Drupal Search: Where are we? Where are we going?

By Robert Douglass

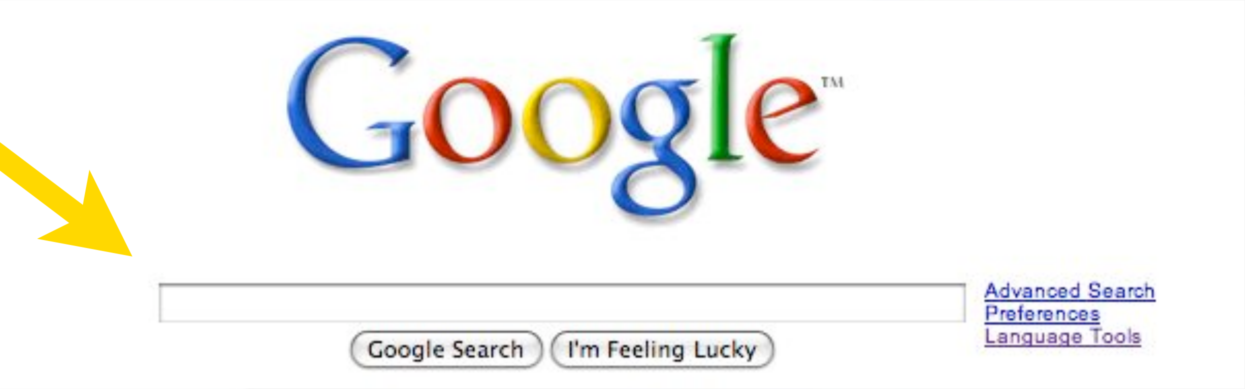
August 28, 2008
Drupalcon Szeged

What is search?

What is search?

- People think of search in a narrow box.

This one!



- (Albeit a very useful and important box)

Why is search important?

- Information retrieval.
- Find documentation or projects on Drupal.org
- Find the product you are trying to sell.
- Part of filter/sort/drill down data mining.
- Part of Mashups.

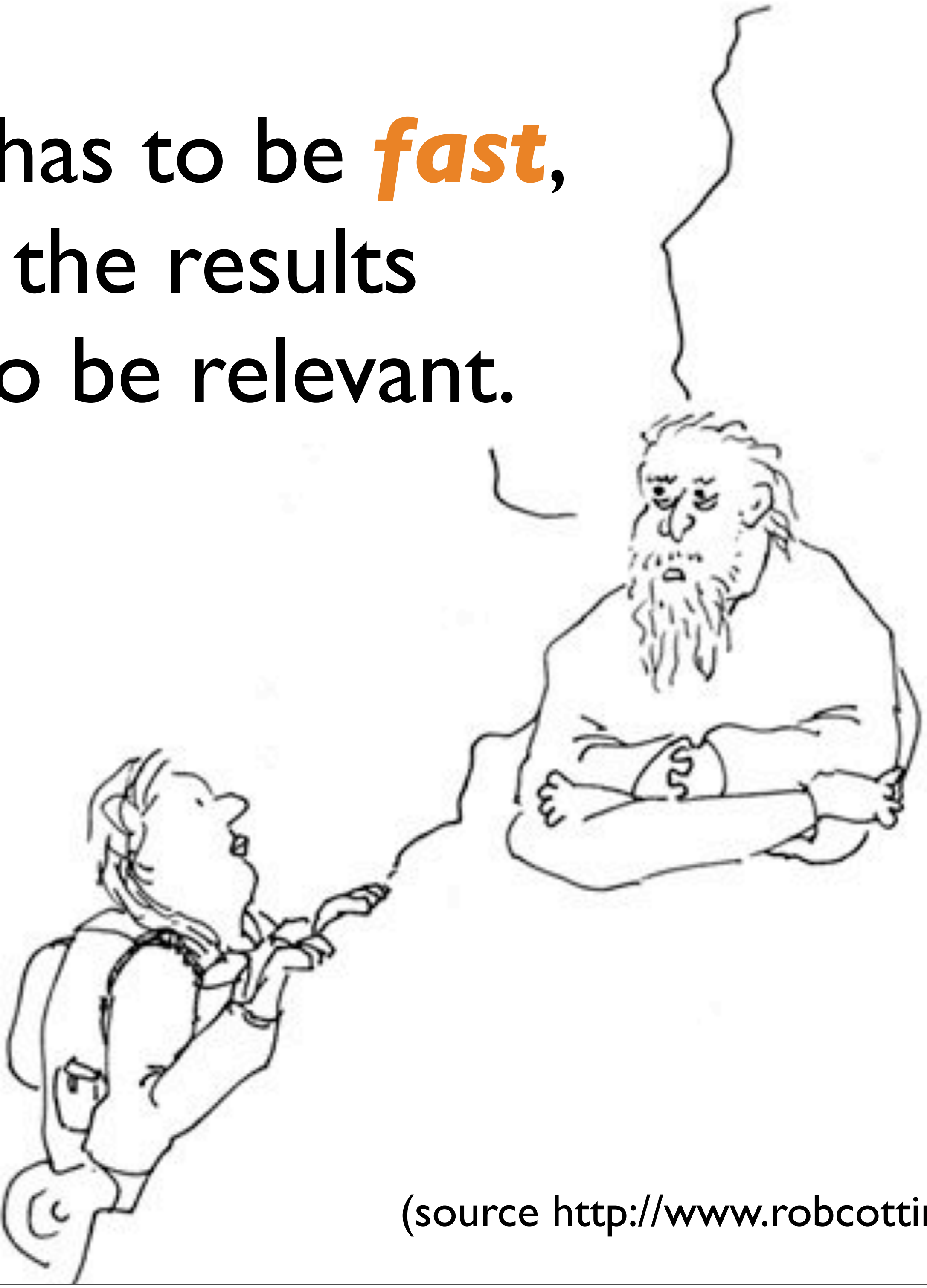
What do we search?

- Nodes
- External data sources
- Uploaded files
- Across many websites (federated search)
- Across many data sources

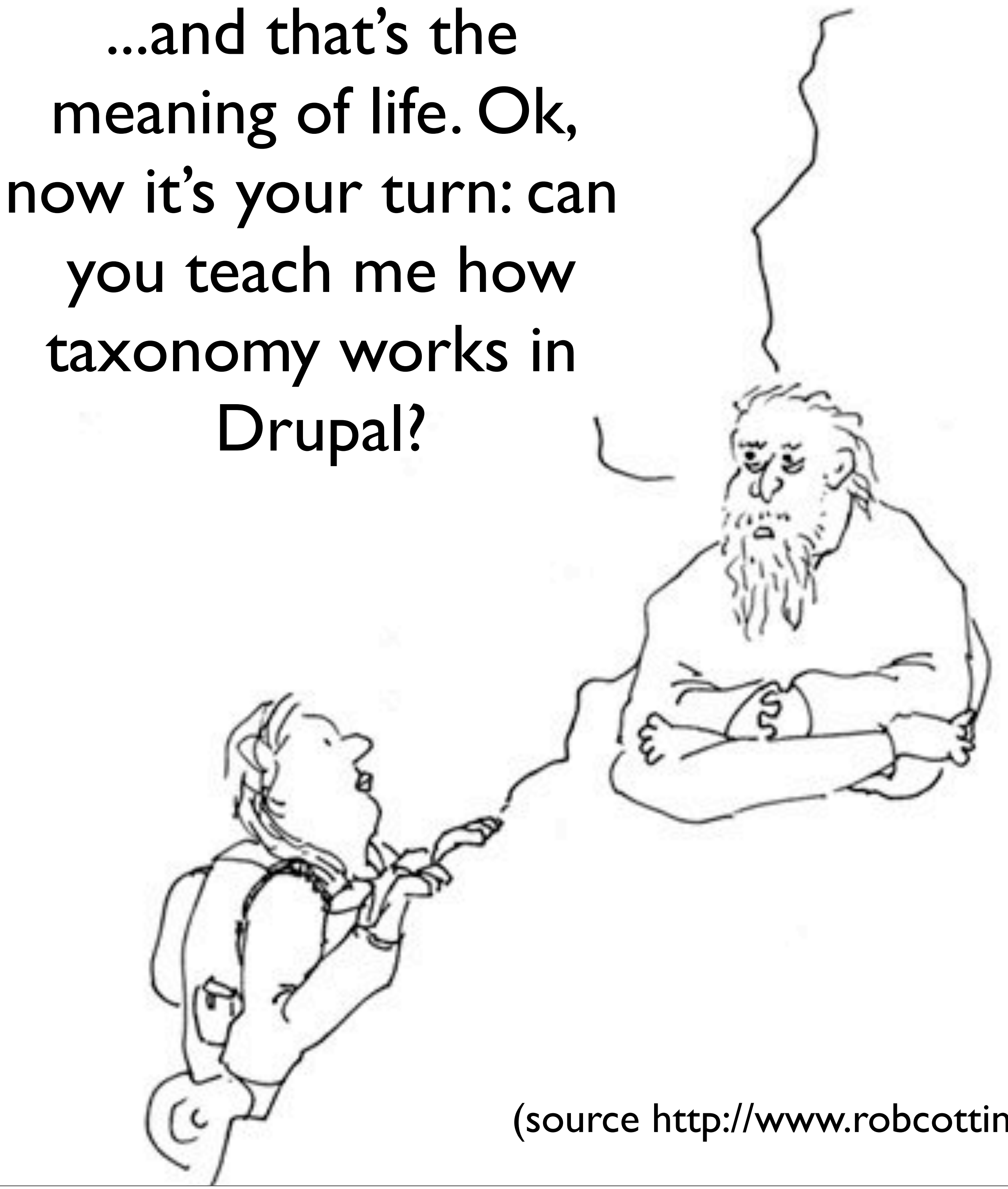
How do we search?

- Drupal core search
- Apache Solr
- Xapian
- Sphinx
- SPARQL

Search has to be ***fast***,
and the results
have to be relevant.



(source http://www.robcottingham.ca)


...and that's the meaning of life. Ok, now it's your turn: can you teach me how taxonomy works in Drupal?

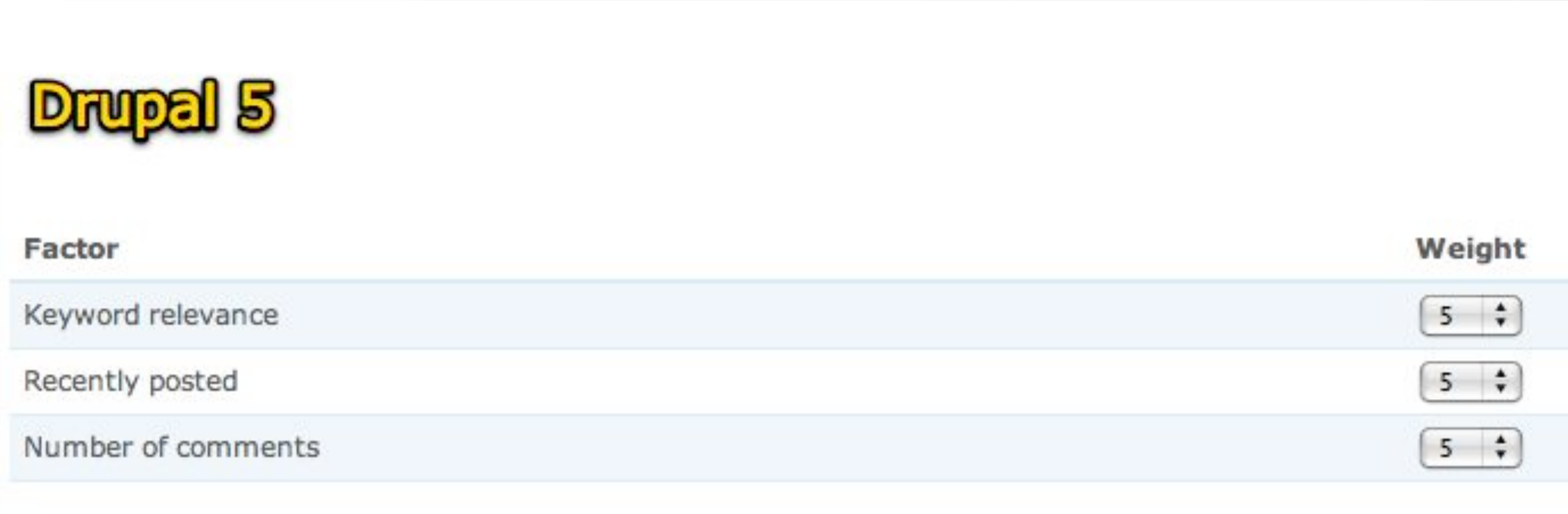
(source http://www.robcottingham.ca)

Minnesota Search Sprint

Goals were:

- Faster core search
- More features: language aware search
- Better relevancy: more scoring factors and open framework for scoring factors
- Code cleanup
- More reusable code for 3rd party search implementations
- Allow core searches to be turned off.

