visionCMS™ Vision Search User Guide

Yuliang Ding

Vision Technology Solutions, LLC.

Oct 5, 2015 – 2015.3.928.0
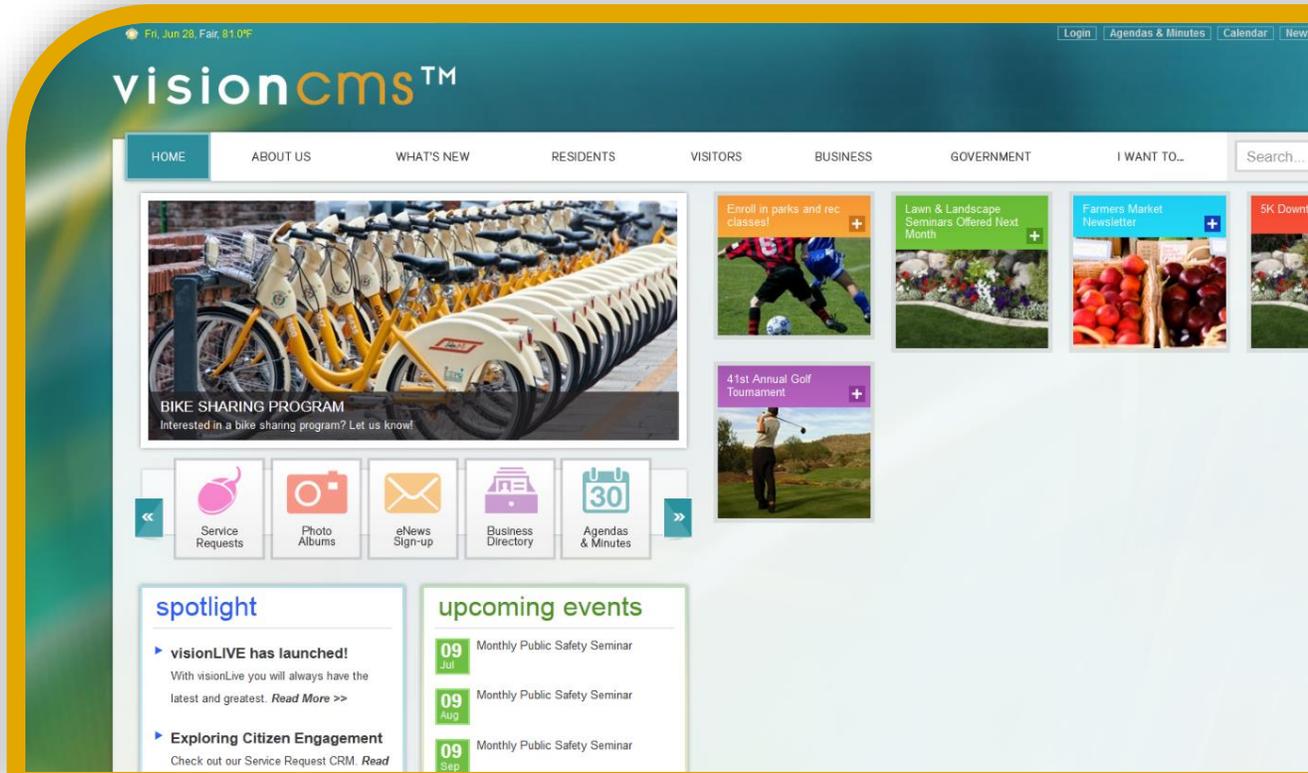
Table of Contents

Introduction

Vision Search is built in-house on top of the popular open source search library Solr which is in turn built on Apache Lucene™. Solr is a very popular search platform used on such websites as eHarmony, Sears, Zappos, AT&T, Ticketmaster, eBay, Instagram, Netflix, Disney, Adobe and StubHub. Building our own search engine on top of Solr makes sense because we can rely on the powerful search technology built into Solr to ensure fast and accurate search results as well as intelligent indexing, while also being able to build and customize and continue to expand the functionality to work specifically with CMS6.

Vision Search includes two parts:

1. Search Engine Management Site
2. Integration with CMS6

Vision Search Management Site Guide

You can access Vision Search live site by: http://search.visioninternet.com/. Currently, only the IT team has the permission to access this site.

User & Security Role
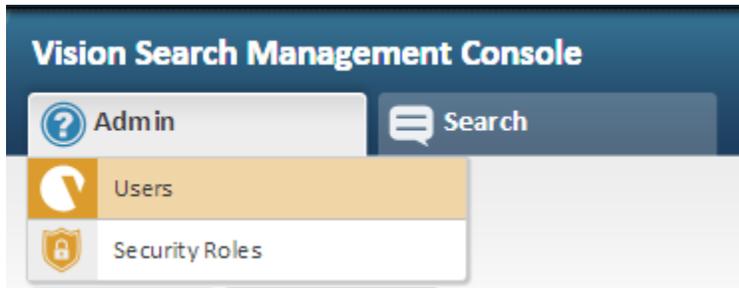
The user and security role logic is the same as CMS6, if you want to add more users into Vision Search Management Site, please add related security roles and assign to the users.

