>
</tr>
</tbody>
</table>
## localhost server: Search Index Content

<table>
<thead>
<tr>
<th>TYPE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexed</td>
<td>250 Items (50 sent but not yet processed)</td>
</tr>
<tr>
<td>Remaining</td>
<td>4700 items (6% has been sent to the server)</td>
</tr>
<tr>
<td>Schema</td>
<td>drupal-3.0-rc2-solr3</td>
</tr>
<tr>
<td>Solr Core Name</td>
<td>d7-core</td>
</tr>
<tr>
<td>Delay</td>
<td>2 min before updates are processed.</td>
</tr>
<tr>
<td>Pending Deletions</td>
<td>0</td>
</tr>
</tbody>
</table>

View more details on the search index contents

**Actions**

- **Index queued content (50)**
  - Indexes just as many items as 1 cron run would do.
- **Index all queued content**
  - Could take time and could put an increased load on
### Pages

<table>
<thead>
<tr>
<th>NAME</th>
<th>PATH</th>
<th>SEARCH ENVIRONMENT</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core Search <em>(Default)</em></td>
<td>search/site</td>
<td>localhost server</td>
<td>Edit</td>
</tr>
<tr>
<td>Taxonomy Search</td>
<td>taxonomy/term/%</td>
<td>&lt;Disabled&gt;</td>
<td>Edit</td>
</tr>
</tbody>
</table>

**Blocks "More Like This"**

<table>
<thead>
<tr>
<th>NAME</th>
<th>SEARCH ENVIRONMENT</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>More like this</td>
<td>localhost server</td>
<td>Configure</td>
</tr>
</tbody>
</table>

*Edit, Clone, Delete*
### Add search environment

<table>
<thead>
<tr>
<th>NAME</th>
<th>URL</th>
<th>CONFIGURATION</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>localhost server (Default)</td>
<td><a href="http://localhost:8983/solr/d7-core">http://localhost:8983/solr/d7-core</a></td>
<td>Facets, Bias, Index, Edit, Clone</td>
<td></td>
</tr>
</tbody>
</table>

- **ADVANCED CONFIGURATION**

- Save configuration
Solr server URL *
http://localhost:8983/solr/d7-core
Example: http://localhost:8983/solr

☑ Make this Solr search environment the default

Description *
D7 server
Machine name: solr

Index write access
- Read and write (normal)
- Read only

*Read only* stops this site from sending updates to this search environment. Useful for development sites.

Save  Test connection  Cancel
Settings for: D7 server (Overview)

- Show facets on non-search pages.

The Blocks realm displays each facet in a separate block. Users are able to refine their searches in a drill-down fashion. For performance reasons, you should only enable facets that you intend to have available to users on the search page.

<table>
<thead>
<tr>
<th>ENABLED</th>
<th>FACET</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Author</td>
<td>configure display</td>
</tr>
<tr>
<td></td>
<td>Filter by author.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Content type</td>
<td>configure display</td>
</tr>
<tr>
<td></td>
<td>Filter by content type.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td>configure display</td>
</tr>
<tr>
<td></td>
<td>Filter by language.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Post date</td>
<td>configure display</td>
</tr>
<tr>
<td></td>
<td>Filter by the date the node was posted.</td>
<td></td>
</tr>
</tbody>
</table>
Configure facet display

DISPLAY SETTINGS

Display widget

Links

Select the display widget used to render this facet.

Soft limit

20

Limits the number of displayed facets via JavaScript.

Prevent crawlers from following facet links

Add the rel="nofollow" attribute to facet links to maximize SEO by preventing crawlers from indexing duplicate content and getting stuck in loops.

Empty facet behavior

Do not display facet

Show row weights

<table>
<thead>
<tr>
<th>SORT</th>
<th>ORDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facet active</td>
<td>Descending</td>
</tr>
<tr>
<td>Count</td>
<td>Descending</td>
</tr>
<tr>
<td>Display value</td>
<td>Ascending</td>
</tr>
<tr>
<td>Indexed value</td>
<td>Ascending</td>
</tr>
</tbody>
</table>
Empty facet behavior

Do not display facet

The action to take when a facet has no items.

GLOBAL SETTINGS

The configuration options below apply to this facet across all realms.

Operator

- AND
- OR

AND filters are exclusive and narrow the result set. OR filters are inclusive and widen the result set.

Hard limit

50

Display no more than this number of facet items.

Minimum facet count

1

Only display facets that are matched in at least this many documents.