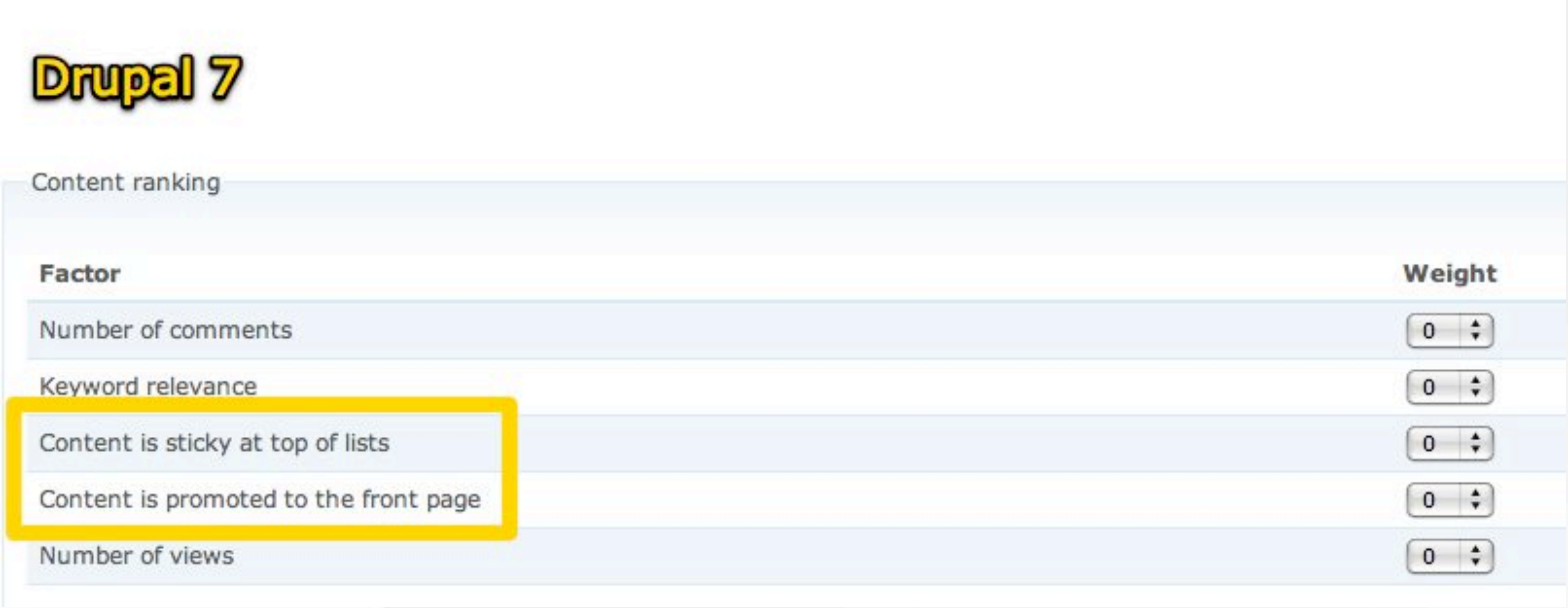
Scoring factors in Drupal 5



Drupal 5

Factor	Weight
Keyword relevance	5
Recently posted	5
Number of comments	5

Scoring factors in Drupal 7



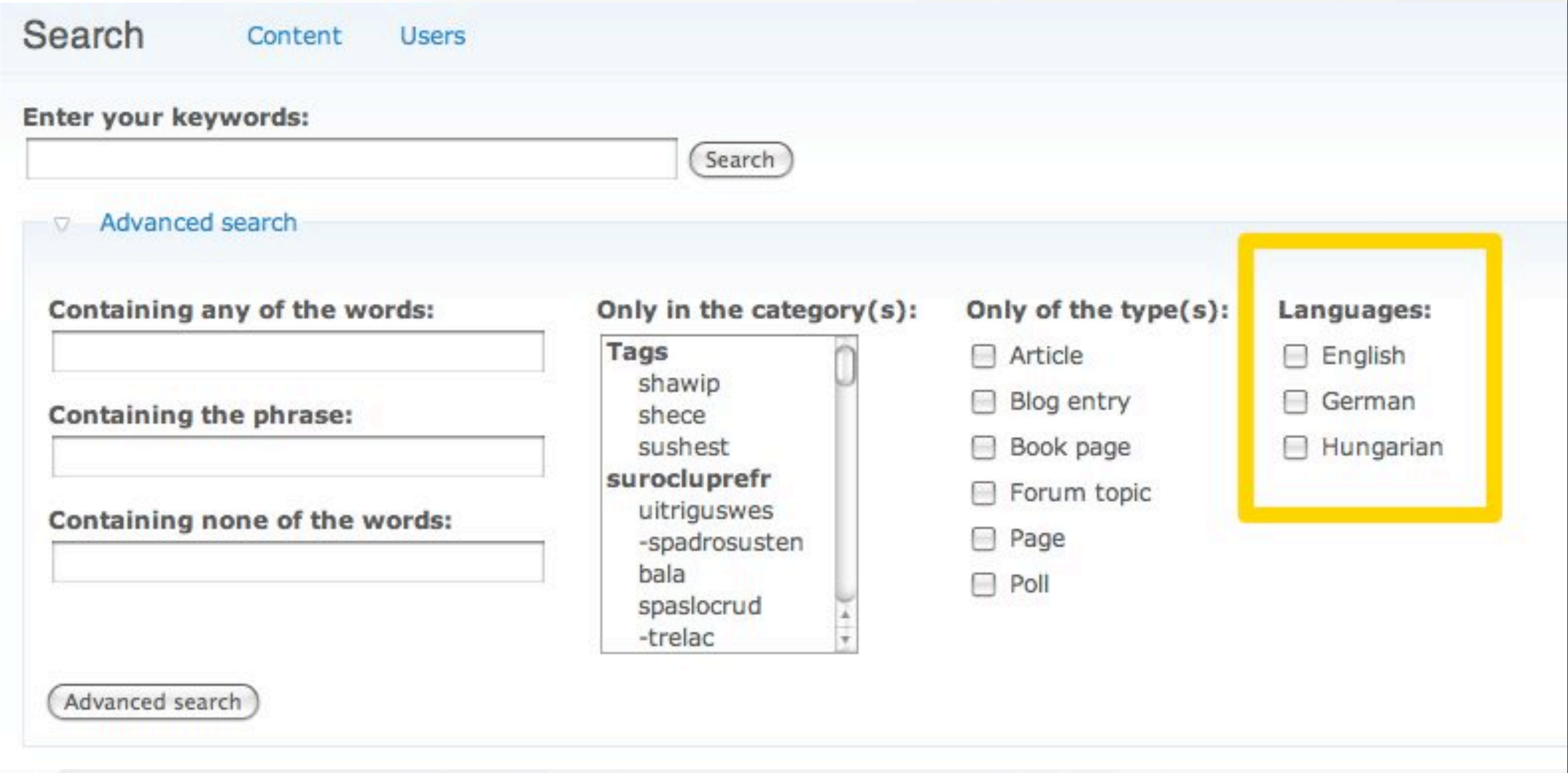
Drupal 7

Content ranking

Factor	Weight
Number of comments	0
Keyword relevance	0
Content is sticky at top of lists	0
Content is promoted to the front page	0
Number of views	0

Plus, modules can now add their own scoring factors.

Search in specific languages



Search   Content   Users

Enter your keywords:

Search

▽ Advanced search

Containing any of the words:

Containing the phrase:

Containing none of the words:

Only in the category(s):

Tags
- shawip
- shece
- sushest

suroclupref r
- uitriguswes
- -spadrosusten
- bala
- spaslocrud
- -trelac

Only of the type(s):

□ Article
□ Blog entry
□ Book page
□ Forum topic
□ Page
□ Poll

Languages:

□ English
□ German
□ Hungarian

Advanced search

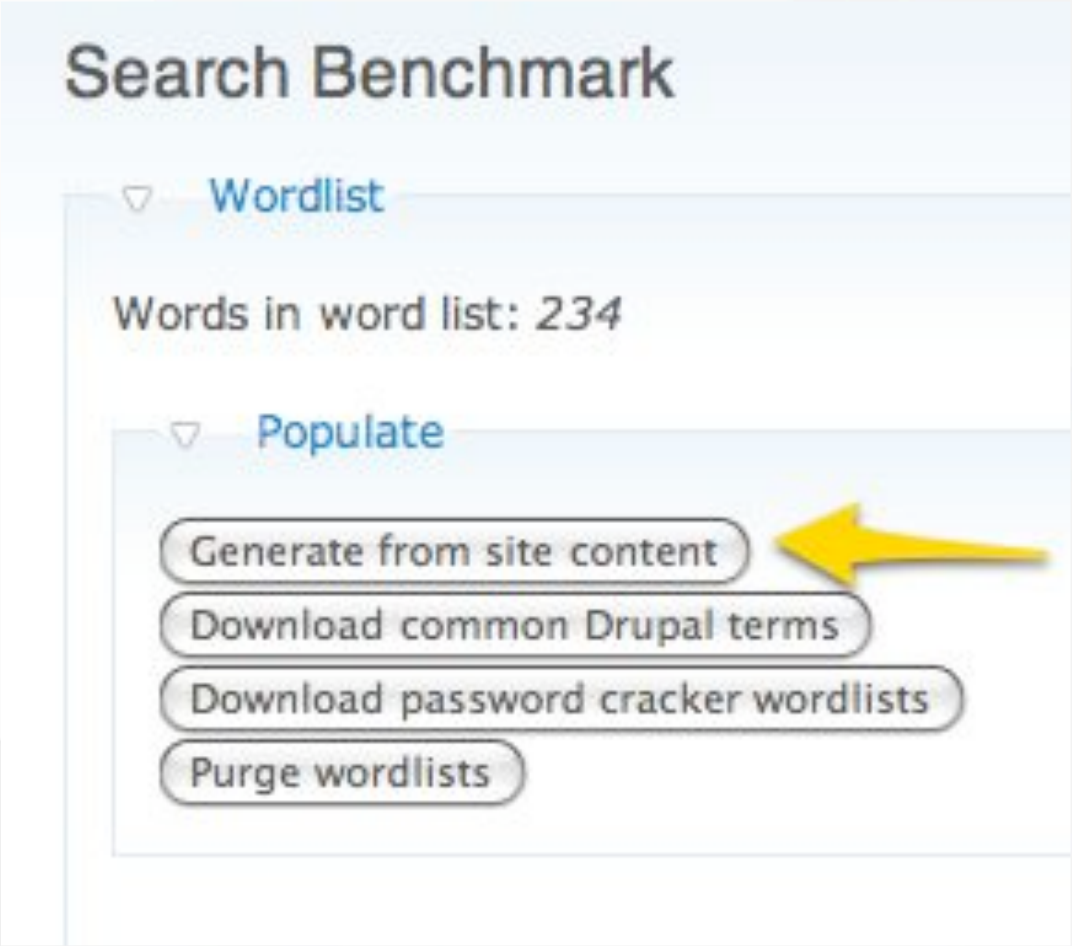
SearchBench Module

- By Jeremy Andrews (Tag1 Consulting)

