# Blog search using Solr

January 14, 2012

- Overview:
    - Why is it different
    - How does it work

- Audience:
    - Intermediate to Advanced

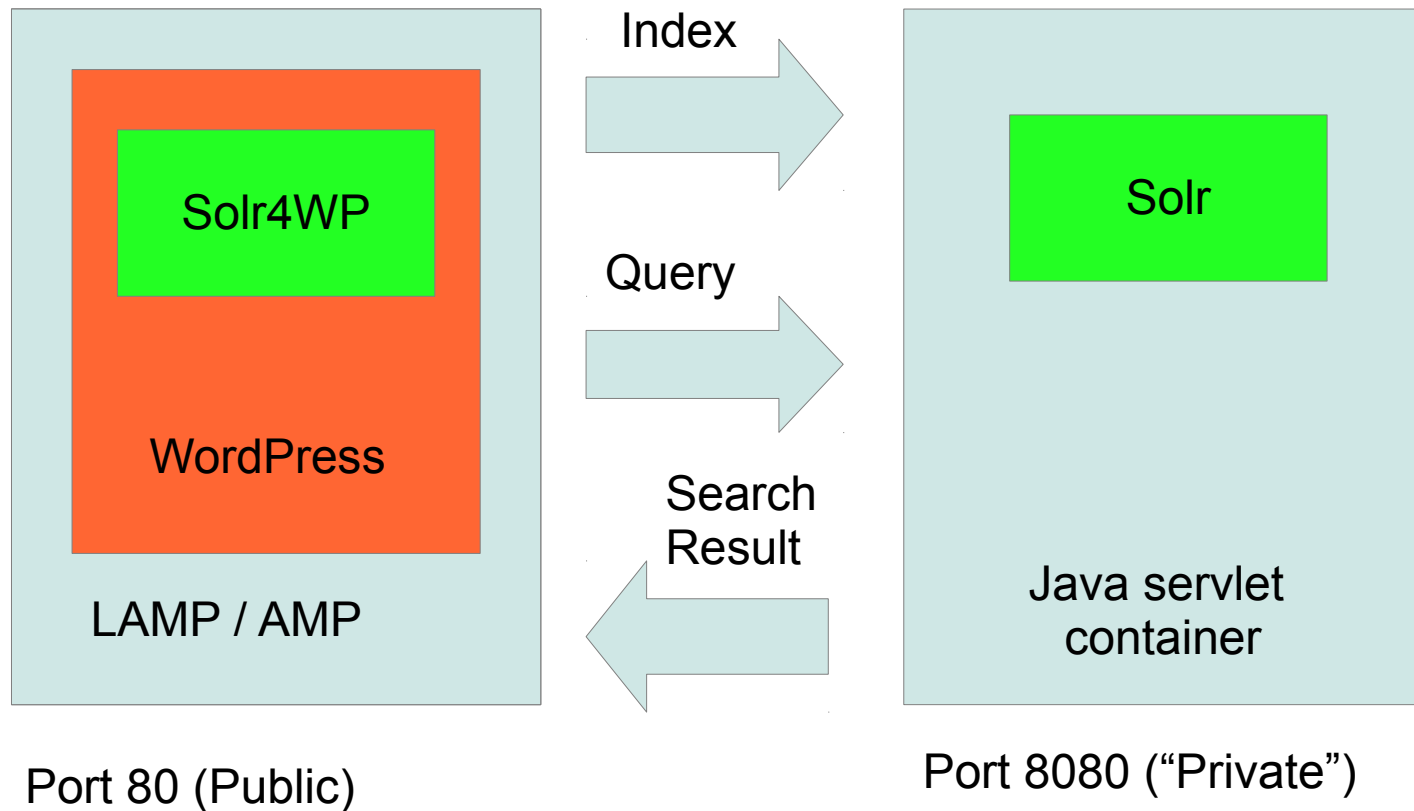- Today's example:
    - Searching WordPress

# Overview

- 1. Installing and configuring Solr:
    - Theory
    - Demonstration

- 2. Indexing a blog (WordPress):
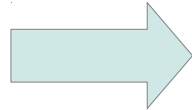    - Theory
    - Demonstration

# Overview continued

- 3. Searching:
  - Theory
  - Demonstration
- Additional topics and ideas
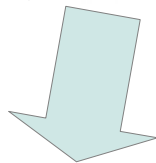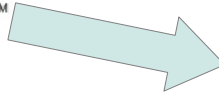- Summary and question time

# Overview – Diagram

# Starts with a story
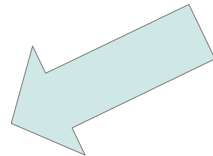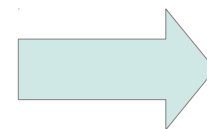
June 2011



Netflix
digg
Zappos
Bibliocommons
/ GVPL
...