Save configuration    Save and go back to realm settings    Cancel
The configuration options have been saved. To enable or arrange the facet blocks, visit the blocks administration page.

Settings for: D7 server (Overview)

- Show facets on non-search pages.

The Blocks realm displays each facet in a separate block. Users are able to refine their searches in a drill-down fashion. For performance reasons, you should only enable facets that you intend to have available to users on the search page.

<table>
<thead>
<tr>
<th>ENABLED</th>
<th>FACET</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Author</td>
<td>configure display</td>
</tr>
<tr>
<td>✅</td>
<td>Filter by author.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Content type</td>
<td>configure display</td>
</tr>
<tr>
<td>✅</td>
<td>Filter by content type.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td>configure display</td>
</tr>
<tr>
<td></td>
<td>Filter by language.</td>
<td></td>
</tr>
<tr>
<td>BLOCK</td>
<td>REGION</td>
<td>OPERATIONS</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>Disabled</td>
<td></td>
<td>configure</td>
</tr>
<tr>
<td>‡  Apache Solr Core: Sorting</td>
<td>- None -</td>
<td>configure</td>
</tr>
<tr>
<td>‡  Apache Solr recommendations: More like this</td>
<td>- None -</td>
<td>configure delete</td>
</tr>
<tr>
<td>‡  Current search: Standard</td>
<td>- None -</td>
<td>configure</td>
</tr>
<tr>
<td>‡  Facet API: Apache Solr environment: D7 server : Author</td>
<td>- None -</td>
<td>configure</td>
</tr>
<tr>
<td>‡  Facet API: Apache Solr environment: D7 server : Content type</td>
<td>- None -</td>
<td>configure</td>
</tr>
<tr>
<td>‡  Facet API: Apache Solr environment: D7 server : Tags</td>
<td>- None -</td>
<td>configure</td>
</tr>
</tbody>
</table>
?q=search/node/ratis  WTH? no facets!

Search

Enter your keywords
ratis

Advanced search

Search results

Erat Haero Jus Minim

... Cogo melior roto saluto. Caecus dui iusto occuro ratis saepius suscipere uxor. Abigo neo occuro z
Brevitas erat ex laoreet ... luctus venio. Causa illum lenis validus. Aptent esca ex ratis refoveo. Aptent te
valetudo. Antehabeo iustum ludus. Damnnum eum ...

hibruk - 05/22/2012 - 22:16 - 9 comments

Duis Scisco
Simple CJK handling

Whether to apply a simple Chinese/Japanese/Korean tokenizer based on overlapping sequences. Turn this off if you want to use an external preprocessor for this instead. Does not affect other languages.

ACTIVE SEARCH MODULES

- Apache Solr search
- Node
- User

Choose which search modules are active from the available modules.

Default search module

- Apache Solr search
- Node
- User
Search

Enter terms
ratis

Did you mean
paratus

Search results

Ratis Saepius Valde Wisi

Diam paulatim quae sudo. Abico bene ludus oppeto qui. Appellatio esse eu gravis huic nimis oppeto ratis scisco verto. Aliquam lucidus nulla olim praesent uxor. At defui elit gilvus usitas vulpes. Eligo gilvus hos iusto singularis. Dolus humo nimis os premo. Ad erat ex meas ... blandit damnum erat lucidus ratis. Eros metu mos pneum refoveo tation typicus utinam valde virtus. Aptent ...

wracha - 05/22/2012 - 22:16 - 10 comments
Easy Content Recommendation

- Uses the MLT handler
- Picks fields from the currently viewed node

---

Search with Big Data Principles

**Submitted by:** Eric Pugh

*nSource Connections*

How many of millions of documents to search? ImportHandler flipping up while indexing? Random threads thrown by Solr Cell during document extraction? Query performance collapsing? Then you've searching at Big Data scale. This talk will focus on the underlying principles of Big Data, and to apply them to Solr. This talk isn't a deep dive into Cloud, though we'll talk about it. It also isn't meant to be a traditional scaling of Solr. Instead we'll talk about how to apply the principles of big data like "Bring the code to the data, not data to the code" to Solr. How to answer the question "How..."
A short diversion...

