



October 7, 2010



# Drupal Integration with Solr for Fabulous CMS Search

**Peter M. Wolanin, Ph.D.**

Principal Engineer, Acquia, Inc.

Drupal contributor [drupal.org/user/49851](http://drupal.org/user/49851)

co-maintainer of the Drupal Apache Solr Search Integration module

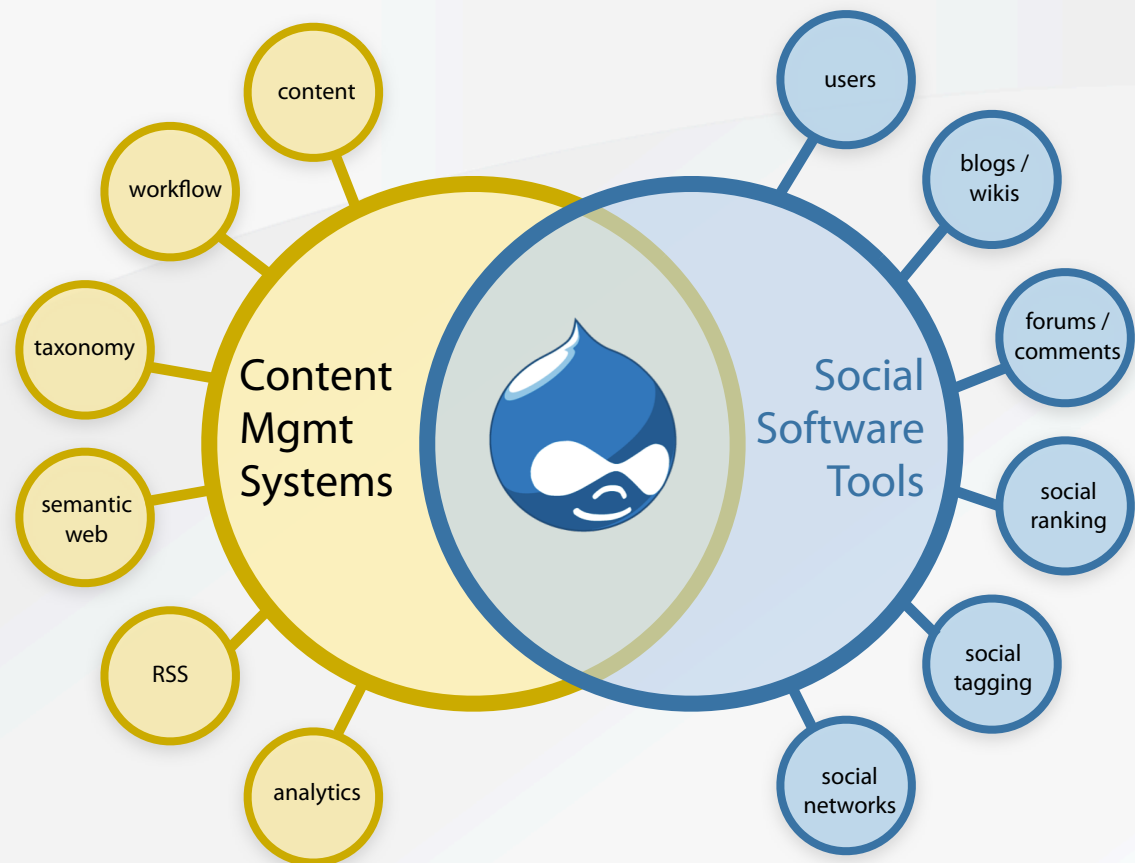
# What You Will Learn

- A little about Drupal
- Drupal terminology
- Examples of Drupal sites using Apache Solr
- How Drupal works with Apache Solr
- Configuration options for searches
- Customization possibilities

# Drupal: Web Application Framework + CMS == Social Publishing Platform

*Drupal "... is as much a Social Software platform as it is a web content management system."*

CMS Watch, The Web CMS Report 2009



# Drupal Has User Accounts, Roles & Permissions



- Define custom roles
- Set granular access controls by role
  - Task / object-based
- Configure user behavior:
  - Registration
  - Email
  - Profiles
  - Pictures