# About Solr

**What Is Apache Lucene?**

The Apache Lucene™ project develops open-source search software, including:

- **Apache Lucene Core**™ (formerly named Lucene Java), our flagship sub-project, provides a Java-based indexing and search implementation, as well as spellchecking, hit highlighting and advanced analysis/tokenization capabilities.
- **Apache Solr**™ is our high performance enterprise search server, with XML/HTTP and JSON/Python/Ruby APIs, hit highlighting, faceted search, caching, replication, distributed search, database integration, web admin and search interfaces.
- Apache PyLucene™ is a Python port of the the Lucene Core project

http://lucene.apache.org/

- Lucene

- Apache license

- Fully featured

  - http://lucene.apache.org/java/docs/features.html

- Java

# 1. Installing and configuring

- Basic example today is running on this laptop – not a production environment

- Solr: Java, servlet container

- WordPress: PHP, MySQL, web server, plug-in

- Open source compatible licenses

# 1. Before

WordPress

LAMP / AMP

Port 80 (Public)

# 1. After

WordPress

LAMP / AMP

Port 80 (Public)

Solr

Java servlet
container

Port 8080 ("Private")

# 1. Demonstration

- Tomcat service, home page

- Web Application Manager

- Solr example servlet

- Security tip – no Solr security implemented in this demonstration
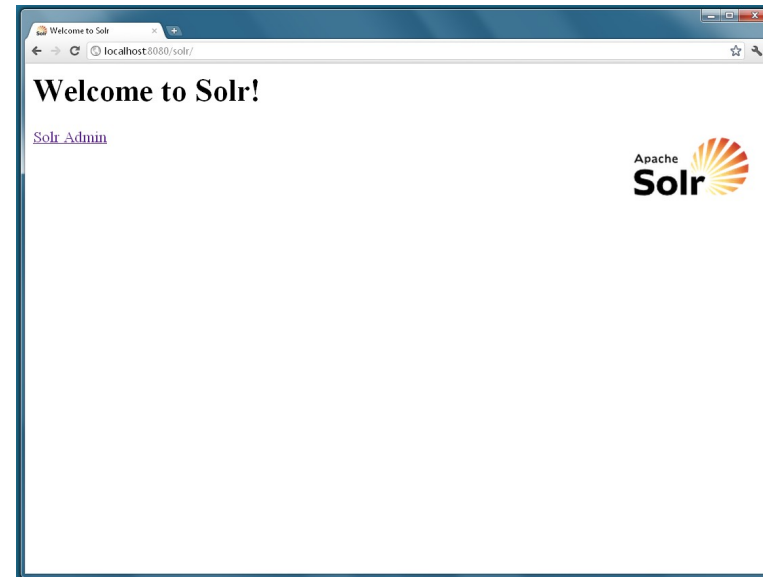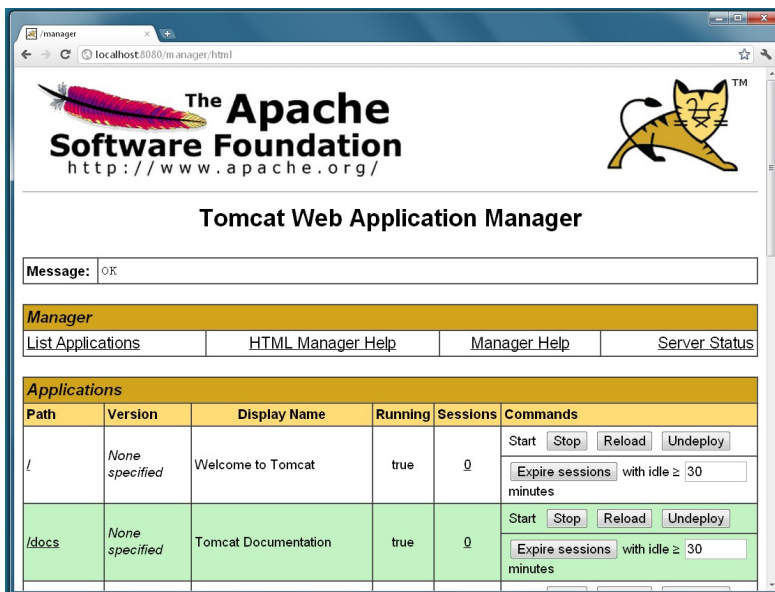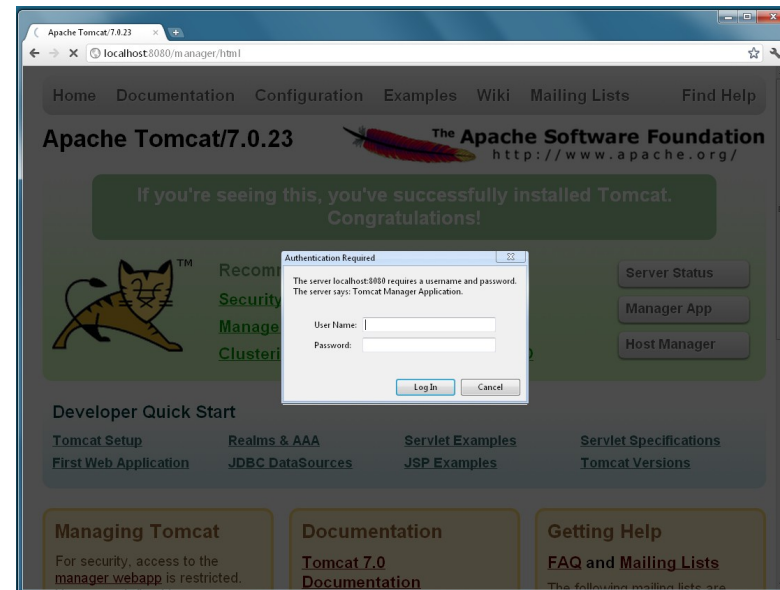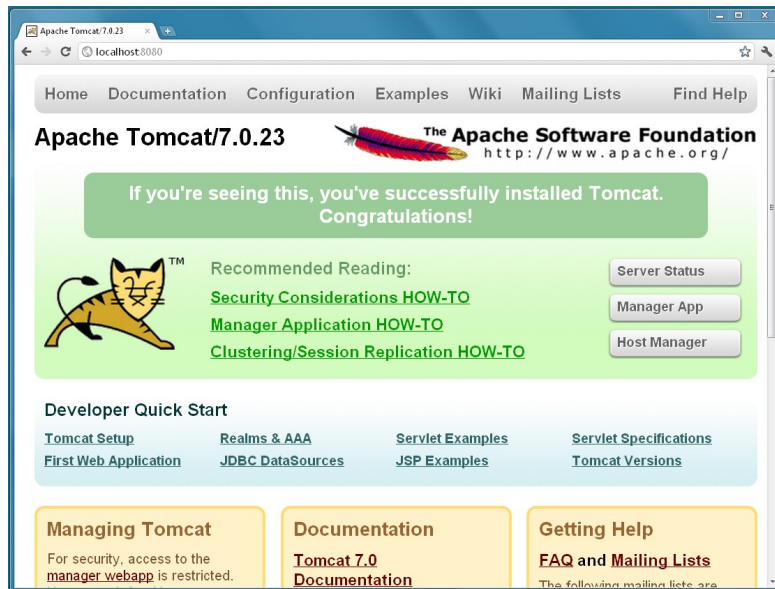
# Solr Admin (example)

