

February 2, 2013

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

co-maintainer of the Drupal Apache Solr Search Integration module

Acquia[™]

Pizza Without Beer?



Pizza Without Beer?

- Ok, Drupal alone is great, but a 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.

Key Questions to Be Answered

- 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?

Solr Interface/API is HTTP

- 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).

Solr Admin (drupal-3.0-0-solr3)

tvwna-ip-f-92.princeton.org:8983

cwd=/Users/Shared/www/apache-solr-3.5.0-quickstart SolrHome=multicore/d7-core/

HTTP caching is ON



Solr	[SCHEMA] [CONFIG] [ANALYSIS] [SCHEMA BROWSER] [STATISTICS] [INFO] [DISTRIBUTION] [PING] [LOGGING]
Cores:	[D6-CORE] [D6.X-3.X-CORE] [d7-core] [EXT]
App server:	[JAVA PROPERTIES] [THREAD DUMP]

Make a Query [\[FULL INTERFACE\]](#)

Query String:

solr

Search

Assistance [\[DOCUMENTATION\]](#) [\[ISSUE TRACKER\]](#) [\[SEND EMAIL\]](#) [\[SOLR QUERY SYNTAX\]](#)

Current Time: Sat Feb 02 10:52:41 EST 2013

Server Start At: Sat Feb 02 10:52:24 EST 2013

Enable the Modules

▼ SEARCH TOOLKIT

ENABLED	NAME	VERSION	DESCRIPTION
<input type="checkbox"/>	Apache Solr Access	7.x-1.1	Integrates node access and other permissions with Solr Requires: Apache Solr framework (enabled)
<input checked="" type="checkbox"/>	Apache Solr framework	7.x-1.1	Framework for searching with Solr Required by: Apache Solr search (enabled), Acquia search (disabled), Apache Solr external index (disabled)
<input checked="" type="checkbox"/>	Apache Solr search	7.x-1.1	Search with Solr Requires: Search (enabled), Apache Solr framework (enabled) Required by: Acquia search (disabled), Apache Solr external index (disabled)
<input checked="" type="checkbox"/>	Current Search Blocks	7.x-1.2	Provides an interface for creating blocks containing search results Requires: Facet API (enabled), Chaos tools (enabled)
<input checked="" type="checkbox"/>	Facet API	7.x-1.2	An abstracted facet API that can be used by various modules Requires: Chaos tools (enabled) Required by: Apache Solr external index (disabled), C

localhost server: Search Index Content

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

[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

Apache Solr search


DEFAULT INDEX

PAGES/BLOCKS


SETTINGS

[+ Add search page](#) [+ Add search block "More Like This"](#)

Pages

NAME 	PATH	SEARCH ENVIRONMENT	OPERATIONS		
Core Search <i>(Default)</i>	search/site	localhost server	Edit	Clone	
Taxonomy Search	taxonomy/term/%	<Disabled>	Edit	Clone	Delete

Blocks "More Like This"

NAME 	SEARCH ENVIRONMENT	OPERATIONS	
More like this	localhost server	Configure	Delete

Apache Solr search

DEFAULT INDEX

PAGES/BLOCKS

SETTINGS

[+ Add search environment](#)

NAME	URL	CONFIGURATION			OPERATIONS	
localhost server (Default)	http://localhost:8983/solr/d7- core	Facets	Bias	Index	Edit	Clone

► **ADVANCED CONFIGURATION**

Save configuration

Apache Solr search

DEFAULT INDEX

PAGES/BLOCKS

SETTINGS

Facets

Index

Bias

Edit

Solr server URL *

Example: <http://localhost:8983/solr>

☒ Make this Solr search environment the default

Description *

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.

ENABLED	FACET	OPERATIONS
<input type="checkbox"/>	Author Filter by author.	configure display ▼
<input type="checkbox"/>	Content type Filter by content type.	configure display ▼
<input type="checkbox"/>	Language Filter by language.	configure display ▼
<input type="checkbox"/>	Post date Filter by the date the node was posted.	configure display ▼

Configure facet display

DISPLAY SETTINGS

[Show row weights](#)

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

	SORT	ORDER
<input checked="" type="checkbox"/>	Facet active Sort by whether the facet is active or not.	Descending
<input checked="" type="checkbox"/>	Count Sort by the facet count.	Descending
<input checked="" type="checkbox"/>	Display value Sort by the value displayed to the user.	Ascending
<input type="checkbox"/>	Indexed value Sort by the raw value stored in the index.	Ascending

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.