  http://tag1consulting.com/

- http://drupal.org/project/searchbench

Generate a wordlist



Search Benchmark

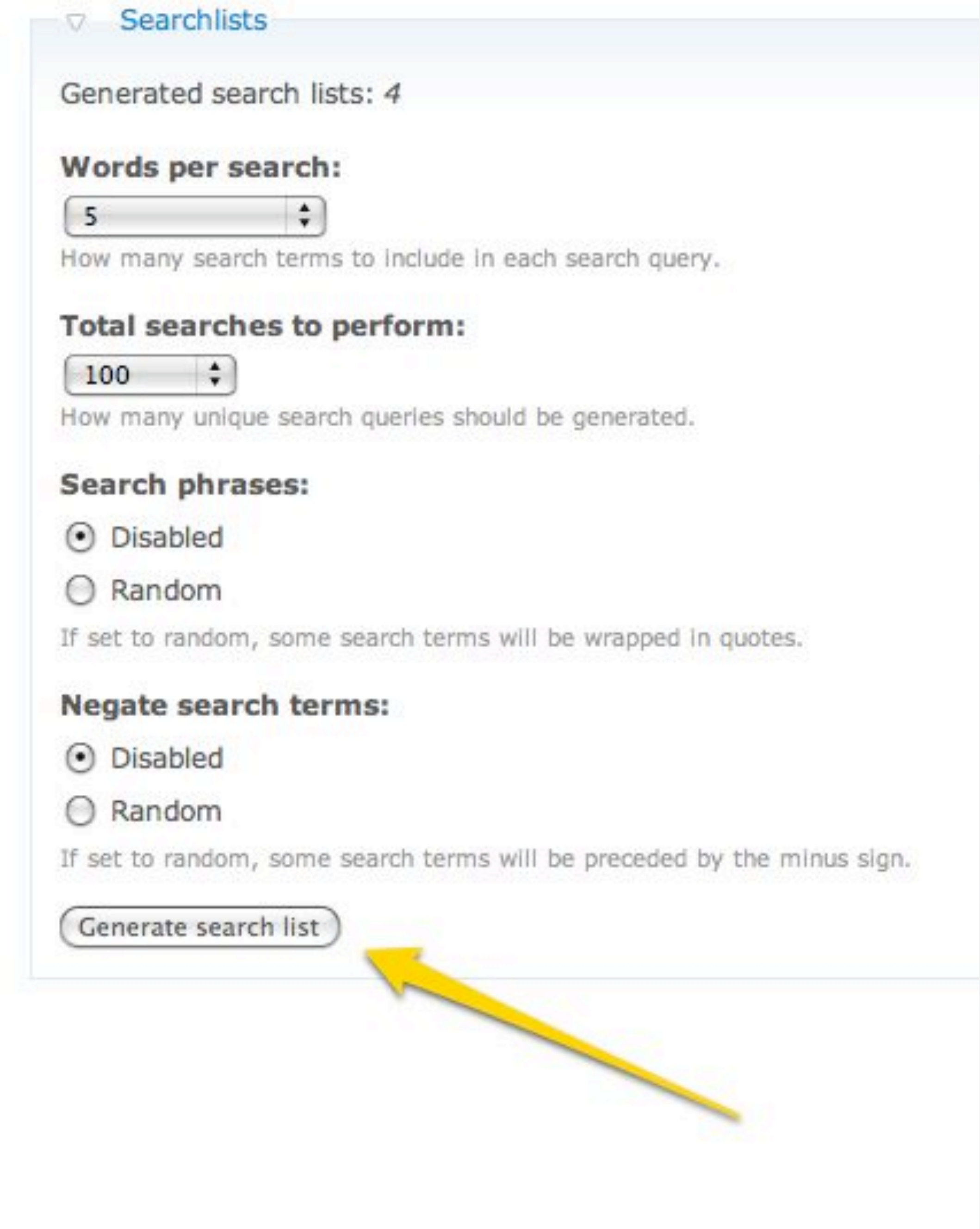
▽ Wordlist

Words in word list: *234*

▽ Populate

Generate from site content
Download common Drupal terms
Download password cracker wordlists
Purge wordlists

Generate a search list



▽ Searchlists

Generated search lists: *4*

Words per search:

5 ⬍

How many search terms to include in each search query.

Total searches to perform:

100 ⬍

How many unique search queries should be generated.

Search phrases:

● Disabled

○ Random

If set to random, some search terms will be wrapped in quotes.

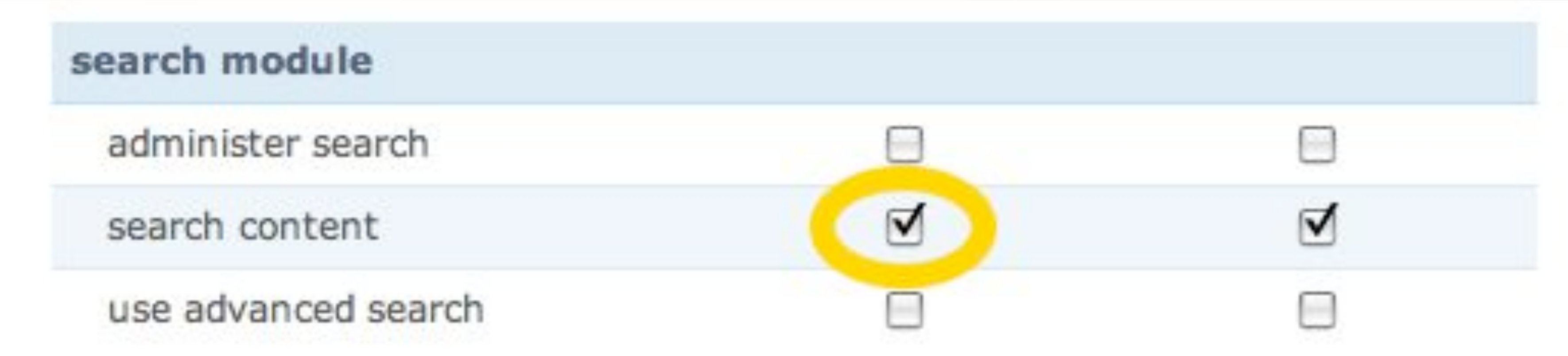
Negate search terms:

● Disabled

○ Random

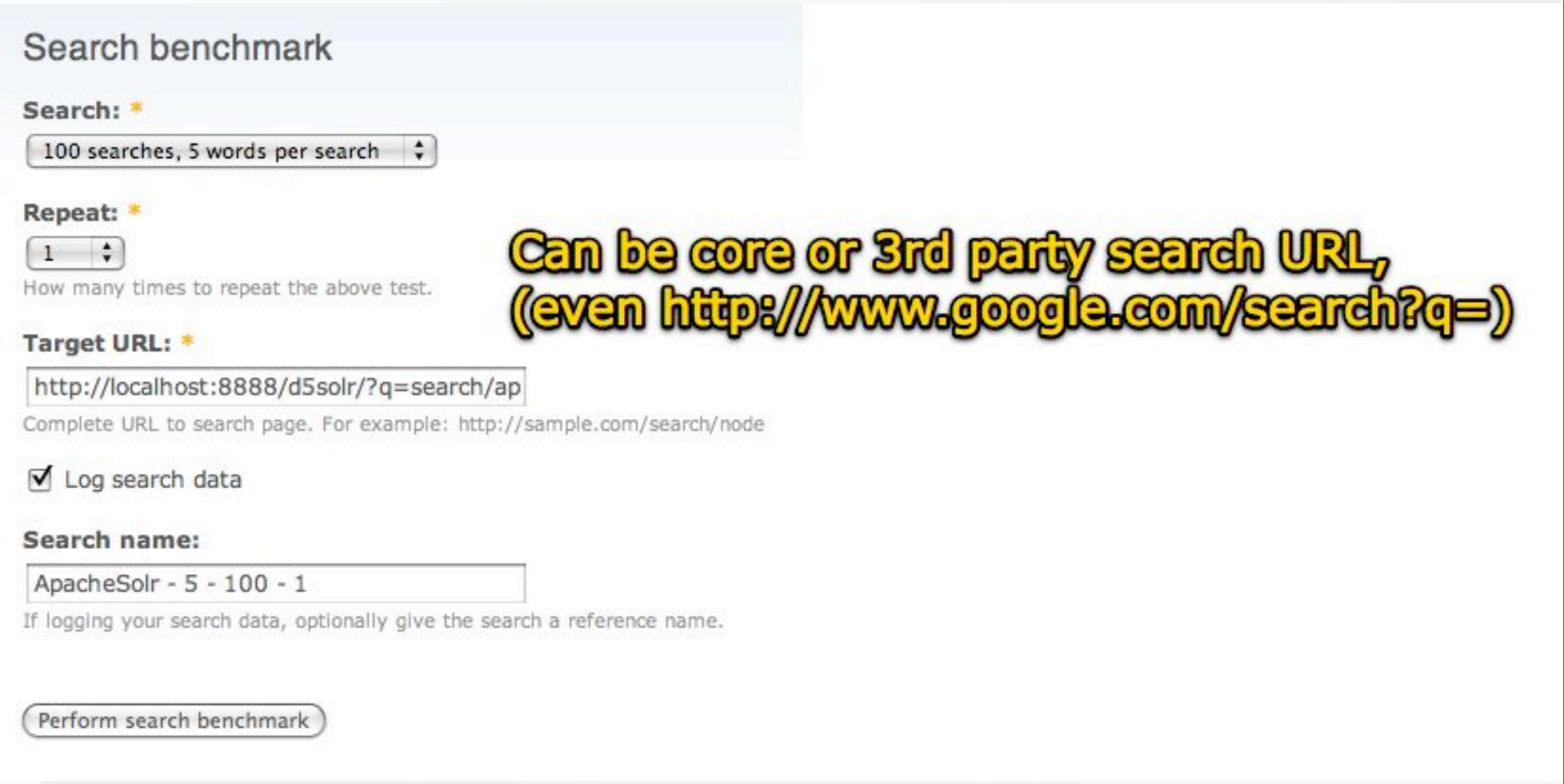
If set to random, some search terms will be preceded by the minus sign.

Set access permissions



search module		
administer search	□	□
search content	☑	☑
use advanced search	□	□

Run a search list



Search benchmark

Search: *

100 searches, 5 words per search

Repeat: *

1

How many times to repeat the above test.