Sam:8080
cwd=C:\Program Files\Apache Software Foundation\Tomcat 7.0 SolrHome=d:\solr\solr\.\
HTTP caching is OFF

Apache **Solr**

| **Solr** | [SCHEMA] [CONFIG] [ANALYSIS] [SCHEMA BROWSER] |
| | [STATISTICS] [INFO] [DISTRIBUTION] [PING] [LOGGING] |
| **App server:** | [JAVA PROPERTIES] [THREAD DUMP] |

## Make a Query [FULL INTERFACE]

Query String:

```
*:*
```

Search

## Assistance

[DOCUMENTATION] [ISSUE TRACKER] [SEND EMAIL]
[SOLR QUERY SYNTAX]

Current Time: Mon Jan 16 10:12:27 PST 2012

Server Start At: Sat Jan 14 09:57:34 PST 2012

# 2. Indexing WordPress

- ## Solr for WordPress plugin

  - http://wordpress.org/extend/plugins/solr-for-wordpress/

  - Version 0.5.1

  - Look at features

A WordPress plugin that replaces the default WordPress search with Solr.

Download Version 0.5.1

Description | Installation | FAQ | Screenshots | Other Notes | Stats

A WordPress plugin that replaces the default WordPress search with Solr. Features include:

- Index pages and posts
- Enable faceting on fields such as tags, categories, author, and page type.
- Indexing and faceting on custom fields
- Multisite support
- Treat the category facet as a taxonomy
- Add special template tags so you can create your own custom result pages to match your theme.
- Completely replaces default WordPress search, just install and configure.
- Completely integrated into default WordPress theme and search widget.
- Configuration options allow you to select pages to ignore, features to enable/disable, and what type of result information you want output.
- i18n Support
- Multi server/core support

Note that this plugin requires you to have an instance of Solr using a schema with the following fields: id, permalink, title, content, numcomments, categories, categoriessrch, tags, tagssrch, author, type, and text. The facet fields (categories, tags, author, and type) should be string fields. You can make tagssrch and categoriessrch of any type you want as they are used for general searching. The plugin is distributed with a Solr schema you can use at `solr-for-wordpress/schema.xml`.

**Author:** mattweber

**Tags:** custom search, search, search integration, search results, solr

**Requires:** 3.0 or higher
**Compatible up to:** 3.2.1
**Last Updated:** 2011-8-28
**Downloads:** 1,640

**Average Rating**
★★★★☆
(8 ratings)

**FYI**
Other Versions »
Subversion Repository »
Donate to this plugin »

Forums Posts
Development Log

**Compatibility**

| | | |
|---|---|---|
| WordPress | 3.3.1 | ▼ |
| + Plugin | 0.5.1 | ▼ |
| = | Not enough data | |

0 people say it works.
0 people say it's broken.

Log in to vote.