Search Environments Reference
Different Servers and/or Config

- Most people need only one to start.
- The most important use is to bundle different sets of enabled facets and their configuration - e.g. for different search pages.
- Can also be used to search multiple servers.
- Each has its own ID and config variables.
# Apache Solr search

## Add search environment

<table>
<thead>
<tr>
<th>NAME</th>
<th>URL</th>
<th>CONFIGURATION</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>D7 server</td>
<td><a href="http://localhost:8983/solr/d7-core">http://localhost:8983/solr/d7-core</a></td>
<td>Facets Bias Index</td>
<td>Edit Clone</td>
</tr>
</tbody>
</table>

### ADVANCED CONFIGURATION

Save configuration
### Add search environment

<table>
<thead>
<tr>
<th>NAME</th>
<th>URL</th>
<th>CONFIGURATION</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D7 server</strong></td>
<td><a href="http://localhost:8983/solr/d7-core">http://localhost:8983/solr/d7-core</a></td>
<td>Facets Bias Index Edit Clone</td>
<td></td>
</tr>
<tr>
<td><em>Default</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D7 server</strong></td>
<td><a href="http://localhost:8983/solr/d7-core">http://localhost:8983/solr/d7-core</a></td>
<td>Facets Bias Index Edit Clone Delete</td>
<td></td>
</tr>
<tr>
<td><em>cloned</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### ADVANCED CONFIGURATION

Save configuration
Block title

Override the default title for the block. Use <none> to display no title, or leave blank to use the default block title.

Block name *

More like this

The block name displayed to site users.

Maximum number of related items to display

5

Search environment

D7 server

Fields for finding related content *

☐ The full, rendered content (e.g. the rendered node body)

☑ Title or label

☐ Path alias
Edit search page

**Label**

Core Search  
Machine name: core_search  
The human-readable name of the search page configuration.

**Description**

Core Search

**Make this Solr Search Page the default**

Useful for eg. making facets to link to this page when they are shown on non-search pages

**SEARCH PAGE INFORMATION**

**Search environment**

D7 server  
This page always uses the current default search environment

**Title**

Site  
You can use %value to place the search term in the title
Apache Solr search

Label *
The human-readable name of the search page configuration.

Description

Make this Solr Search Page the default
Useful for eg. making facets to link to this page when they are shown on non-search pages

SEARCH PAGE INFORMATION

Search environment
<Disabled>
The environment that is used by this search page. If no environment is selected, this page will be disabled.

Title *
You can use %value to place the search term in the title

Search Type
Custom Field
The Module Has a Pipeline for Indexing Drupal Content to Solr

- Drupal entities are processed into one (or more) document objects. Each document object is converted to XML and sent to Solr.

**Node object**
- title
- nid
- type

**Drupal callbacks & hooks**

**Document object**
- entity_type
- label
- entity_id
- bundle

**XML string**

```xml
<doc>
  <field name="entity_type">node</field>
  <field name="label">Hello Drupal</field>
  <field name="entity_id">101</field>
  <field name="bundle">session</field>
</doc>
```
<table>
<thead>
<tr>
<th>Key</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>String, 16 characters</td>
<td>a520t1/node/2203</td>
</tr>
<tr>
<td>site</td>
<td>String, 20 characters</td>
<td><a href="http://drupal-7.dev/">http://drupal-7.dev/</a></td>
</tr>
<tr>
<td>hash</td>
<td>String, 6 characters</td>
<td>a520t1</td>
</tr>
<tr>
<td>entity_id</td>
<td>String, 4 characters</td>
<td>2203</td>
</tr>
<tr>
<td>entity_type</td>
<td>String, 4 characters</td>
<td>node</td>
</tr>
<tr>
<td>bundle</td>
<td>String, 7 characters</td>
<td>article</td>
</tr>
<tr>
<td>bundle_name</td>
<td>String, 7 characters</td>
<td>Article</td>
</tr>
<tr>
<td>path</td>
<td>String, 9 characters</td>
<td>node/2203</td>
</tr>
<tr>
<td>url</td>
<td>String, 37 characters</td>
<td><a href="http://drupal-7.dev/node-2203-article">http://drupal-7.dev/node-2203-article</a></td>
</tr>
<tr>
<td>language</td>
<td>String, 3 characters</td>
<td>und</td>
</tr>
<tr>
<td>path_alias</td>
<td>String, 17 characters</td>
<td>node-2203-article</td>
</tr>
<tr>
<td>label</td>
<td>String, 21 characters</td>
<td>Consectetuer Pertineo</td>
</tr>
<tr>
<td>content</td>
<td>String, 2937 characters</td>
<td>Abluo aliquam eu facilisi...</td>
</tr>
<tr>
<td>teaser</td>
<td>String, 296 characters</td>
<td>Abluo aliquam eu facilisi...</td>
</tr>
<tr>
<td>ss_name</td>
<td>String, 7 characters</td>
<td>brutrod</td>
</tr>
<tr>
<td>tos_name</td>
<td>String, 7 characters</td>
<td>brutrod</td>
</tr>
</tbody>
</table>
Entity Meta-data Gives Automatic Facets