Target URL: *

http://localhost:8888/d5solr/?q=search/ap

Complete URL to search page. For example: http://sample.com/search/node

☑ Log search data

Search name:

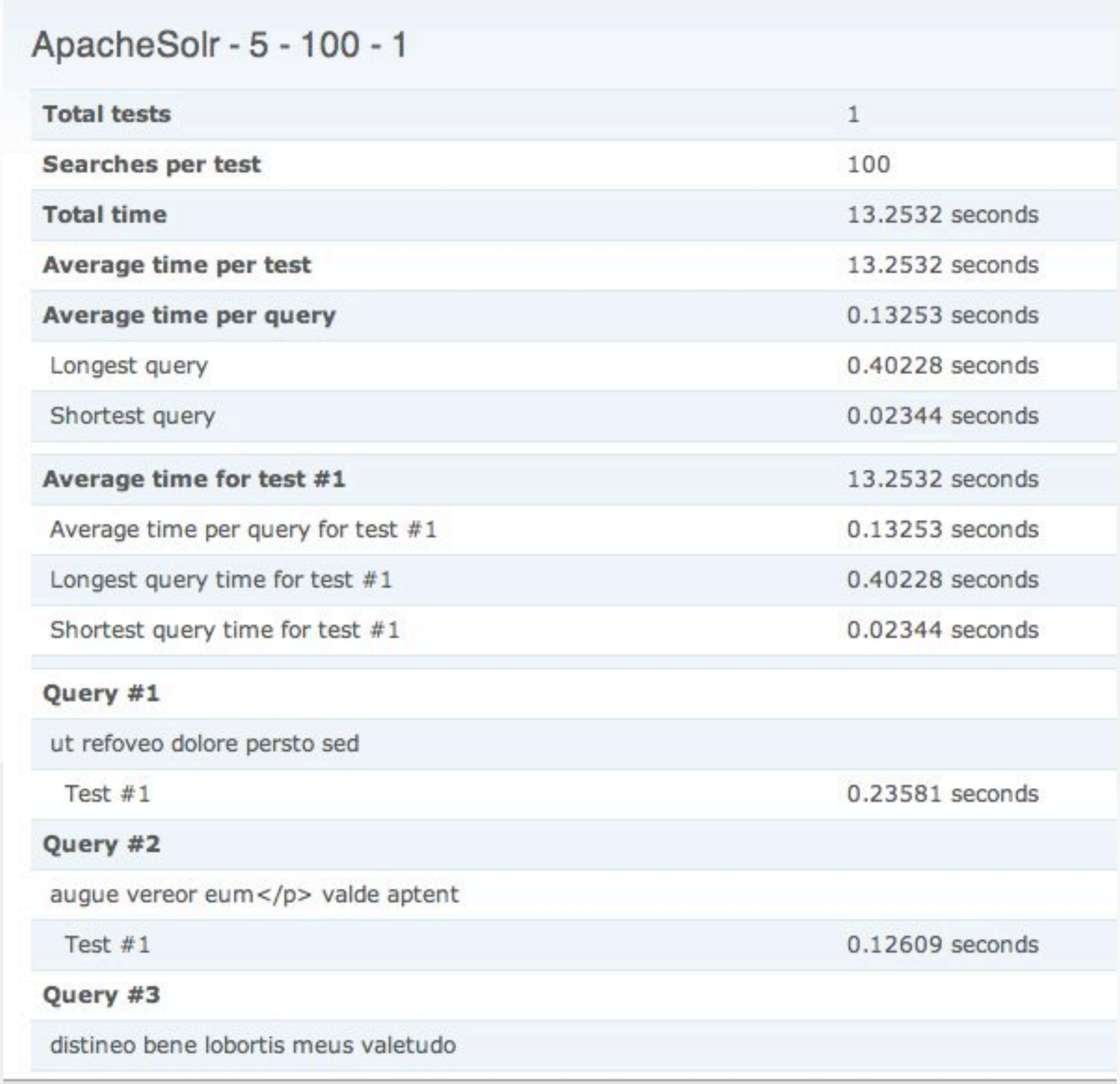
ApacheSolr - 5 - 100 - 1

If logging your search data, optionally give the search a reference name.

Perform search benchmark

Can be core or 3rd party search URL, (even http://www.google.com/search?q=)

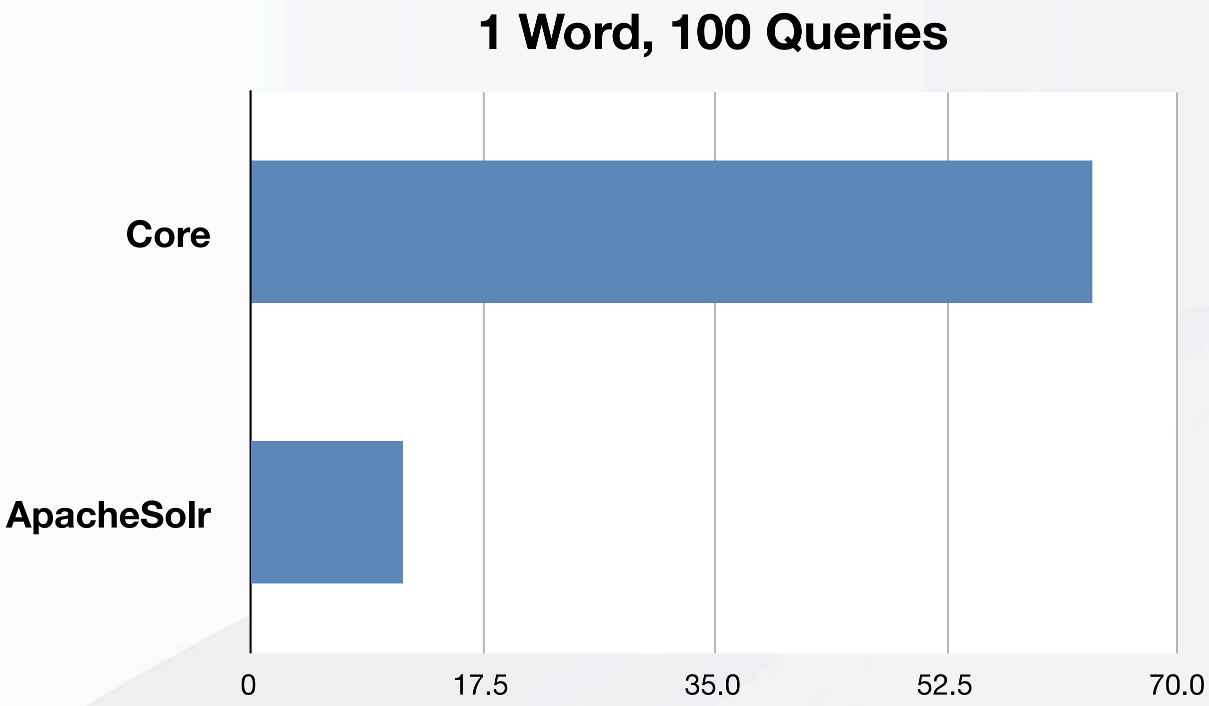
Generate search results



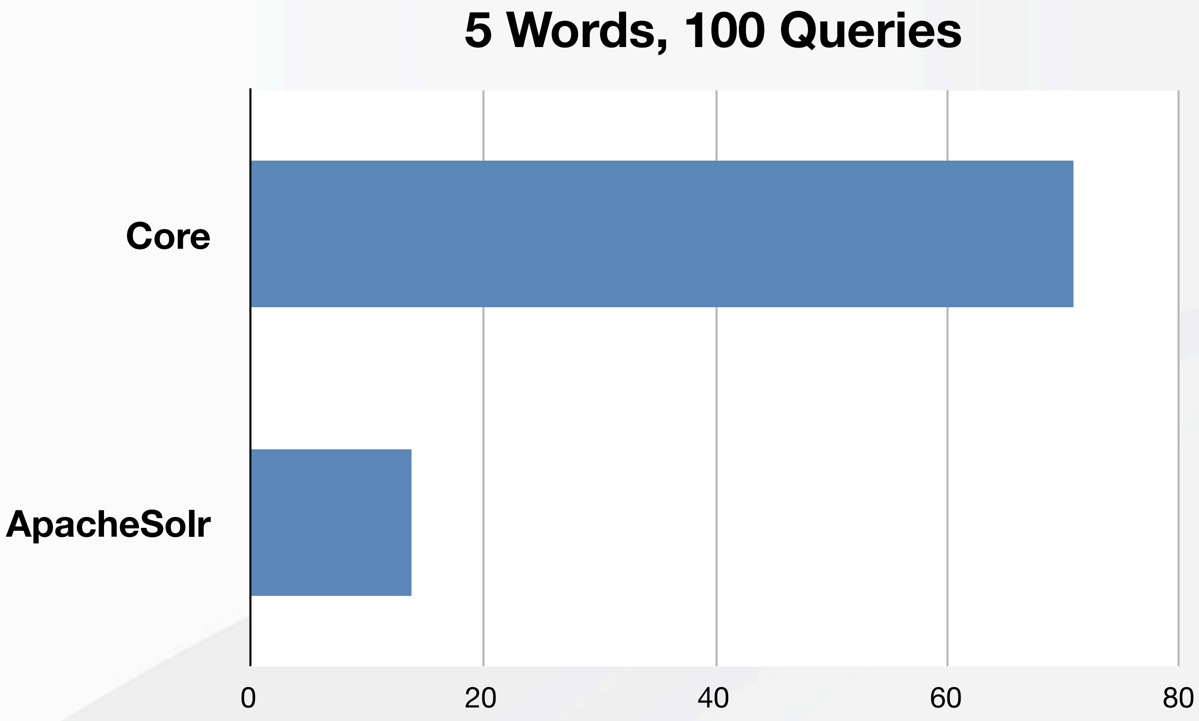
ApacheSolr - 5 - 100 - 1

Total tests	1
Searches per test	100
Total time	13.2532 seconds
Average time per test	13.2532 seconds
Average time per query	0.13253 seconds
Longest query	0.40228 seconds
Shortest query	0.02344 seconds
Average time for test #1	13.2532 seconds
Average time per query for test #1	0.13253 seconds
Longest query time for test #1	0.40228 seconds
Shortest query time for test #1	0.02344 seconds
Query #1	
ut refoveo dolore persto sed	
Test #1	0.23581 seconds
Query #2	
augue vereor eum</p> valde aptent	
Test #1	0.12609 seconds
Query #3	
distineo bene lobortis meus valetudo	