ENABLED	FACET	OPERATIONS
<input checked="" type="checkbox"/>	Author Filter by author.	configure display ▼
<input checked="" type="checkbox"/>	Content type Filter by content type.	configure display ▼
<input type="checkbox"/>	Language Filter by language.	configure display ▼

BLOCK	REGION	OPERATIONS
Disabled		
⊕ Apache Solr Core: Sorting	- None -	configure
⊕ Apache Solr recommendations: More like this	- None -	configure delete
⊕ Current search: Standard	- None -	configure
⊕ Facet API: Apache Solr environment: D7 server : Author	- None -	configure
⊕ Facet API: Apache Solr environment: D7 server : Content type	- None -	configure
⊕ Facet API: Apache Solr environment: D7 server : Tags	- None -	configure

?q=search/node/ratis

WTH? no facets!

Search

Content

Site

Users

Enter your keywords

ratis



▸ Advanced search

Search results

Erat Haero Jus Minim

... Cogo melior roto saluto. Caecus duis ideo 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. Damnum eum ...

[hibruk](#) - 05/22/2012 - 22:16 - 9 comments

Duis Scisco

Search settings | drupal-7.de

drupal-7.dev/admin/config/search/settings

Home

Dashboard

Content

Structure

Appearance

People

Modules

Configuration

Reports

Help

Hello peter

Log out

☒ 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

20

?q=search/site/ratis

Search

Content

Site

Users

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 meus ... blandit damnum erat lucidus **ratis**. Eros metuo mos pneum refoveo tation typicus utinam valde virtus. Aptent ...

[wracha](#) - 05/22/2012 - 22:16 - 10 comments

Filter by author:

- ☐ [Anonymous \(160\)](#)
- ☐ [brutrod \(103\)](#)
- ☐ [viuaclochok \(102\)](#)
- ☐ [reprukuc \(97\)](#)
- ☐ [wribra \(97\)](#)
- ☐ [wufrophetriw \(96\)](#)
- ☐ [nagatikijeb \(94\)](#)
- ☐ [spokanesha \(93\)](#)
- ☐ [stedrijo \(92\)](#)
- ☐ [lilisw \(90\)](#)

[Show more](#)

Filter by content type:

- [Article \(2481\)](#)
- [Basic page \(1590\)](#)

Filter by tags:

► [wriiuloki \(1806\)](#)

Easy Content Recommendation

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



Scaling Search with Big Data Principles

Presented by: Eric Pugh

LinkedInSource Connections

hundreds of millions of documents to search?

NullPointerException blowing up while indexing? Random threads thrown by Solr Cell during document extraction? Query performance collapsing? Then you're searching at Big Data scale.

This talk will focus on the underlying principles of Big Data, and how to apply them to Solr. This talk isn't a deep dive into the Cloud, though we'll talk about it. It also isn't meant to be a replacement for traditional scaling of Solr. Instead we'll talk about how to apply principles of big data like "Bring the code to the data, not the data to the code" to Solr. How to answer the question "How

More like this

- [Scaling Search with Big Data and Solr](#)
- [Indexing Wikipedia as a Benchmark of Single Machine Performance Limits](#)
- [NetDocuments - Journey from FAST to Solr](#)
- ["Stump The Chump": Get On The Spot Solutions To Your Real Life Solr/Lucene Challenges](#)
- [Indexing Big Data in the Cloud](#)

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

DEFAULT INDEX

PAGES/BLOCKS

SETTINGS

[+ Add search environment](#)

NAME	URL	CONFIGURATION			OPERATIONS	
D7 server <i>(Default)</i>	http://localhost:8983/solr/d7- core	Facets	Bias	Index	Edit	Clone



► **ADVANCED CONFIGURATION**

Save configuration

Apache Solr search

DEFAULT INDEX

PAGES/BLOCKS

SETTINGS

[+ Add search environment](#)

NAME	URL	CONFIGURATION			OPERATIONS		
D7 server (Default)	http://localhost:8983/solr/d7-core	Facets	Bias	Index	Edit	Clone	
D7 server [cloned]	http://localhost:8983/solr/d7-core	Facets	Bias	Index	Edit	Clone	Delete

► **ADVANCED CONFIGURATION**

Save configuration

'*Apache Solr recommendations: More like this*' block

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 *

The block name displayed to site users.