# 2. Fields and connector

- Schema overview

- SolrPHPClient

```xml
<fields>
    <field name="id" type="string" indexed="true" stored="true" required="true" />

        <field name="blogid" type="string" indexed="true" stored="true"/>
        <field name="blogdomain" type="string" indexed="true" stored="true"/>
        <field name="blogpath" type="string" indexed="true" stored="true"/>
        <field name="wp" type="string" indexed="true" stored="true"/>

        <field name="permalink" type="string" indexed="true" stored="true"/>
    <field name="title" type="text_lws" indexed="true" stored="true"/>
        <field name="content" type="text" indexed="true" stored="true"/>
        <field name="numcomments" type="integer" indexed="true" stored="true"/>
        <field name="comments" type="text" indexed="true" stored="true" multiValued="true"/>

    <field name="categories" type="string" indexed="true" stored="true" multiValued="true"/>
        <field name="categoriessrch" type="text_lws" indexed="true" stored="false" multiValued="true"/>

    <field name="tags" type="string" indexed="true" stored="true" multiValued="true"/>
        <field name="tagssrch" type="text_lws" indexed="true" stored="false" multiValued="true"/>

    <field name="author" type="string" indexed="true" stored="true"/>
        <field name="type" type="string" indexed="true" stored="true"/>

        <!-- Date fields used for sorting and searching -->
        <field name="date" type="date" indexed="true" stored="false"/>
        <field name="modified" type="date" indexed="true" stored="false"/>

        <!-- Fields used to display the date -->
        <field name="displaydate" type="string" indexed="false" stored="true"/>
        <field name="displaymodified" type="string" indexed="false" stored="true"/>

        <!-- For spell checking, did you mean? -->
        <field name="spell" type="textSpell" indexed="true" stored="true" multiValued="true"/>

    <!-- composite field -->
    <field name="text" type="text" indexed="true" stored="false" multiValued="true"/>

    <!-- dynamic fields -->
    <dynamicField name="*_str" type="string" indexed="true" stored="true" multiValued="true"/>
```

# 2. Before



WordPress

LAMP / AMP

Port 80 ("Public")

Solr

Java servlet
container

Port 8080 ("Private")

# 2. After

Index →

**Solr4WP**

**WordPress**

LAMP / AMP

Port 80 (Public)

**Solr**

Java servlet container

Port 8080 ("Private")

# 2. Demonstration

- Plugin install

- Plugin activate

- Solr configuration

- Ping – running or not?

- Index data (note about Firefox)

- Plugin options

- View in Solr admin

**Window 1 — Plugins**

WhiskyIntelligence.com | + New | Howdy,

Appearance
Plugins
Installed Plugins
Add New
Editor
WP-SpamFree
Users
Tools
Settings
Contact
Collapse menu

Akismet — Activate | Edit | Delete
Used by millions, Akismet is quite possibly the best way in the world to **protect your blog from comment and trackback spam**. It keeps your site protected from spam even while you sleep. To get started: 1) Click the "Activate" link to the left of this description, 2) Sign up for an Akismet API key, and 3) Go to your Akismet configuration page, and save your API key.
Version 2.5.5 | By Automattic | Visit plugin site

Contact Form 7 — Settings | Deactivate | Edit
Just another contact form plugin. Simple but flexible.
Version 3.0.2.1 | By Takayuki Miyoshi | Visit plugin site

Hello Dolly — Activate | Edit | Delete
This is not just a plugin, it symbolizes the hope and enthusiasm of an entire generation summed up in two words sung most famously by Louis Armstrong: Hello, Dolly. When activated you will randomly see a lyric from Hello, Dolly in the upper right of your admin screen on every page.
Version 1.5 | By Matt Mullenweg | Visit plugin site

Sociable — Deactivate | Edit
Automatically add links on your posts, pages and RSS feed to your favorite social bookmarking sites.
Version 4.1.5 | By Blogplay | Visit plugin site

Solr for WordPress — Activate | Edit | Delete
Indexes, removes, and updates documents in the Solr search engine.
Version 0.5.1 | By Matt Weber | Visit plugin site

WP-SpamFree — Settings | Deactivate | Edit
An extremely powerful anti-spam plugin that virtually eliminates comment spam. Finally, you can enjoy a spam-free WordPress blog! Includes spam-free contact form feature as well.
Version 2.0.0.3 | By Scott Allen, aka WebGeek | Visit plugin site