ApacheSolr is 5-6x faster

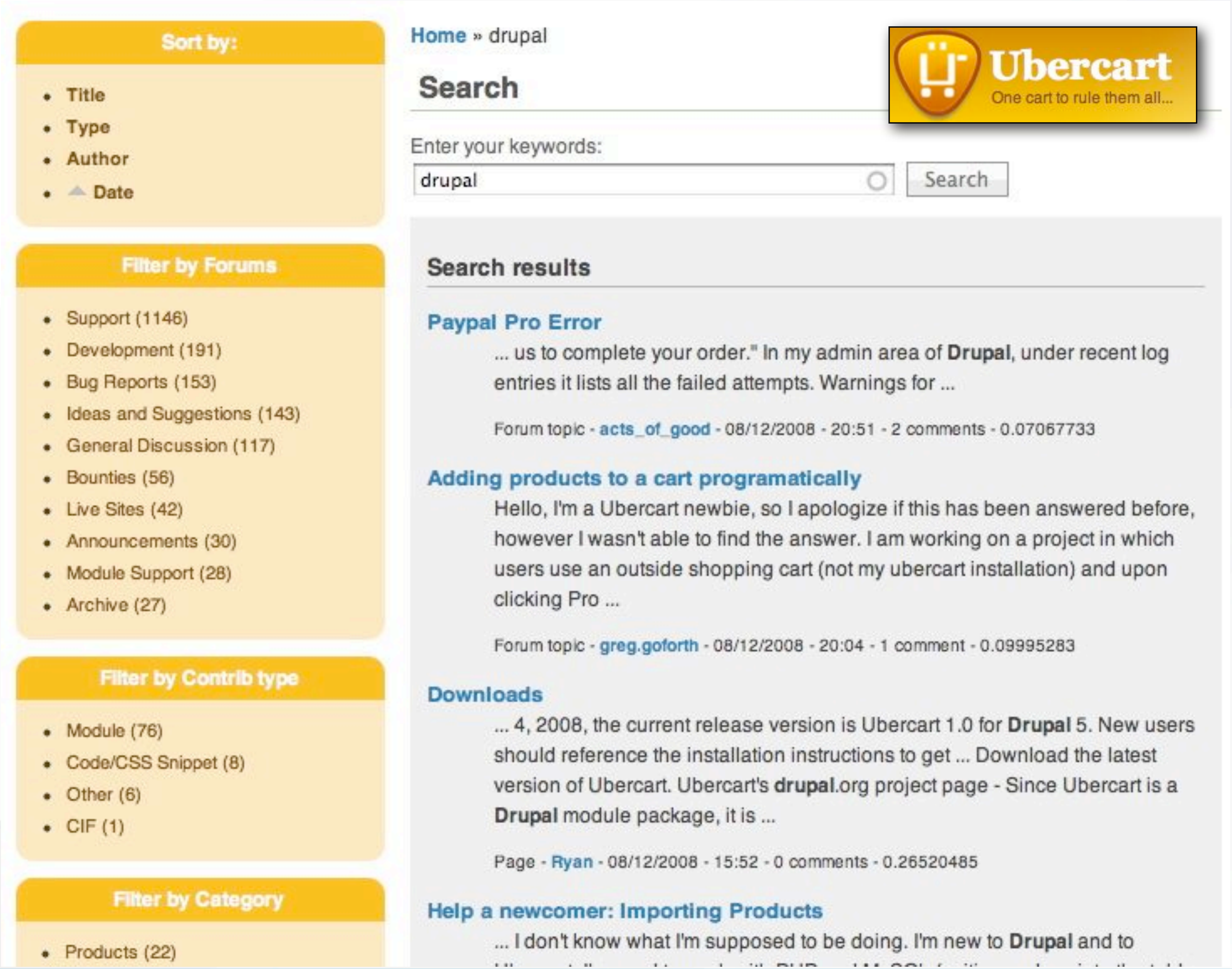


ApacheSolr is 5-6x faster



Solr search

- Java application (runs in Tomcat, Jetty, Resin, JBoss, etc.)
- Provides web service layer (indexing and searching) on top of Apache Lucene
- Faster indexing and searching than Drupal core
- Faceted search
- Support for searching uploaded files
- Views filter integration



Sort by:

- Title
- Type
- Author
- ▲ Date

Filter by Forums

- Support (1146)
- Development (191)
- Bug Reports (153)
- Ideas and Suggestions (143)
- General Discussion (117)
- Bounties (56)
- Live Sites (42)
- Announcements (30)
- Module Support (28)
- Archive (27)

Filter by Contrib type

- Module (76)
- Code/CSS Snippet (8)
- Other (6)
- CIF (1)

Filter by Category

- Products (22)

Home » drupal

Search

Enter your keywords:

drupal

Search

Search results

Paypal Pro Error

... us to complete your order." In my admin area of **Drupal**, under recent log entries it lists all the failed attempts. Warnings for ...

Forum topic - acts_of_good - 08/12/2008 - 20:51 - 2 comments - 0.07067733

Adding products to a cart programatically

Hello, I'm a Ubercart newbie, so I apologize if this has been answered before, however I wasn't able to find the answer. I am working on a project in which users use an outside shopping cart (not my ubercart installation) and upon clicking Pro ...

Forum topic - greg.goforth - 08/12/2008 - 20:04 - 1 comment - 0.09995283

Downloads

... 4, 2008, the current release version is Ubercart 1.0 for **Drupal** 5. New users should reference the installation instructions to get ... Download the latest version of Ubercart. Ubercart's **drupal**.org project page - Since Ubercart is a **Drupal** module package, it is ...

Page - Ryan - 08/12/2008 - 15:52 - 0 comments - 0.26520485

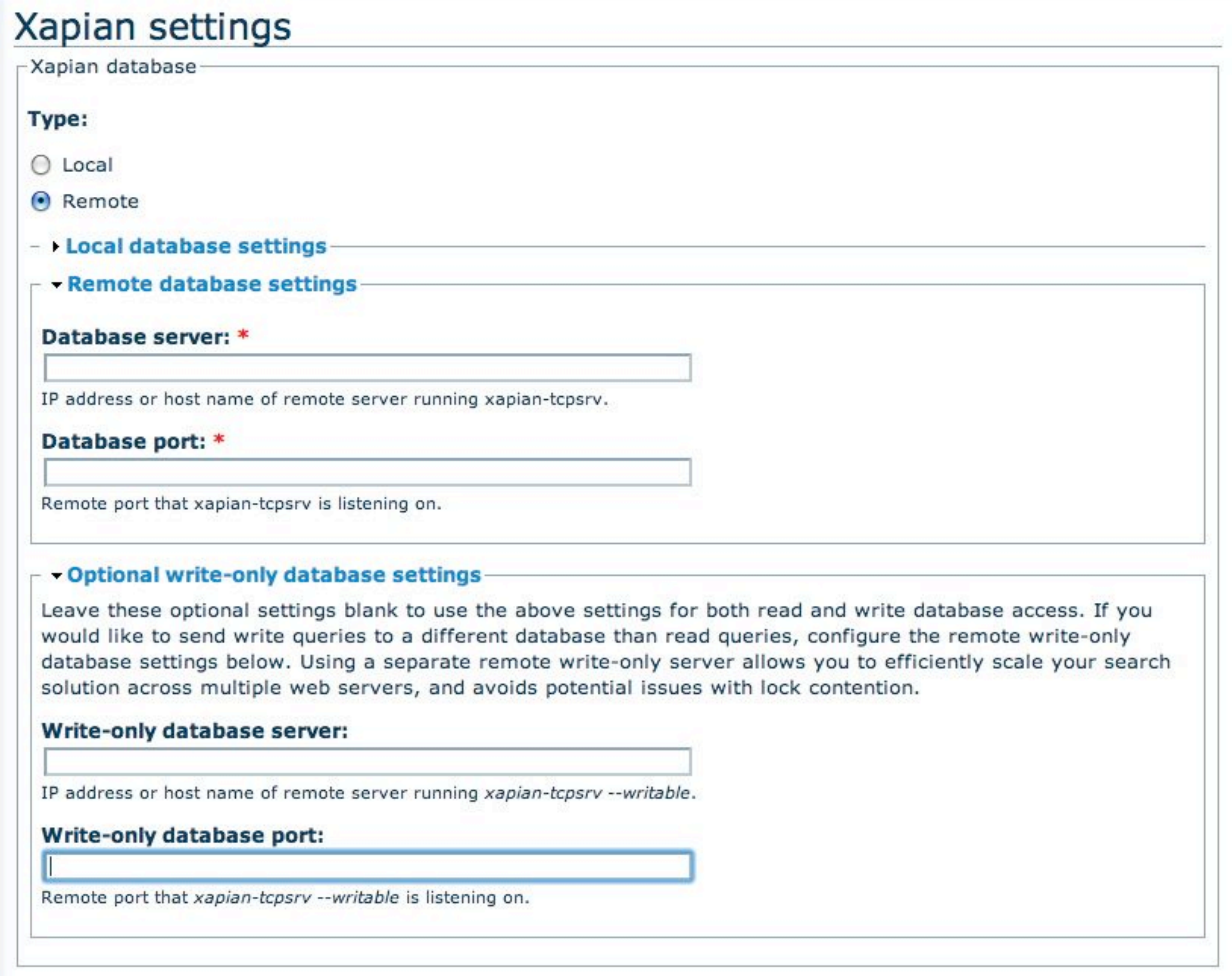
Help a newcomer: Importing Products

... I don't know what I'm supposed to be doing. I'm new to **Drupal** and to

Xapian

- In use on Drupal.org
- Written in C++ with bindings to most common languages.
- http://xapian.org
- http://drupal.org/project/xapian

Xapian settings #1



Xapian settings

Xapian database

Type:

○ Local

● Remote

▸ **Local database settings**

▾ **Remote database settings**

Database server: *

IP address or host name of remote server running xapian-tcpsrv.

Database port: *

Remote port that xapian-tcpsrv is listening on.

▾ **Optional write-only database settings**

Leave these optional settings blank to use the above settings for both read and write database access. If you would like to send write queries to a different database than read queries, configure the remote write-only database settings below. Using a separate remote write-only server allows you to efficiently scale your search solution across multiple web servers, and avoids potential issues with lock contention.

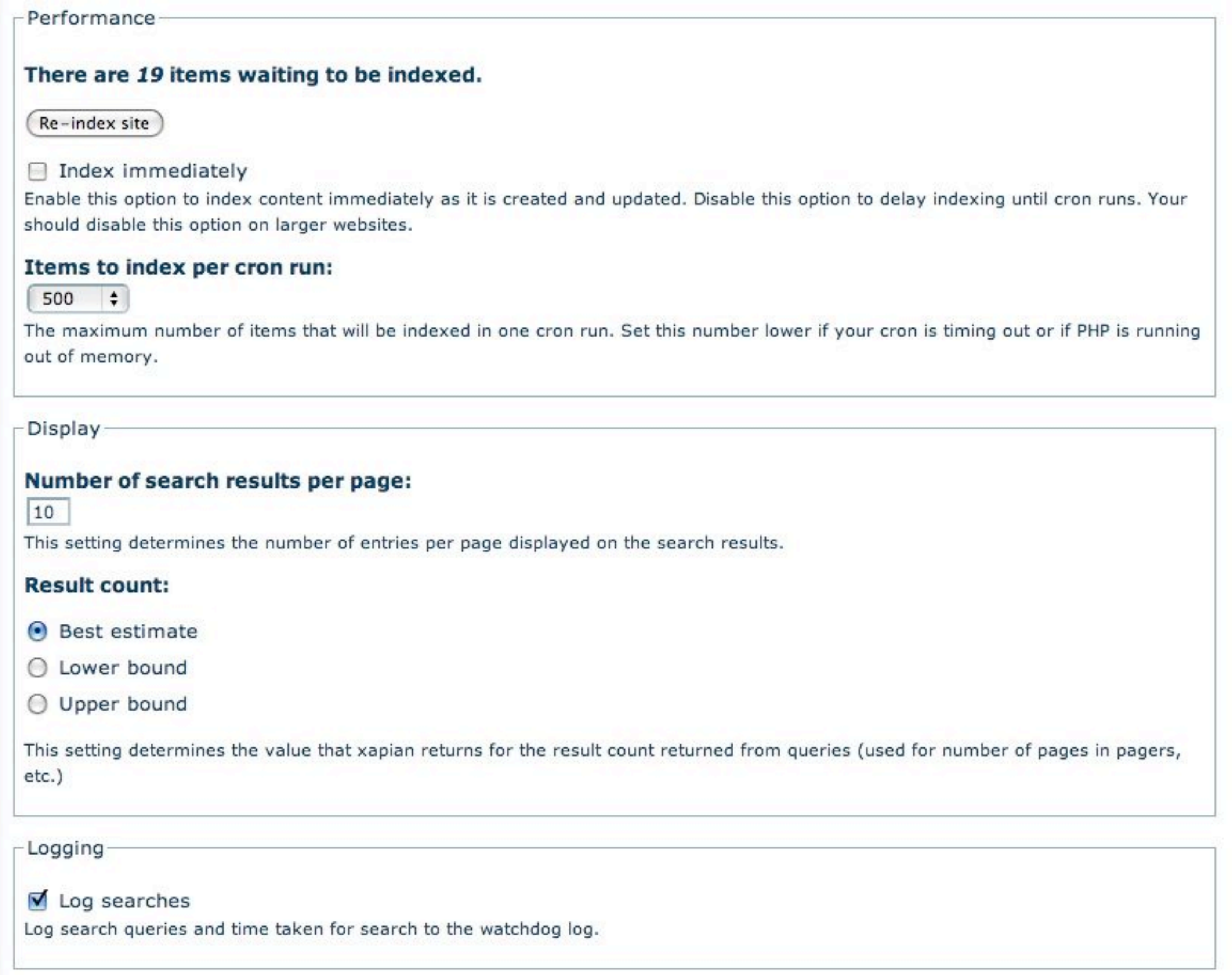
Write-only database server:

IP address or host name of remote server running *xapian-tcpsrv --writable*.