Maximum number of related items to display

Search environment

Fields for finding related content *

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

☒ Title or label

☐ Path alias

Edit search page

Label *

Machine name: core_search

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



This page always uses the current default search environment

Title *

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



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

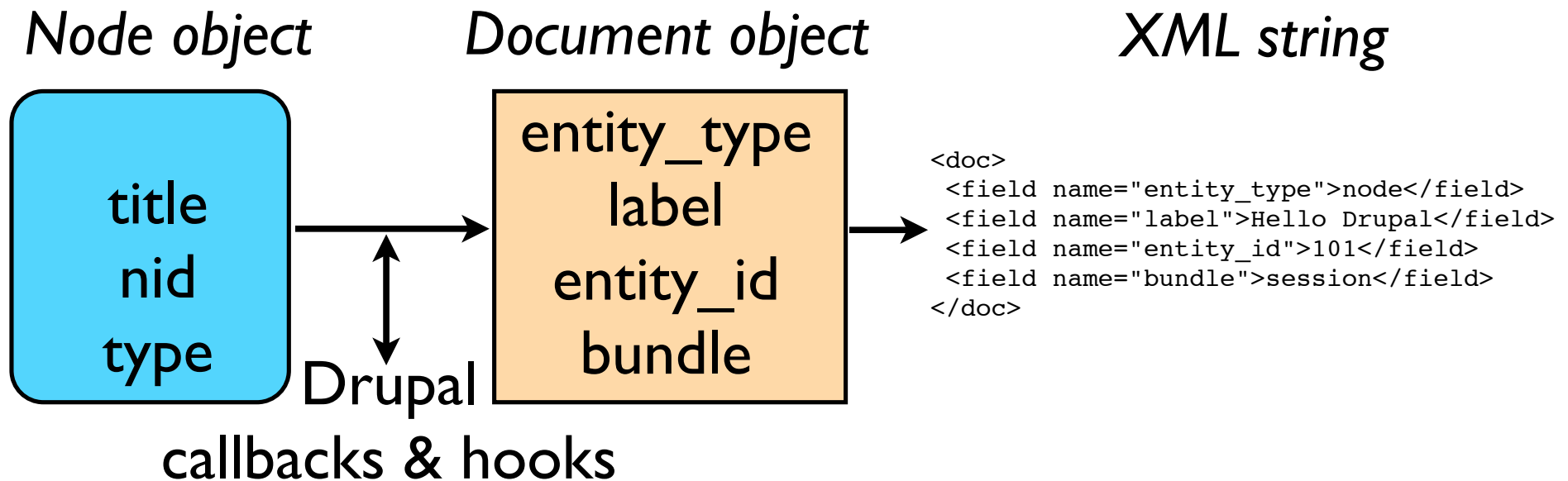
Title *

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

Search Type

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.



... (Array, 36 elements)

id (String, 16 characters) a520t1/node/2203

site (String, 20 characters) http://drupal-7.dev/

hash (String, 6 characters) a520t1

entity_id (String, 4 characters) 2203

entity_type (String, 4 characters) node

bundle (String, 7 characters) article

bundle_name (String, 7 characters) Article

path (String, 9 characters) node/2203

url (String, 37 characters) http://drupal-7.dev/node-2203-article

language (String, 3 characters) und

path_alias (String, 17 characters) node-2203-article

label (String, 21 characters) Consectetuer Pertineo

content (String, 2937 characters) Abluo aliquam eu facilisi...

teaser (String, 296 characters) Abluo aliquam eu facilisi...

ss_name (String, 7 characters) brutrod

tos_name (String, 7 characters) brutrod

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

Filter by content type:

- [Session \(29\)](#)
- [Basic page \(2\)](#)
- [Speaker Bio \(2\)](#)
- [Blog entry \(1\)](#)

Filter by author:

- [peter \(29\)](#)
- [Eliza Not \(2\)](#)
- [ameena.syeda \(1\)](#)
- [Joe Smalls \(1\)](#)
- [Robert \(1\)](#)

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

Type biasing and exclusion

Field biases


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.

Save configuration

Reset to defaults

Settings for: D7 server ([Overview](#))

Result biasing

Type biasing and exclusion

Field biases

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

▼ SEARCH TOOLKIT

ENABLED	NAME	VERSION	DESCRIPTION
<input type="checkbox"/>	Apache Solr Access	7.x-1.1	Integrates node access and other permissions w Requires: Apache Solr framework (enabled)
<input checked="" type="checkbox"/>	Apache Solr framework	7.x-1.1	Framework for searching with Solr Required by: Apache Solr search (enabled), Acquia Solr search attachments (disabled), Apache Solr ext
<input checked="" type="checkbox"/>	Apache Solr search	7.x-1.1	Search with Solr Requires: Search (enabled), Apache Solr framework Required by: Acquia search (disabled), Apache Solr
<input type="checkbox"/>	Apache Solr search attachments	7.x-1.2	Search file attachments with Solr Requires: Apache Solr framework (enabled)

Apache Solr search

DEFAULT INDEX

ATTACHMENTS

PAGES/BLOCKS

SETTINGS

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.

Acquia is Hiring!

- 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