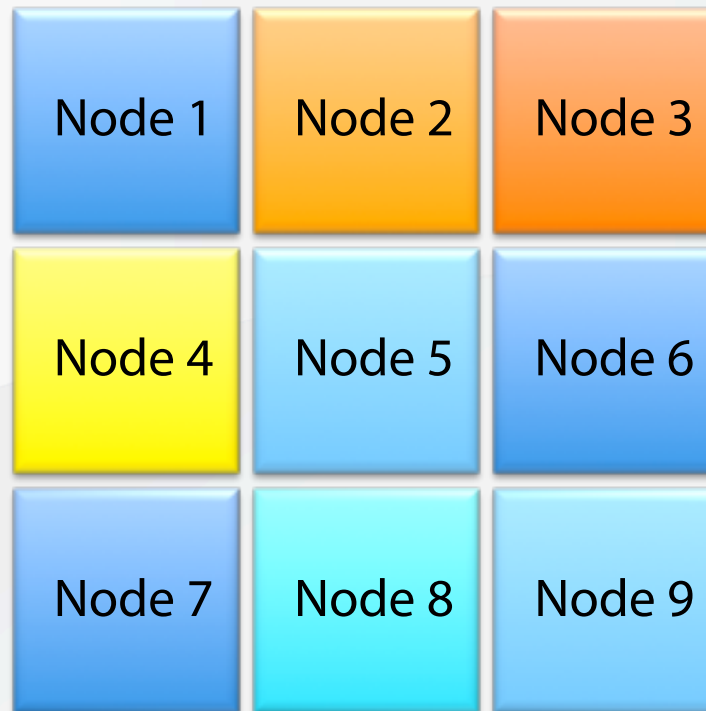
Permission	anonymous user	authenticated user
<b>aggregator module</b>		
access news feeds	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
administer news feeds	<input type="checkbox"/>	<input type="checkbox"/>
<b>block module</b>		
administer blocks	<input type="checkbox"/>	<input type="checkbox"/>
use PHP for block visibility	<input type="checkbox"/>	<input type="checkbox"/>
<b>blog module</b>		
create blog entries	<input type="checkbox"/>	<input checked="" type="checkbox"/>
delete any blog entry	<input type="checkbox"/>	<input type="checkbox"/>
delete own blog entries	<input type="checkbox"/>	<input checked="" type="checkbox"/>
edit any blog entry	<input type="checkbox"/>	<input type="checkbox"/>
edit own blog entries	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<b>comment module</b>		
access comments	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
administer comments	<input type="checkbox"/>	<input type="checkbox"/>
post comments	<input type="checkbox"/>	<input checked="" type="checkbox"/>
post comments without approval	<input type="checkbox"/>	<input checked="" type="checkbox"/>





# Drupal Nodes are Content + Data

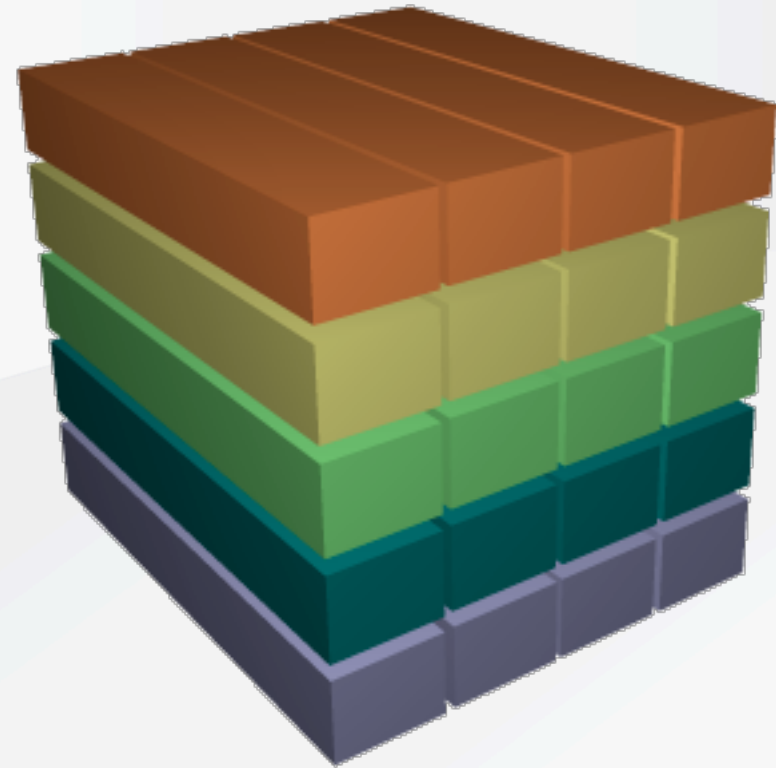
- Nodes are the basic unit of content
- The node system is extensible - can represent any data
- Examples of content stored within Drupal
  - Text
  - Images
  - MP3s
  - Node reference