Write-only database port:

Remote port that *xapian-tcpsrv --writable* is listening on.

Xapian settings #2



Performance

There are *19* items waiting to be indexed.

□ Index immediately

Enable this option to index content immediately as it is created and updated. Disable this option to delay indexing until cron runs. Your should disable this option on larger websites.

Items to index per cron run:

500

The maximum number of items that will be indexed in one cron run. Set this number lower if your cron is timing out or if PHP is running out of memory.

Display

Number of search results per page:

10

This setting determines the number of entries per page displayed on the search results.

Result count:

● Best estimate
○ Lower bound
○ Upper bound

This setting determines the value that xapian returns for the result count returned from queries (used for number of pages in pagers, etc.)

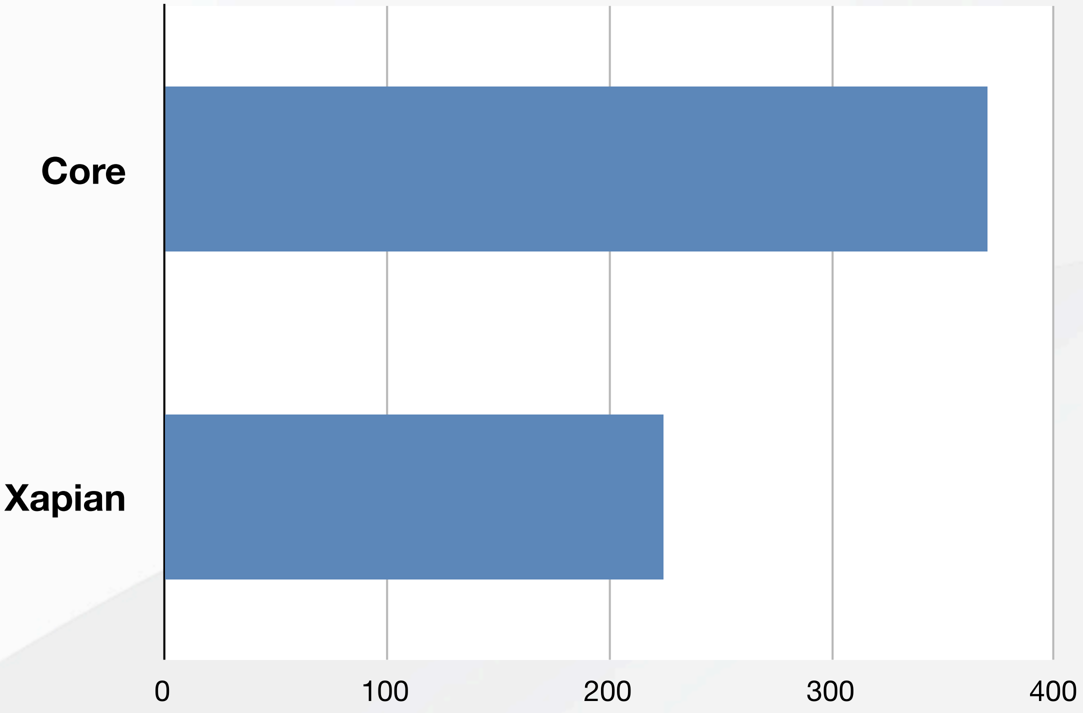
Logging

☑ Log searches

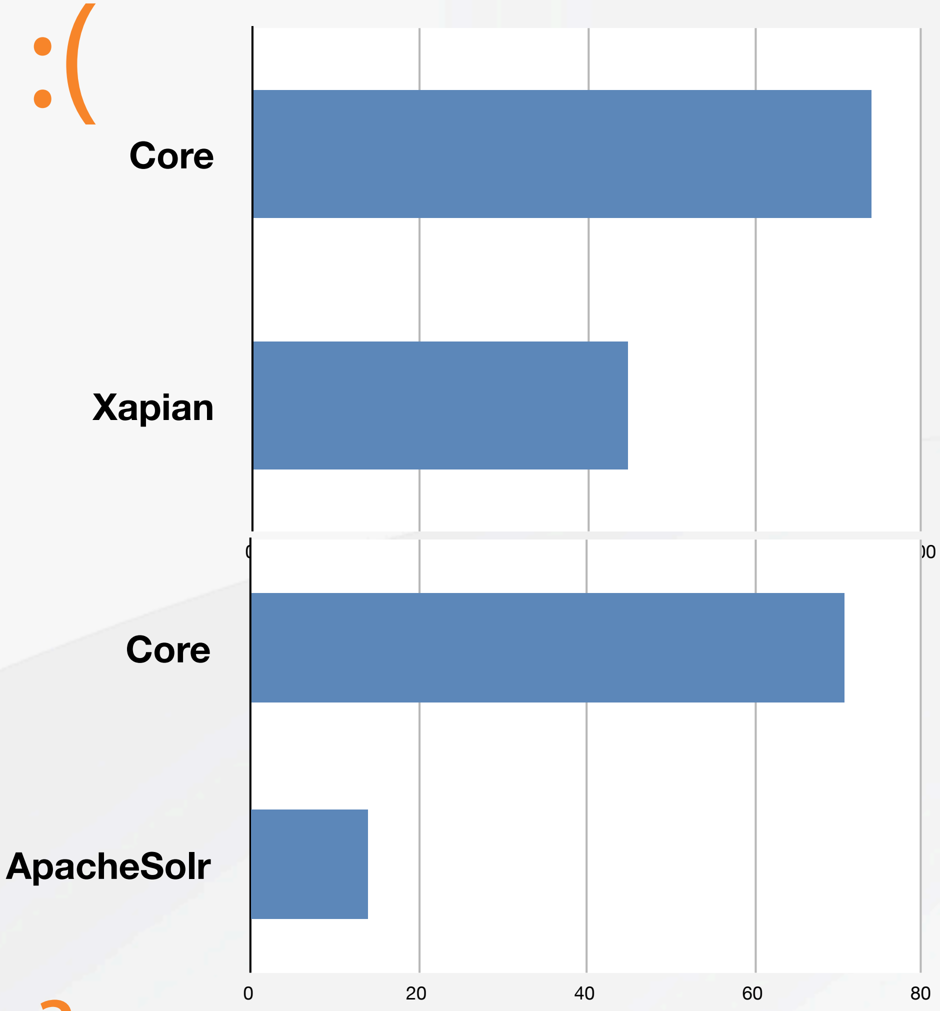
Log search queries and time taken for search to the watchdog log.

Xapian search is 40% faster

1-5 Words, no phrases, no negation



No direct Xapian vs Solr Benchmarks :(



Solr *maybe* faster?

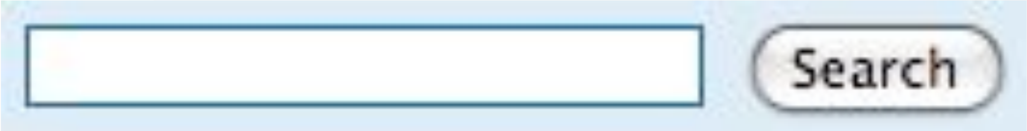
Sphinx

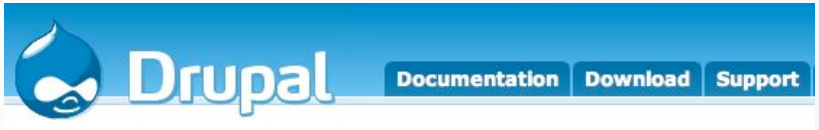
- Used on NowPublic.com, among others.
- http://www.sphinxsearch.com/
- http://drupal.org/project/sphinx

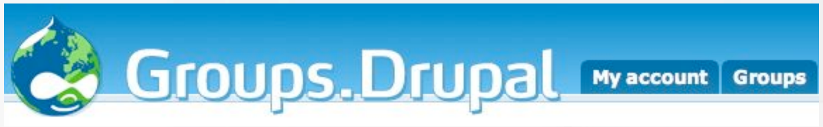
SPARQL search

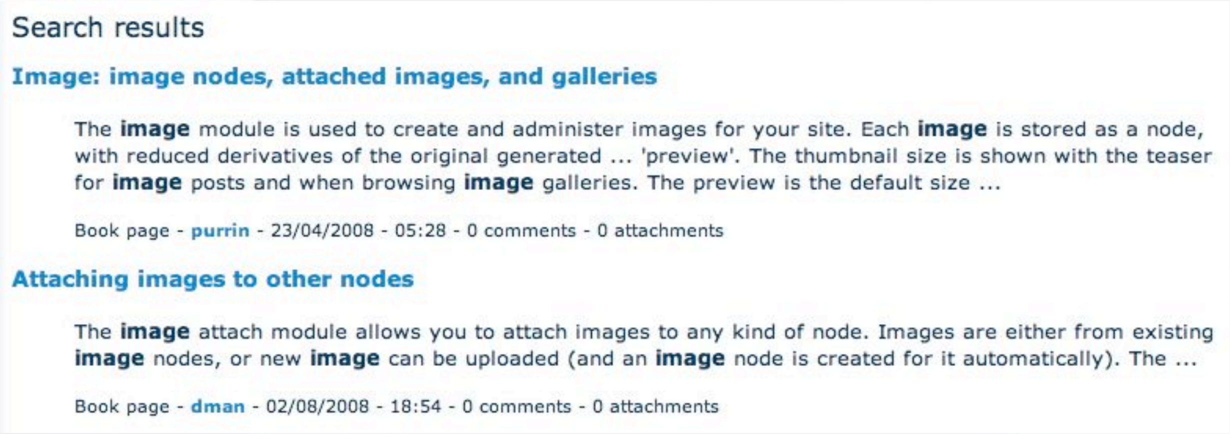
- Query local or remote RDF data stores
- Can combine heterogenous data sources as long as they share any connection on the graph.
- Is ideal for highly flexible federated search.