Plugin | Description

Bulk Actions | Apply          6 items

Thank you for creating with WordPress.          Version 3.3.1
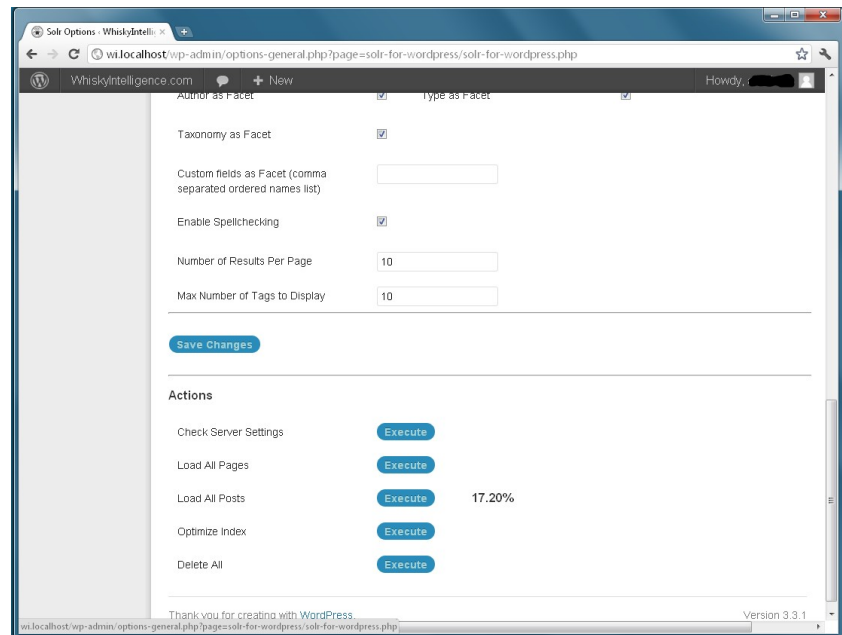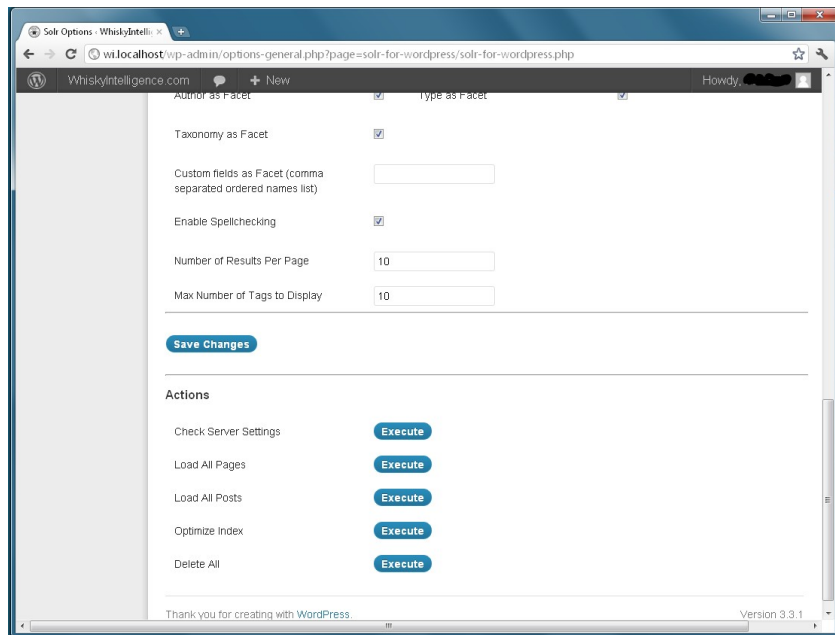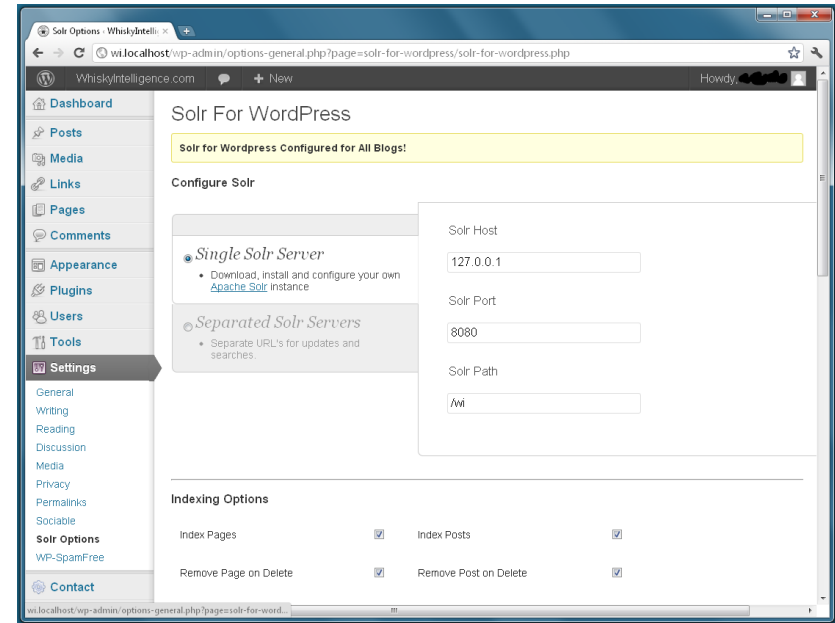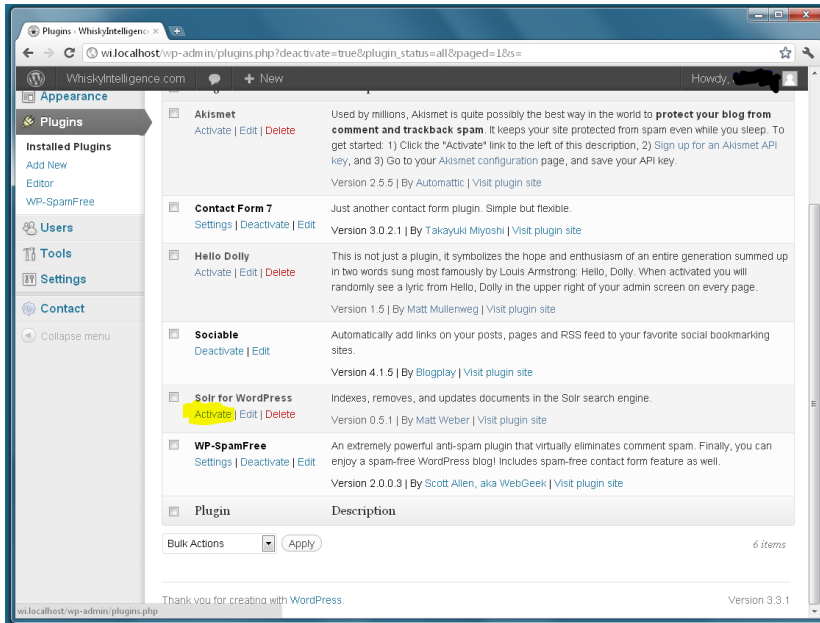
**Window 2 — Solr Options**

Dashboard
Posts
Media
Links
Pages
Comments
Appearance
Plugins
Users
Tools
Settings
General
Writing
Reading
Discussion
Media
Privacy
Permalinks
Sociable
Solr Options
WP-SpamFree
Contact

Solr For WordPress

Solr for Wordpress Configured for All Blogs!