- Content types
- Taxonomy terms per field
- Content authors
- Posted and modified dates
- Text and numbers selected via select list/radios/check boxes
Updates to an Entity or Related Meta-data Cause Reindexing

- Drupal entities are indexed during Drupal cron.
- By using a specialized tracking table, content can automatically be queued for reindex when changed, and subsets of content can potentially be sent to different Solr indexes.
- Entities include many ID-based reference fields (e.g. the User ID of the node author). Changes to the referenced data is also watched.
Finding the “Right” Results

- A big frustration is when the result you expect for a keyword or set of keywords is not first, or even on the first page.
- Apache Solr has very flexible result scoring - you just need to know how to tune it.
- Different sites have different needs - the default settings may be poor for yours.
- acquia.com/blog/delivering-right-search-results
Settings for: D7 server (Overview)

Result biasing

Give bias to certain properties when ordering the search results. Any value except *Ignore* will increase the score of the given type in search results. Choose *Ignore* to ignore any given property.

**Sticky at top of lists**

*Ignore*

Select additional bias to give to nodes that are set to be 'Sticky at top of lists'.

**Promoted to home page**

*Ignore*

Select additional bias to give to nodes that are set to be 'Promoted to home page'.

**More recently created**

*Ignore*
Settings for: D7 server (Overview)

Result biasing

Type biasing and exclusion

Field biases

Article type content bias

Ignore

Basic page type content bias

Ignore

Specify here which node types should get a higher relevancy score in searches. Any value except ignore will increase the score of the given type in search results.
Specify here which fields are more important when searching. Give a field a greater numeric value to make it more important. If you omit a field, it will not be searched.

**The full, rendered content (e.g. the rendered node body)**

1.0

**Title or label**

5.0

**Path alias**

Omit

**Body text inside links (A tags)**

Omit
More Modules Available to Add More Features

A few examples:

- ApacheSolr Attachments
- Apache Solr Multisite Search
- Apache Solr Organic Groups Integration
- Apachesolr User indexing
- Apachesolr Commerce
## Attachments Too

<table>
<thead>
<tr>
<th>Enabled</th>
<th>Name</th>
<th>Version</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Apache Solr Access</td>
<td>7.x-1.1</td>
<td>Integrates node access and other permissions. Requires: Apache Solr framework (enabled)</td>
</tr>
<tr>
<td>✔️</td>
<td>Apache Solr framework</td>
<td>7.x-1.1</td>
<td>Framework for searching with Solr. Required by: Apache Solr search (enabled), Acquia Solr search attachments (disabled), Apache Solr extra</td>
</tr>
<tr>
<td>✔️</td>
<td>Apache Solr search</td>
<td>7.x-1.1</td>
<td>Search with Solr. Requires: Search (enabled), Apache Solr framework. Required by: Acquia search (disabled), Apache Solr extra</td>
</tr>
<tr>
<td></td>
<td>Apache Solr search attachments</td>
<td>7.x-1.2</td>
<td>Search file attachments with Solr. Requires: Apache Solr framework (enabled)</td>
</tr>
</tbody>
</table>
Excluded file extensions

aif art avi bmp gif ico jpg mov mp3 mp4 mpg oga ogv png psd ra ram rgb tif wmv

File extensions that are excluded from indexing. Separate extensions with a space and do not include the leading dot. Extensions are internally mapped to a MIME type, so it is not necessary to put variations that map to the same type (e.g. tif is sufficient for tif and tiff)

Extract using

- Tika (local java application)
- Solr (remote server)

Extraction will be faster if run locally using tika.

Tika directory path

The full path to the tika directory. All library jars must be in the same directory. If on Windows, use forward slashes in the path.

Tika jar file

tika-app-1.1.jar
To Wrap Up

- Drupal has extensive Apache Solr integration already, and it is highly customizable in the UI.
- Apache Solr Search Integration offers more robust integration as compared to Search API Solr and both Drupal 6 and 7 support.
- Acquia includes a secure, hosted Solr index with every support subscription. Get started fast with a 30 day free trial.
Do you love Drupal, Solr, the LAMP stack, DevOps or anything related, and working at a fast-growing and successful startup?

- Boston, Portland, D.C. area U.S. offices.
- Some remote opportunities as well.
- Come talk to me!
  peter.wolanin@acquia.com
  pwolanin in IRC #drupal-apachesolr