Some things that we’re missing









Search results

Image: image nodes, attached images, and galleries

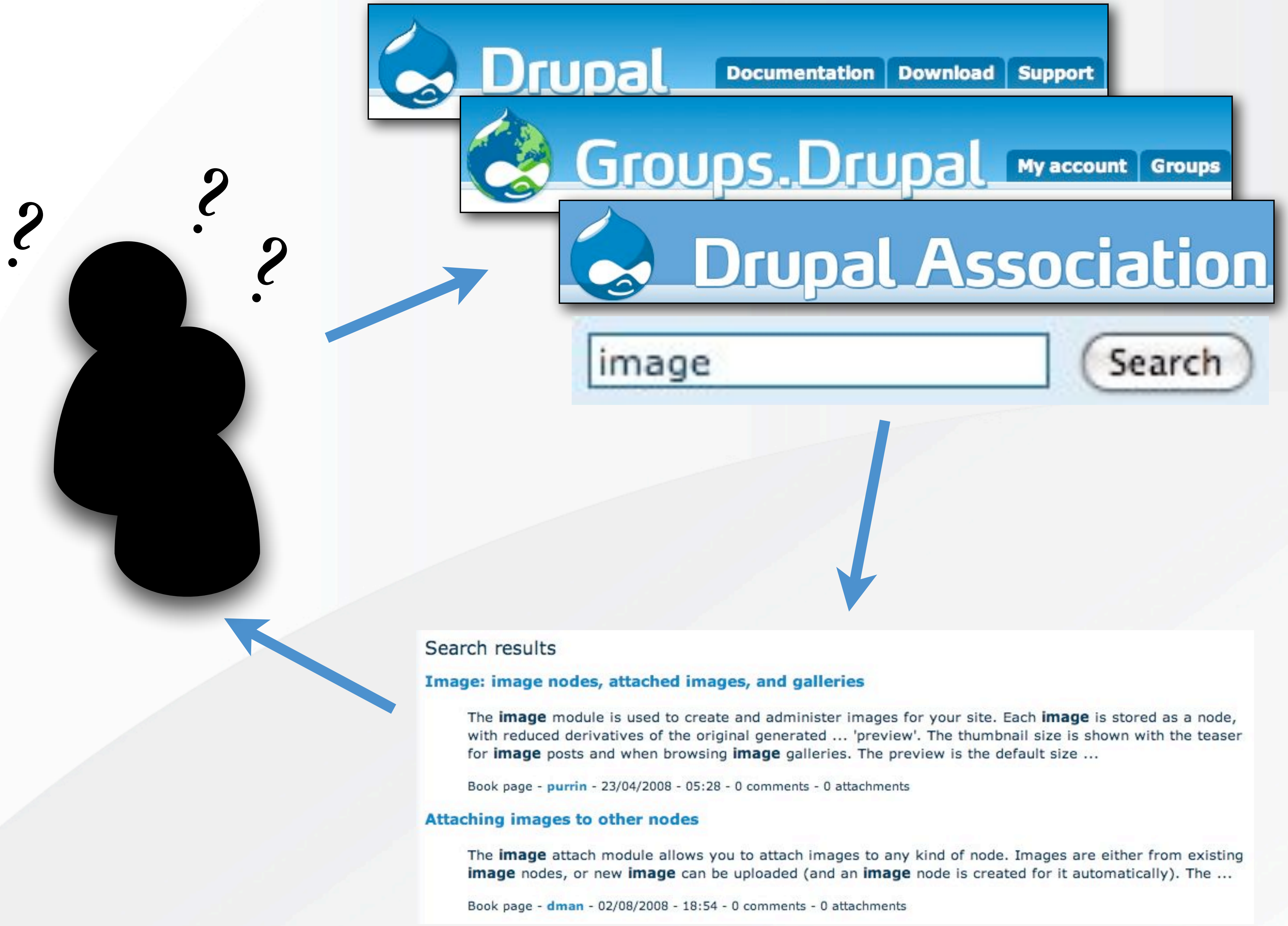
The **image** module is used to create and administer images for your site. Each **image** is stored as a node, with reduced derivatives of the original generated ... 'preview'. The thumbnail size is shown with the teaser for **image** posts and when browsing **image** galleries. The preview is the default size ...

Book page - **purrin** - 23/04/2008 - 05:28 - 0 comments - 0 attachments

Attaching images to other nodes

The **image** attach module allows you to attach images to any kind of node. Images are either from existing **image** nodes, or new **image** can be uploaded (and an **image** node is created for it automatically). The ...

Book page - **dman** - 02/08/2008 - 18:54 - 0 comments - 0 attachments



Search results

Image: image nodes, attached images, and galleries

The **image** module is used to create and administer images for your site. Each **image** is stored as a node, with reduced derivatives of the original generated ... 'preview'. The thumbnail size is shown with the teaser for **image** posts and when browsing **image** galleries. The preview is the default size ...

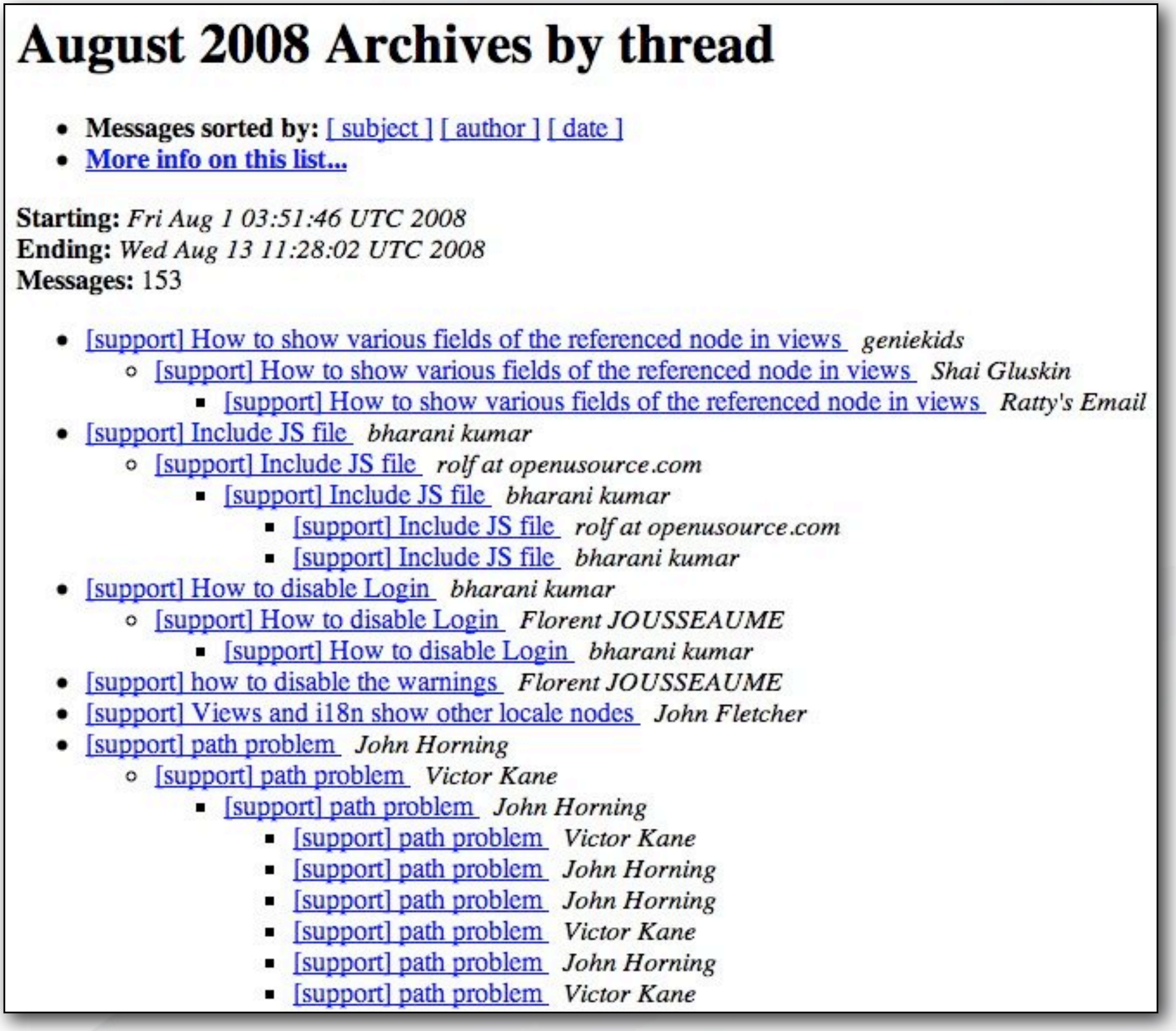
Book page - purrin - 23/04/2008 - 05:28 - 0 comments - 0 attachments

Attaching images to other nodes

The **image** attach module allows you to attach images to any kind of node. Images are either from existing **image** nodes, or new **image** can be uploaded (and an **image** node is created for it automatically). The ...

Book page - dman - 02/08/2008 - 18:54 - 0 comments - 0 attachments

We don’t search everything on Drupal.org

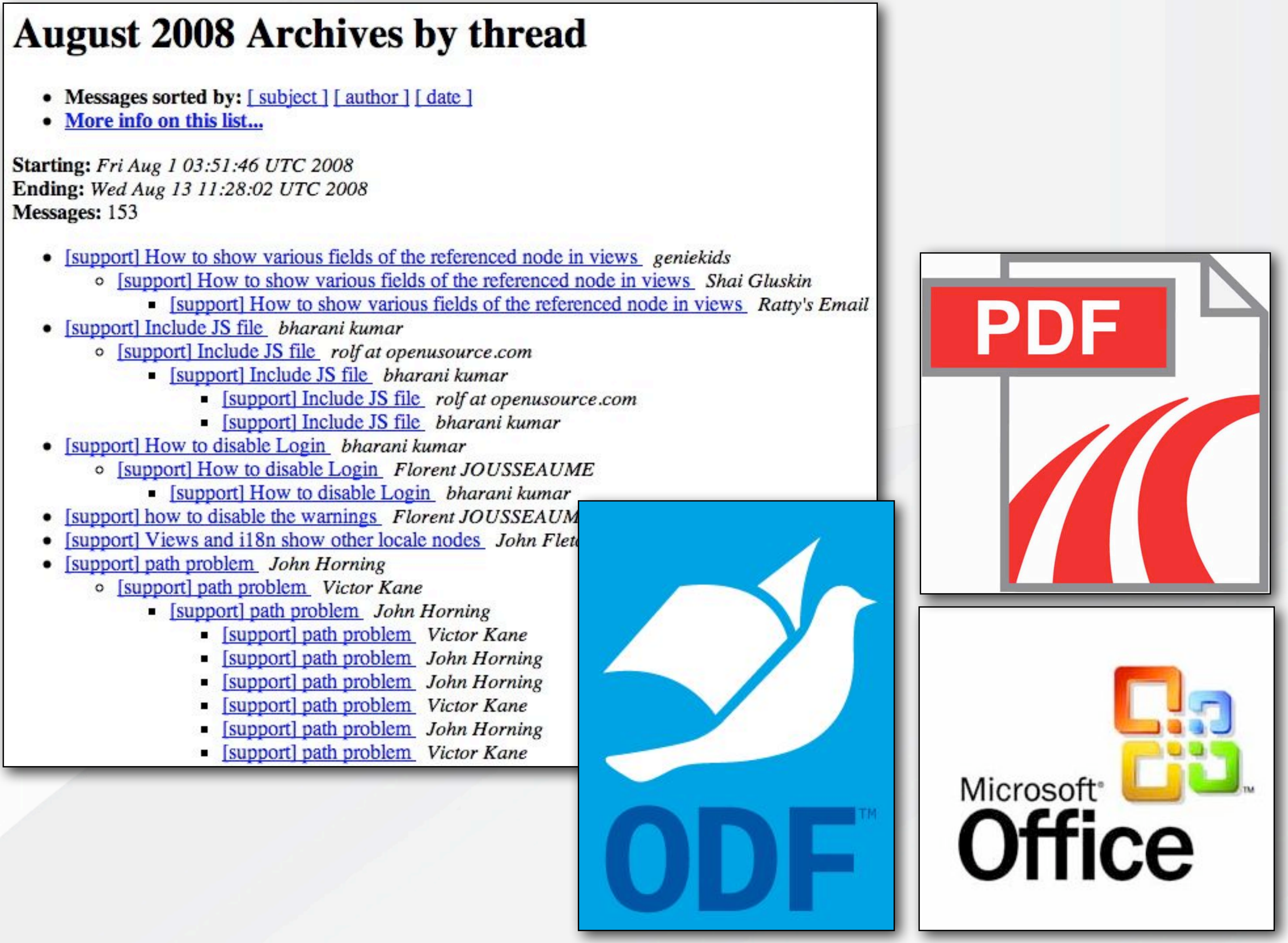


August 2008 Archives by thread

- **Messages sorted by:** [subject] [author] [date]
- **More info on this list...**