Configure Solr

• Single Solr Server
  • Download, install and configure your own Apache Solr instance

○ Separated Solr Servers
  • Separate URL's for updates and searches.

Solr Host
127.0.0.1

Solr Port
8080

Solr Path
/wi

Indexing Options

Index Pages ☑        Index Posts ☑
Remove Page on Delete ☑   Remove Post on Delete ☑

**Window 3 — Solr Options**

Author as Facet ☑        Type as Facet ☑
Taxonomy as Facet ☑
Custom fields as Facet (comma separated ordered names list) [ ]
Enable Spellchecking ☑
Number of Results Per Page [10]
Max Number of Tags to Display [10]

Save Changes

Actions
Check Server Settings — Execute
Load All Pages — Execute
Load All Posts — Execute
Optimize Index — Execute
Delete All — Execute

Thank you for creating with WordPress.          Version 3.3.1

**Window 4 — Solr Options**

Author as Facet ☑        Type as Facet ☑
Taxonomy as Facet ☑
Custom fields as Facet (comma separated ordered names list) [ ]
Enable Spellchecking ☑
Number of Results Per Page [10]
Max Number of Tags to Display [10]

Save Changes

Actions
Check Server Settings — Execute
Load All Pages — Execute
Load All Posts — Execute   17.20%
Optimize Index — Execute
Delete All — Execute

Thank you for creating with WordPress.          Version 3.3.1

# Solr Statistics: (wpdata)

Sam

Apache **Solr**

| Category | [CORE] [CACHE] [QUERY] [UPDATE] [HIGHLIGHTING] [OTHER] |
|---|---|
| | Current Time: Mon Jan 16 10:28:38 PST 2012 |
| | Server Start Time: Sat Jan 14 09:57:35 PST 2012 |

## CORE

| name: | core |
|---|---|
| class: | |
| version: | 1.0 |
| description: | SolrCore |
| stats: | coreName :<br>startTime : Sat Jan 14 09:57:35 PST 2012<br>refCount : 2<br>aliases : [] |
| name: | searcher |
| class: | org.apache.solr.search.SolrIndexSearcher |
| version: | 1.0 |
| description: | index searcher |
| stats: | searcherName : Searcher@6ff3d99c main<br>caching : true<br>numDocs : 4356<br>maxDoc : 4356<br>reader :<br>SolrIndexReader{this=7db1592b,r=ReadOnlyDirectoryReader@7db1592b,refCnt=1,segments=3}<br>readerDir : org.apache.lucene.store.MMapDirectory@D:\solr\wi\data\index |

# 3. Searching

- Solr queries

- Facets to filter results

- Sorting

- Theme search template (optional customization)
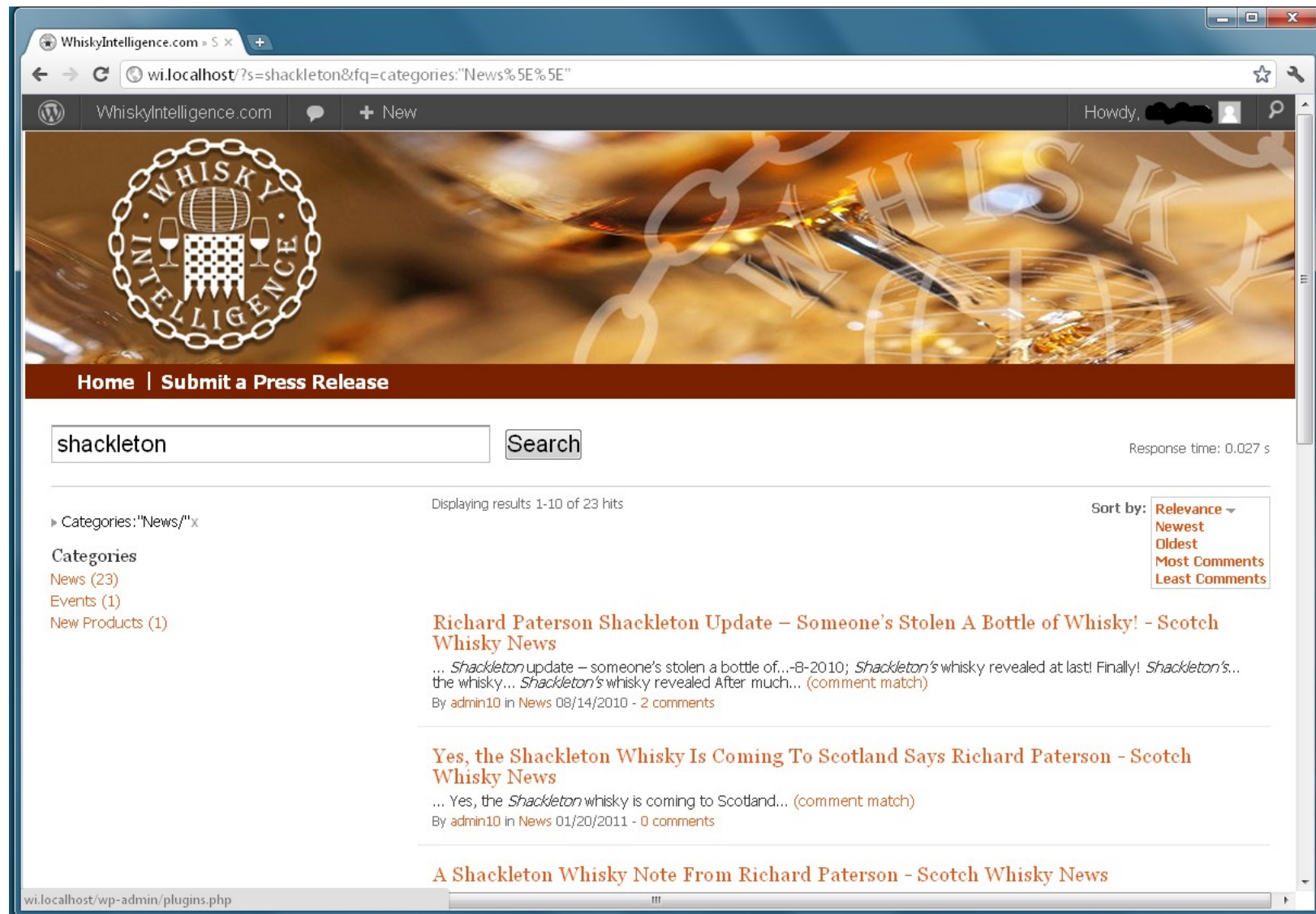
# 3. Demonstration