Search Engine
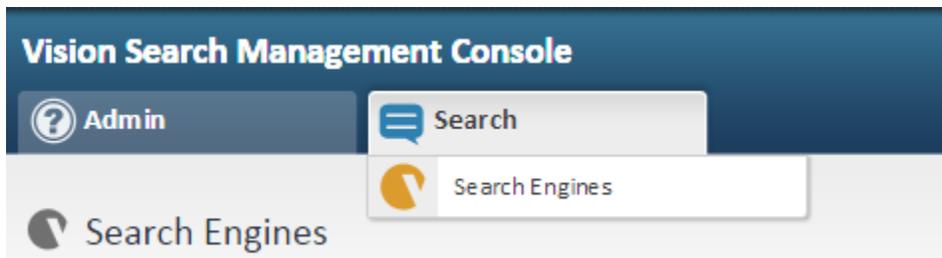
We can create a search engine for each site, or for each different purpose of each site, e.g. different departments in one city may need different search engines. These search engines' settings won't be affected by each other.

How to setup a search engine

Collect Requirements

Get the requirements from client:

1. What site URL does the client want to create search engine for?
2. If the client wants to only index specific pages, instead of whole website, please get the requirement that what kinds of pages on the site does the client want to index? E.g. all subpages of one page …
3. About the indexing frequency, we will set it weekly by default. I don't suggest to ask client weekly or daily, as daily will use more CPU.

Based on above requirements, we can know:

1. Root URLs
    Which URLs we should start to crawl the site.
2. Allow Paths
    Which kinds of URL paths we can allow to crawl.
3. Indexing Allow Paths
    Which kinds of URL paths we can allow to index.

Create a search engine

1. Go to Vision Search Management Site, Search -> Search Engines, Click "ADD SEARCH ENGINES" button:

2. Put the basic information in the form:
    a. ***Name***: is required, and should be unique.

b. **_Description_**: put any information here which is helpful to understand the purpose of the search engine.
c. **_Root URLs_**: The URLs we should start to crawl the site, which is multiple lines; we can add multiple URLs, even they are in different sites.
d. **_Allow Paths_**: The URLs paths or regular expressions we can allow to crawl, which is multiple lines; we can add URL path or regular expression for each line.
e. **_Disallow Paths_**: The URLs paths or regular expressions we don't allow to crawl, which is multiple lines; we can add URL path or regular expression for each line.
Usually, for CMS6 website, we don't need change this setting, just use the default one.
f. **_Indexing Allow Paths_**: The URLs paths or regular expressions we can allow to index, which is multiple lines; we can add URL path or regular expression for each line.
This setting is different from "Allow Paths". Only when the indexed pages are different from the crawled pages, we have to setup this setting, otherwise, leave it as EMPTY.
g. **_Allow Formats_**: The file formats we can allow to crawl and index.
For CMS6 website, usually, we don't need to change this setting, just use the default one.

After you have setup all the settings based on our requirements, click "SAVE" button. After you have saved, it will redirect back to the search engine list page.

Some settings will be generated by default, e.g. Search Engine Key, Search Engine Secret, Content Type Rule, Scanner Settings, Proxy Settings. Usually, we don't need to change these settings for our CMS6 websites, these settings have been optimized for our CMS6 websites.

Search Engine Indexing Activities

After we have created a search engine, we have to setup the scheduled search engine indexing activities.

1. Click "Edit", choose "Indexing" tab to setup the scheduled indexing activities.

Here, we have two options to setup the scheduled indexing activities.

a. Setup recurring/one time scheduled indexing activities.

This is for future scheduled indexing activities, we can setup a one-time schedule, or a recurring schedule. You have to click the "SAVE" button after you have chosen the scheduled options.

b. Setup "schedule now" indexing activities.

After we have setup "schedule now" indexing activities, the search engine will run this activity ASAP.
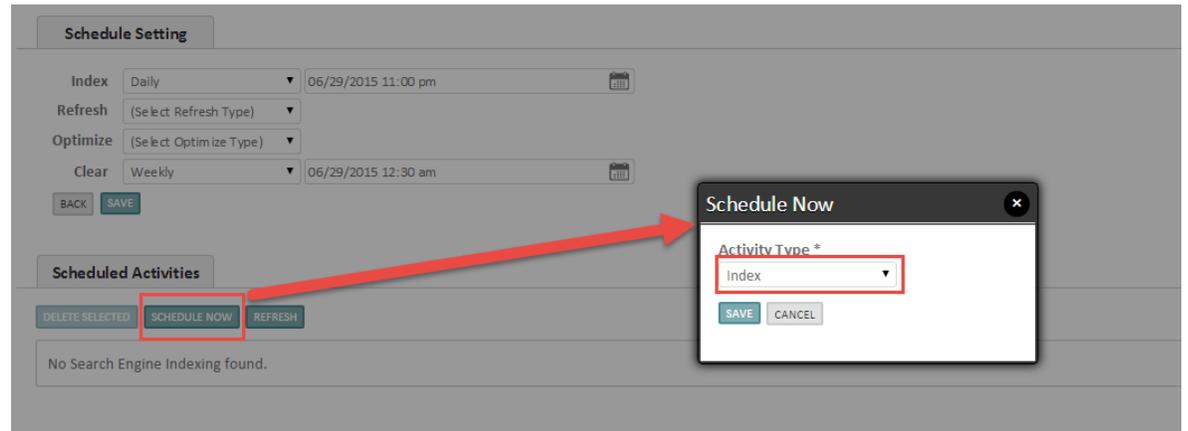
CMS6 Best Practice:
- a. Once we have created the search engine, please "schedule now" at least one time.
- b. For the Scheduled Indexing activities, please :
  - ➢ Index -> Daily -> the date with time after 11:00 pm.
  - ➢ Clear -> Weekly -> the date with time after 12:00 pm.
2. Now, we have finished the setup of a search engine. But if the site is big, which means there are lots of pages in that site, I would suggest to make another change to the search engine: "Spider Max. Depth": by default, it's 2. If the site is big, we can set it as 3 or 4.