# Node Types are Enriched With CCK Fields



- Define new data fields within a node using the CCK module.
  - Text, images, integers, date, reference, etc
- Flexible and configurable in the UI
- No programming required (many existing modules define fields)



# Apply Taxonomy Vocabulary & Terms

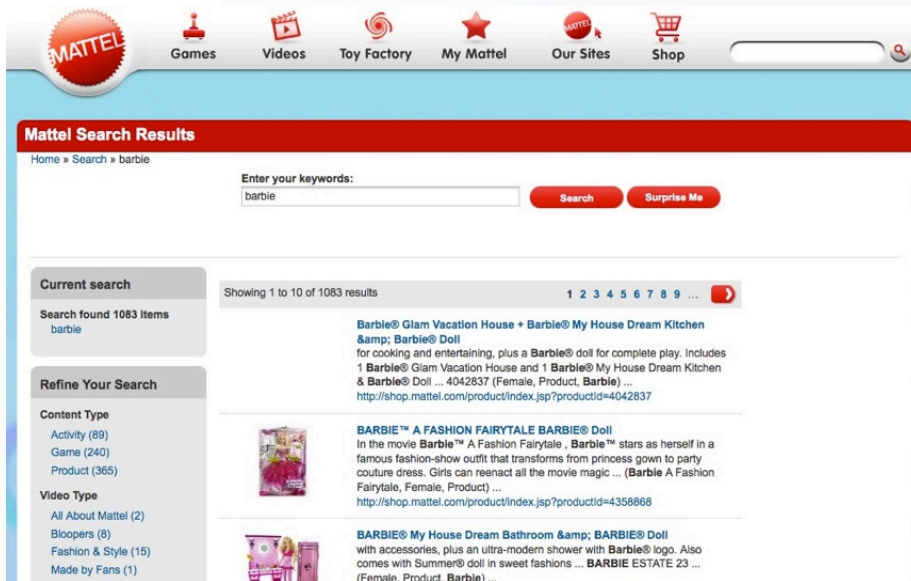
- Provide a strong framework for content classification
- Modules provide taxonomy-based appearance, access control
- Standard input options include free tagging, flat-controlled, and hierarchical-controlled

Terms in *Documentation topics* [List](#) [Add term](#)

*Documentation topics* is a single hierarchy vocabulary. You may organize the terms in the *Documentation topics* vocabulary by using the handles on the left side of the table. To change the name or description of a term, click the *edit* link next to the term. [\[more help...\]](#)

Name	Operations
+ A record	<a href="#">edit</a>
+ about	<a href="#">edit</a>
+ about us page	<a href="#">edit</a>
+ access	<a href="#">edit</a>
+ add this	<a href="#">edit</a>
+ AddThis	<a href="#">edit</a>
+ administer	<a href="#">edit</a>
+ alias	<a href="#">edit</a>
+ analytics	<a href="#">edit</a>
+ appearance	<a href="#">edit</a>
+ attribution	<a href="#">edit</a>
+ author	<a href="#">edit</a>
+ avatar	<a href="#">edit</a>
+ back end	<a href="#">edit</a>

# Drupal + Solr Powers Search for Businesses, Governments and NGOs



<http://www.whitehouse.gov/search/site/>

[http://www.mattel.com/search/apachesolr\\_search/](http://www.mattel.com/search/apachesolr_search/)

[http://opensource.com/search/apachesolr\\_search/](http://opensource.com/search/apachesolr_search/)

<https://www.ethicshare.org/publications/>

[http://www.nypl.org/search/apachesolr\\_search/](http://www.nypl.org/search/apachesolr_search/)

[http://www.mylifetime.com/community/search/apachesolr\\_search/](http://www.mylifetime.com/community/search/apachesolr_search/)

[http://www.restorethegulf.gov/search/apachesolr\\_search/](http://www.restorethegulf.gov/search/apachesolr_search/)