- Search comparison
  - Plugin inactive
  - Plugin active
- Facet / filter
  - View URL
- Sorting
- Pagination
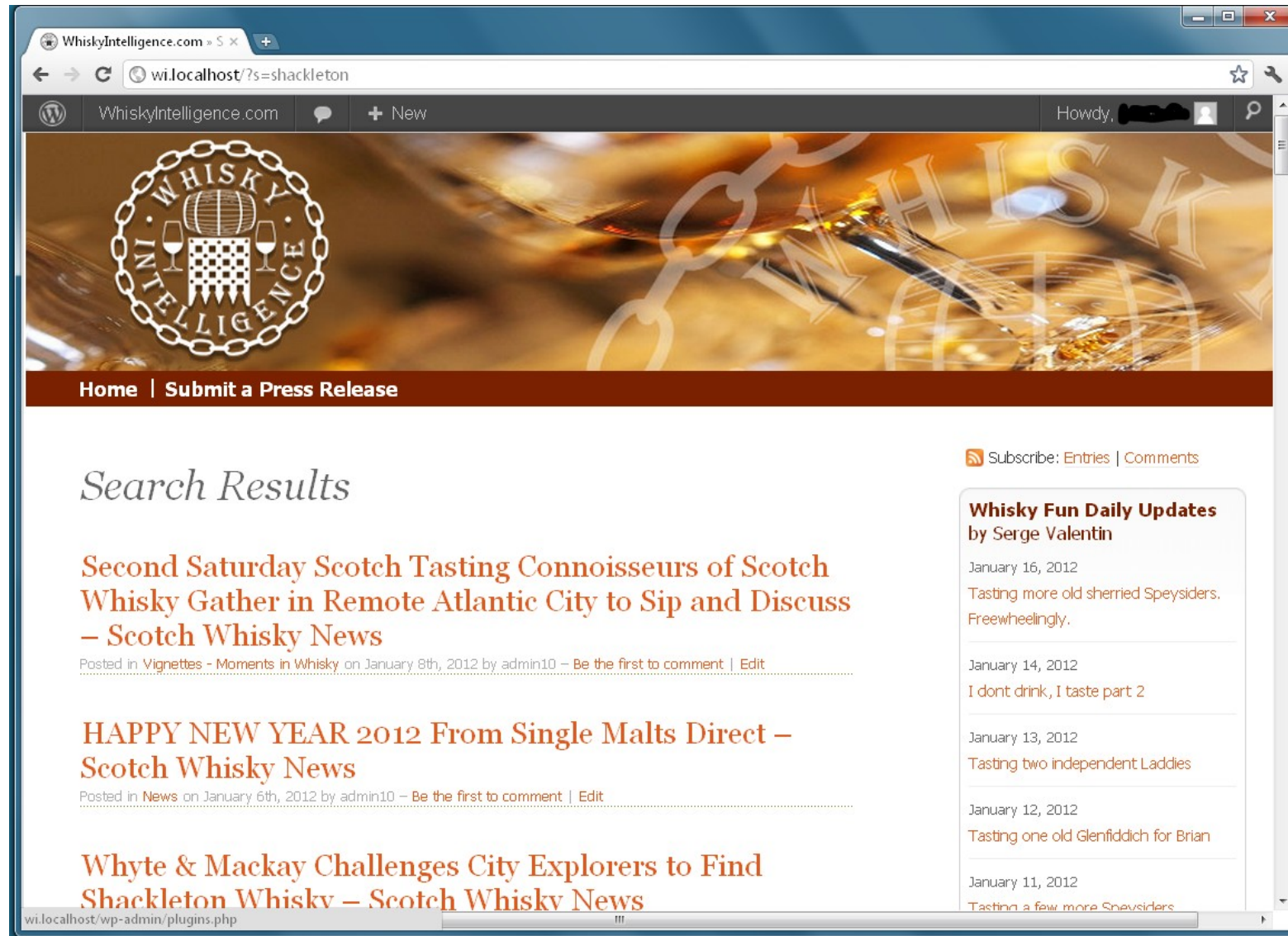
Before
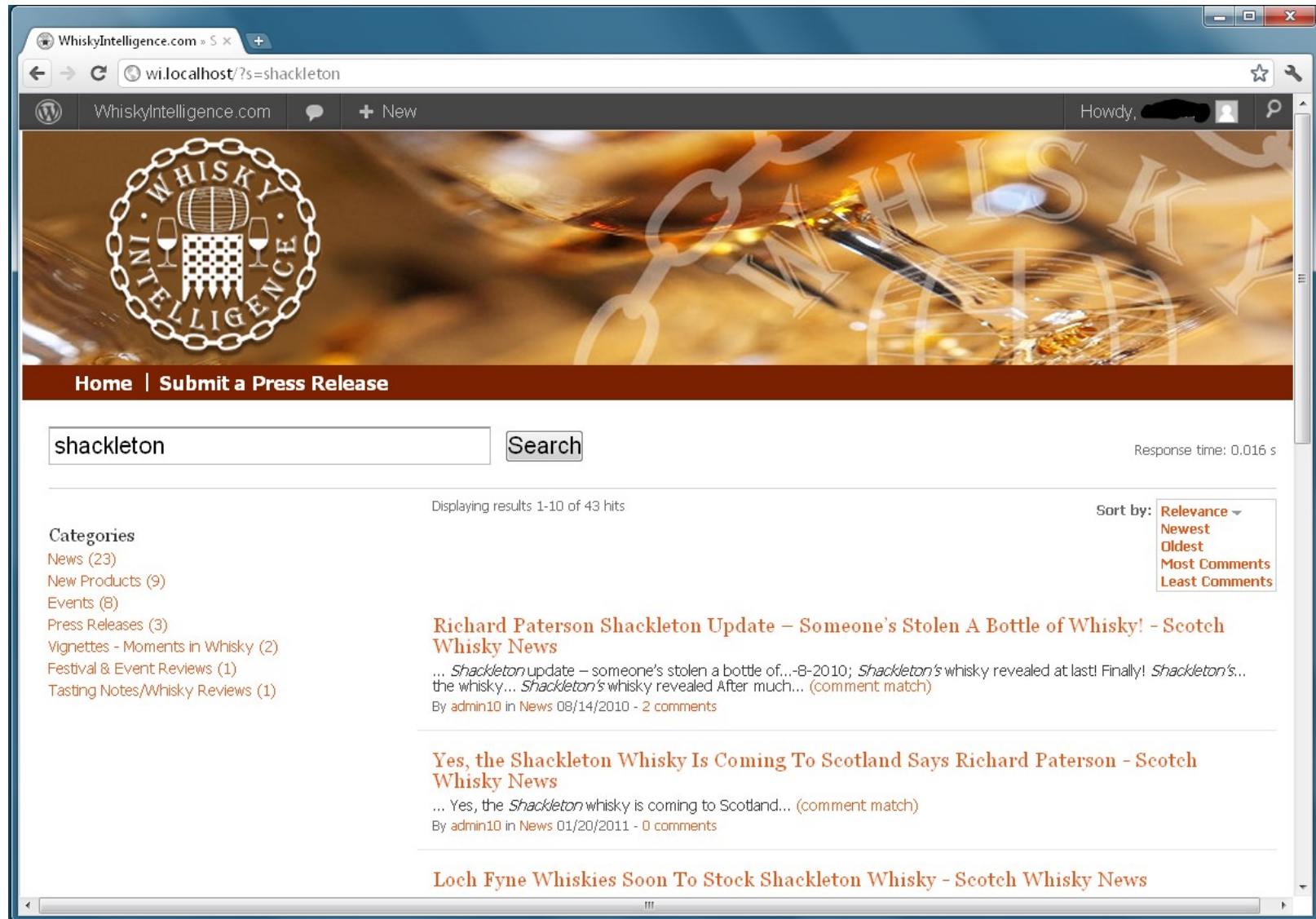
After

News category facet search

# Additional topics and ideas

- Tags – example blog did not use them

- Dynamic fields / custom post types

- Customized theme

- Aggregate multiple blogs (*access privileges to consider)

- Multiple Solr servers

- Hosted Solr solutions

# Summary and question time

- Questions

- Review slides

- More demonstrations

    - Create a new post -> indexing -> search

    - How to create a new Solr index

    - ...

# Thanks!

- Enjoy WordCamp Victoria 2012

- Thanks to the organizers

- Available later for questions

- Graham Nott
- Star Global
- http://www.starglobal.ca
- gnott@starglobal.ca