Starting: *Fri Aug 1 03:51:46 UTC 2008*
Ending: *Wed Aug 13 11:28:02 UTC 2008*
Messages: 153

- [support] How to show various fields of the referenced node in views *geniekids*
    - [support] How to show various fields of the referenced node in views *Shai Gluskin*
        - [support] How to show various fields of the referenced node in views *Ratty's Email*
- [support] Include JS file *bharani kumar*
    - [support] Include JS file *rolf at openusource.com*
        - [support] Include JS file *bharani kumar*
            - [support] Include JS file *rolf at openusource.com*
            - [support] Include JS file *bharani kumar*
- [support] How to disable Login *bharani kumar*
    - [support] How to disable Login *Florent JOUSSEAUME*
        - [support] How to disable Login *bharani kumar*
- [support] how to disable the warnings *Florent JOUSSEAUME*
- [support] Views and i18n show other locale nodes *John Fletcher*
- [support] path problem *John Horning*
    - [support] path problem *Victor Kane*
        - [support] path problem *John Horning*
            - [support] path problem *Victor Kane*
            - [support] path problem *John Horning*
            - [support] path problem *John Horning*
            - [support] path problem *Victor Kane*
            - [support] path problem *John Horning*
            - [support] path problem *Victor Kane*



August 2008 Archives by thread

- **Messages sorted by:** [subject] [author] [date]
- **More info on this list...**

Starting: *Fri Aug 1 03:51:46 UTC 2008*
Ending: *Wed Aug 13 11:28:02 UTC 2008*
Messages: 153

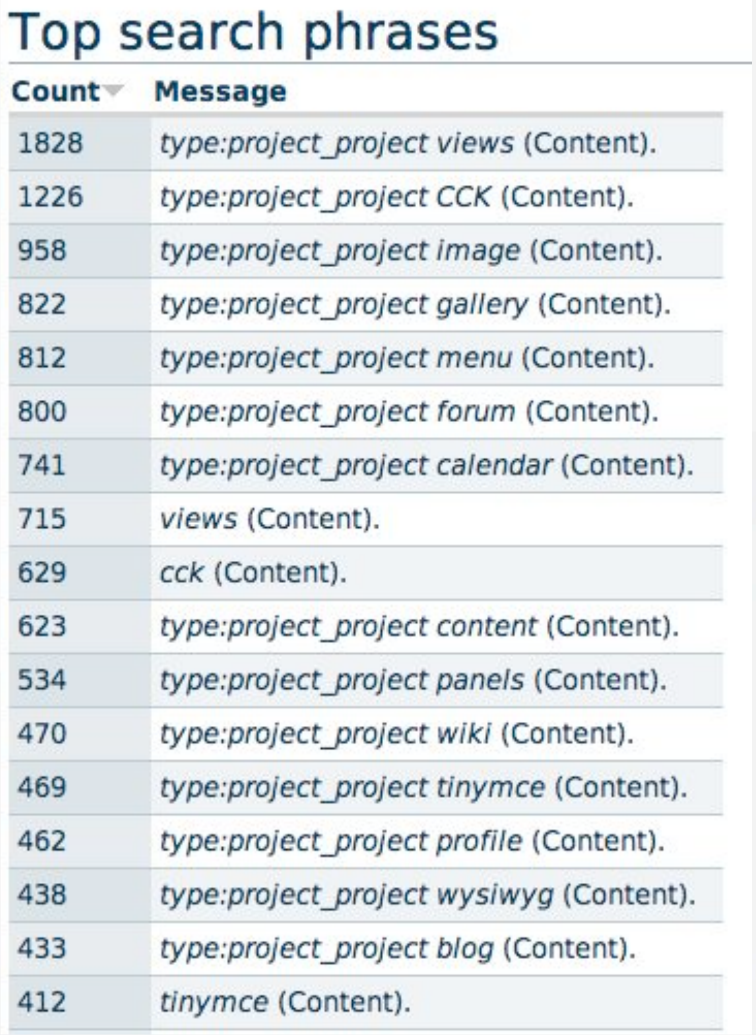
- [support] How to show various fields of the referenced node in views *geniekids*
    - [support] How to show various fields of the referenced node in views *Shai Gluskin*
        - [support] How to show various fields of the referenced node in views *Ratty's Email*
- [support] Include JS file *bharani kumar*
    - [support] Include JS file *rolf at openusource.com*
        - [support] Include JS file *bharani kumar*
            - [support] Include JS file *rolf at openusource.com*
            - [support] Include JS file *bharani kumar*
- [support] How to disable Login *bharani kumar*
    - [support] How to disable Login *Florent JOUSSEAUME*
        - [support] How to disable Login *bharani kumar*
- [support] how to disable the warnings *Florent JOUSSEAUM*
- [support] Views and i18n show other locale nodes *John Flet*
- [support] path problem *John Horning*
    - [support] path problem *Victor Kane*
        - [support] path problem *John Horning*
            - [support] path problem *Victor Kane*
            - [support] path problem *John Horning*
            - [support] path problem *John Horning*
            - [support] path problem *Victor Kane*
            - [support] path problem *John Horning*
            - [support] path problem *Victor Kane*

Search metrics:

Search metrics:

What we have ...

Search metrics:



Top search phrases

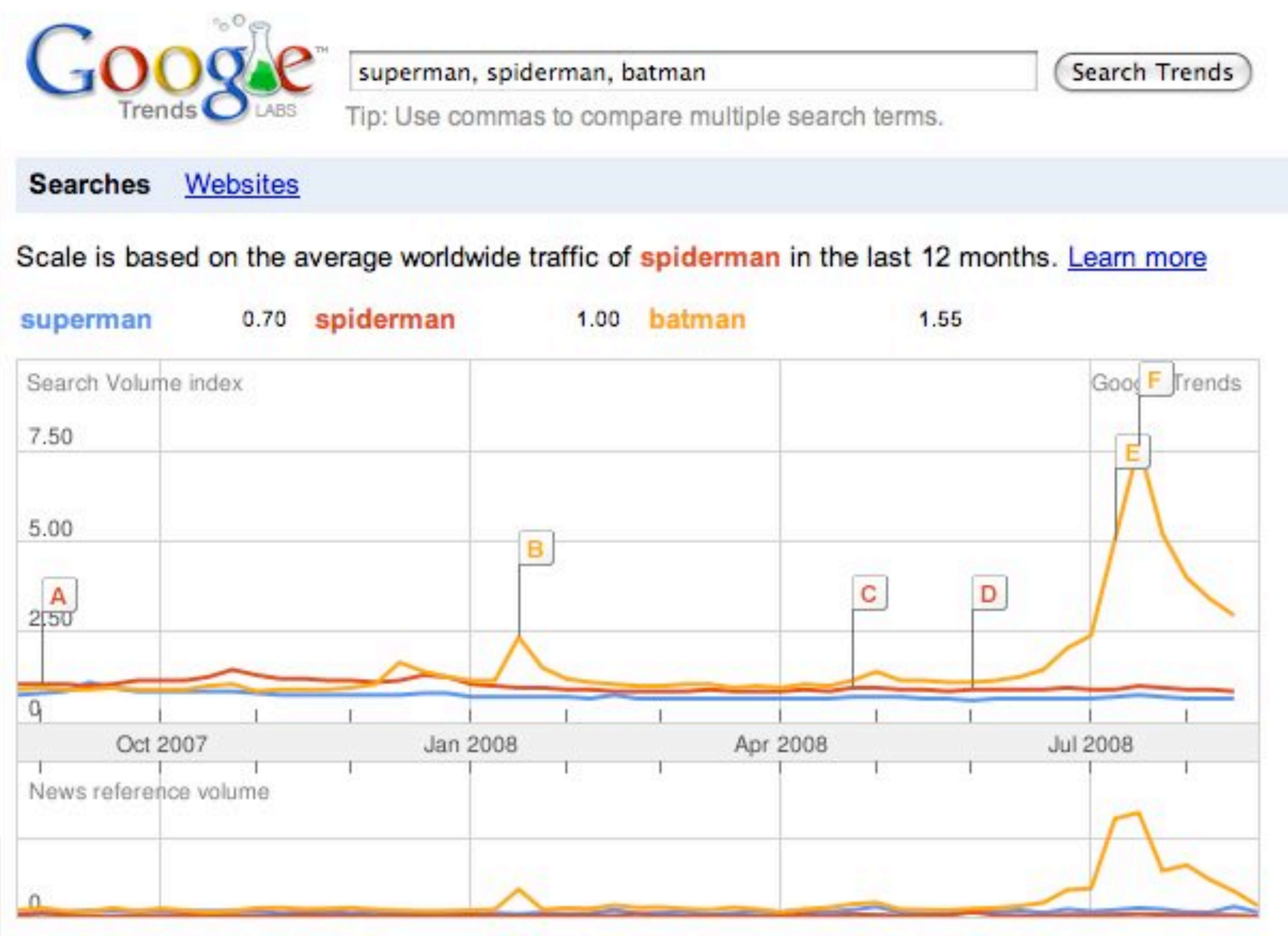
Count	Message
1828	*type:project_project views* (Content).
1226	*type:project_project CCK* (Content).
958	*type:project_project image* (Content).
822	*type:project_project gallery* (Content).
812	*type:project_project menu* (Content).
800	*type:project_project forum* (Content).
741	*type:project_project calendar* (Content).
715	*views* (Content).
629	*cck* (Content).
623	*type:project_project content* (Content).
534	*type:project_project panels* (Content).
470	*type:project_project wiki* (Content).
469	*type:project_project tinymce* (Content).
462	*type:project_project profile* (Content).
438	*type:project_project wysiwyg* (Content).
433	*type:project_project blog* (Content).
412	*tinymce* (Content).

Search metrics:

Search metrics:

What’s possible ...

Search metrics:



superman, spiderman, batman

Tip: Use commas to compare multiple search terms.

Scale is based on the average worldwide traffic of spiderman in the last 12 months. Learn more

superman	0.70	spiderman	1.00	batman	1.55

Some solutions

- ApacheSolr indexes PDF, Word, Text
- Search Files Module
  http://drupal.org/project/search_files
- Swish-E Indexer
  http://drupal.org/project/swish

How to help?

- Testing! There are few unit tests for search.
- See patch on http://drupal.org/node/258998 for the secret to testing the advanced search form.
- Use the SearchBench Module for performance analysis.
- Encourage the RDF and SPARQL modules and similar efforts.

How to get involved?

- Join the search group: http://groups.drupal.org/node/4102
- Review the list of patches: http://groups.drupal.org/node/10569

Some important ones...

- drupal.org/node/286263
  Drupal 6 indexer fails
- drupal.org/node/256792
  Advanced search refactor to open it up
- drupal.org/node/258998
  do_search validation and performance
- drupal.org/node/257033
  Test coverage for search_simplify

Any questions?

robert.douglass@acquia.com