[http://www.hrw.org/en/search/apachesolr\\_search/](http://www.hrw.org/en/search/apachesolr_search/)

[http://www.poly.edu/search/apachesolr\\_search/](http://www.poly.edu/search/apachesolr_search/)

# Drupal + Solr Has Immediate Benefits



- *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

# Node Data Gives Automatic Facets

- Content types
- Taxonomy terms per vocabulary
- Content authors
- Posted and modified dates
- Text and numbers selected via select list/radios/check boxes

**Filter by author**

- rcross (31)
- christefano (28)
- ultimike (21)
- Chris Charlton (20)
- eliza411 (13)
- drewish (11)
- robeano (11)
- Walt Esquivel (10)
- Slurpee (9)
- Amazon (8)

Show more

**Filter by Sitewide tags**

- DrupalCamp (36)
- meetup (36)
- Drupal (13)
- drupalcon (12)
- user group (11)
- Austin (8)
- training (8)
- events (7)
- meeting (7)
- drupal camp (6)

Show more

**Filter by post date**

- 2006 (43)
- 2007 (108)

# Easily Add Content Recommendation

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

## Who is Drupal Gardens for?

Drupal Gardens is designed for people who want a quick, affordable way to build a website with a great design and lots of flexibility without the need for programming or IT maintenance. Core Drupal is the right choice for a large, complex site with custom functionality or integration with other systems. Drupal Gardens is a right choice for smaller sites that need to be built fast and that are satisfied by the features included in the service.

[Add new comment](#)

General

## Related content

- [Top 10 Reasons for Marketers to build Microsites](#)
- [Drupal Gardens update - helping the designer](#)
- [New feature: duplicate your site and create your own templates](#)
- [Quick Start Guide](#)
- [Removing the mittens in Drupal Gardens' June 10 update](#)



# Configure Each MLT Block in the UI



## 'Apache Solr recommendations: Related content' block

### Block specific settings

#### Block name: \*

The block name displayed to site users.

#### Maximum number of related items to display:

#### Fields for finding related content: \*

- Body text - the full, rendered content
- Author name
- Path alias
- All taxonomy term names
- Title
- Taxonomy term names only from the *Documentation topics* vocabulary
- Taxonomy term names only from the *FAQ categories* vocabulary
- Taxonomy term names only from the *Forums* vocabulary
- Taxonomy term names only from the *Forum topic status* vocabulary

Choose the fields to be used in calculating similarity. The default combination of *All taxonomy term names* and *Title* will provide relevant results for typical sites.

—▷ [Advanced configuration](#)

# Advanced Solr Features Plus Configuration in the UI

- Dynamic fields in schema.xml index CCK and custom node data fields
- Query-time boosting options available in the UI
- Dismax handler used for easy keyword searching and per-field boosts
- Add a Drupal modules for attachment indexing
- Another module for multi-site search

Here's a quick look at the admin interface:

- Apache Solr: Your site has contacted the Apache Solr server.
- Apache Solr PHP Client Library: Correct version "Revision: 22".

**Solr host name:** \*

Host name of your Solr server, e.g. localhost or example.com.

**Solr port:** \*

Port on which the Solr server listens. The Jetty example server is 8983, while Tomcat is 8080 by default.

**Solr path:**

Path that identifies the Solr request handler to be used.

**Number of items to index per cron run:**

The maximum number of items indexed in each pass of a [cron maintenance task](#). If necessary, reduce the number of items to prevent timeouts and memory errors.

**Results per page:**

The number of results that will be shown per page.

**On failure:**

What to display if Apache Solr search is not available.

**[Add a new content recommendation block](#)**

You currently have 1 block.

—▷ [Advanced configuration](#)

 **Enable spellchecker and suggestions**

Enable spellchecker and get word suggestions. Also known as the "Did you mean ... ?" feature.

## Apache Solr

[Settings](#)[Search Index](#)[Enabled filters](#)[Content bias settings](#)[Search fields](#)

The search index is generated by [running cron](#). 100% of the site content has been sent to the server. There are 0 items left to send.

Using schema.xml version: **drupal-0.9.4**

Solr core name: **ad**

*The server has a 30 sec delay before updates are processed.*

Number of documents in index: 1486

Number of pending deletions: 0

[View more details on the search index contents](#)

### Index controls

Re-index all content

Re-indexing will add all content to the index again (overwriting the index), but existing content in the index will remain searchable.

Delete the index

Deletes all of the documents in the Solr index. This is rarely necessary unless your index is corrupt or you have installed a new schema.xml.

# Apache Solr

[Settings](#)[Search index](#)[Enabled filters](#)[Content bias settings](#)[Search fields](#)

## Result biasing

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

### 'Sticky at top of lists' weight:

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

### 'Promoted to home page' weight:

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

### 'More recently created' bias:

This setting will change the result scoring so that nodes created more recently may appear before those with higher keyword matching.

### 'More comments' bias:

This setting will change the result scoring so that nodes with more comments may appear before those with higher keyword matching.

### 'More recent comments' bias:

This setting will change the result scoring so that nodes with the most recent comments (or most recent updates to the node itself) may appear before those with higher keyword matching.

▼ Type biasing and exclusion

**Blog type content bias:**

Ignore ▾

**Case study type content bias:**

8.0 ▾

**Page type content bias:**

Ignore ▾

**Story 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.

**Types to exclude from the search index:**

Blog

Case study

Page

Story

Specify here which node types should be totally excluded from the search index. Content excluded from the index will never appear in any search results.

Save configuration

Reset to defaults

▼ **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,

**Body text - the full, rendered content:**

1.0

**Author name:**

3.0

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

Omit

**Body text inside H1 tags:**

5.0

**Body text inside H2 or H3 tags:**

3.0

**Body text inside H4, H5, or H6 tags:**

2.0

**Body text in inline tags like EM or STRONG:**

1.0

**All taxonomy term names:**

2.0

**Title:**

5.0

# Drupal Modules Implement *hooks* to Control Indexing and Retrieval of Data

```
hook_apachesolr_update_index(&$document, $node,  
$namespace)
```

- By creating a Drupal module (in PHP), you can implement module and theme hooks to extend or alter Drupal behavior.
- Change or replace the data normally indexed
- Modify the search results and their appearance

# Image Data Using Dynamic Fields

```
/**
 * Implementation of hook_apachesolr_update_index().
 */
function apachesolr_image_apachesolr_update_index(&$document, $node, $namespace) {
  if ($node->type == 'image' && $document->entity == 'node') {
    $areas = array();
    $sizes = image_get_derivative_sizes($node->images['_original']);
    foreach ($sizes as $name => $info) {
      $areas[$name] = $info['width'] * $info['height'];
    }
    asort($areas);
    $image_path = FALSE;
    foreach ($areas as $preset => $size) {
      $image_path = $node->images[$preset];
      break;
    }
    if ($image_path) {
      $document->ss_image_relative = $image_path;
      // Support multi-site too.
      $document->ss_image_absolute = file_create_url($image_path);
    }
  }
}

/**
 * Implementation of hook_apachesolr_modify_query().
 */
function apachesolr_image_apachesolr_modify_query(&$query, &$params, $caller) {
  // Also retrieve image thumbnail links.
  $params['fl'] .= ',ss_image_relative';
}
```

# Image Data Using Dynamic Fields

```
if ($image_path) {
    $document->ss_image_relative =
$image_path;
}

/**
 * Implement hook_apachesolr_modify_query().
 */
function
apachesolr_image_apachesolr_modify_query(
&$query, &$params, $caller) {
    // Also retrieve image thumbnail links.
    $params['fl'] .= ',ss_image_relative';
}
```

## To Wrap Up

- Drupal has extensive Apache Solr integration already, and is highly customizable.
- The Drupal platform is widely adopted, and the Drupal community drives rapid innovation.
- Acquia provides Enterprise Drupal support and a network of partners.
- Acquia includes a secure, hosted Solr index with every support subscription.

## Resources ... Questions?

- <http://drupal.org/project/apachesolr>
- [http://drupal.org/project/apachesolr\\_attachments](http://drupal.org/project/apachesolr_attachments)
- <http://drupal.org/project/luceneapi> (pure PHP)
- SF video: <http://www.archive.org/details/ApacheSolrSearchMastery>
- <http://acquia.com/category/tags/apachesolr>
- <http://groups.drupal.org/lucene-nutch-and-solr>

# Drupal Adapts to You!

