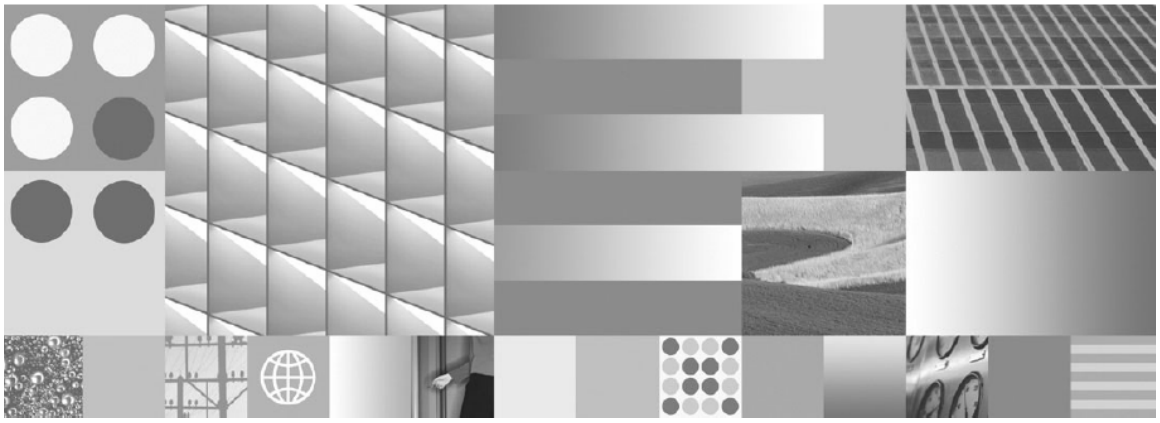
IBM OmniFind Yahoo! Edition

Version 8.4



Adding IBM’s OmniFind Yahoo! Edition to Your Web Site

By: Todd Leyba (Software Architect)

December 13th, 2006

Table of Contents

Introduction

IBM® and Yahoo! have partnered together to offer a free downloadable search engine that is easy to set up and use. IBM's OmniFind™ Yahoo! Edition (referred to as just OmniFind in this article) can crawl and index up to half a million Web pages and/or file system documents and make them available for search through a simple to use Web interface. You may have already downloaded OmniFind and discovered how quick and easy it is to set up an index and start searching.

But you are beyond that point. You are now investigating how to integrate OmniFind's search functionality into your Web site. In this article I will explore four ways to accomplish this. Namely:

1. Linking directly to the OmniFind search page
2. Using your own search box and button
3. Presenting search results as HTML snippets
4. Using XSLT to transform OmniFind XML results into HTML

The methods presented here build upon each other, adding increased flexibility at the expense of increased level of effort.

Integration Scenario

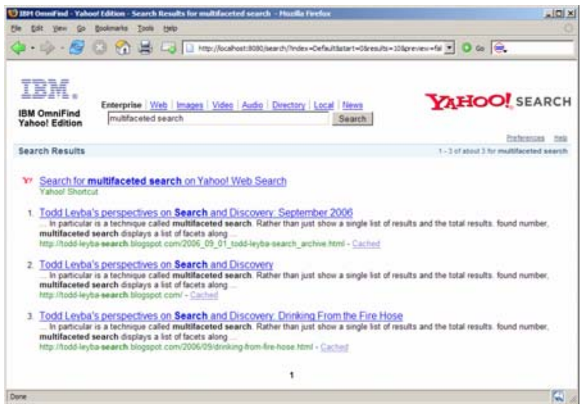
The scenario I will follow is to add OmniFind search functionality to a blog site. If you have ever read a blog you will notice that not all of the blogger's postings are presented on the first page. Most blog hosting facilities list the most recent postings and provide an archive link to the older postings which are typically organized by month. So if you want to view a posting that is not listed on the first page, you would need to click through the previous months to find the article. In this integration scenario, I use OmniFind to crawl and index my personal blog site and then allow users to search my OmniFind blog index to find previous postings. I use Google's Blogger to host my real blog (Todd Leyba's Perspectives on Search and Discovery) and Blogger already provides such a search capability. But Blogger's search feature is immutable. As a result, I will demonstrate in this article how to replace Blogger's search facility with OmniFind and then show how to customize the overall search experience.

Linking to OmniFind Directly

The easiest integration option is to provide a link to the OmniFind search page in a conspicuous place somewhere in the blog. All changes to my blog site are made through a Blogger template. The template defines the overall structure of my blog and is written in standard HTML. As such it is a simple matter for me to insert a link to the OmniFind search page – in this case right above the archive section. The link to use would be similar to the following but with the host and port changed to the OmniFind installation site.

`http://omniFindhost:8080/search/`

The overall look and feel of the OmniFind search interface (shown to the right) can then be customized by selecting from several layout options that are offered in the OmniFind administrator's console. With no programming involved you can change the banners and images to those of your company, change the text of various labels and buttons, and even choose which features are to appear or not (such as summaries, footers, etc.)


IBM OmniFind - Yahoo! Edition - Search Results for multifaceted search - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost:8080/search/?index=Default&start=0&results=10&preview=fal
IBM OmniFind Yahoo! Edition
Enterprise | Web | Images | Video | Audio | Directory | Local | News
multifaceted search
Search
YAHOO! SEARCH
Preferences Help
Search Results
1 - 3 of about 3 for multifaceted search

Search for **multifaceted search** on Yahoo! Web Search
Yahoo! Shortcut

1. Todd Leyba's perspectives on **Search** and Discovery: September 2006
... In particular is a technique called **multifaceted search**. Rather than just show a single list of results and the total results. found number, **multifaceted search** displays a list of facets along ...
http://todd-leyba-**search**.blogspot.com/2006_09_01_todd-leyba-search_archive.html - Cached

2. Todd Leyba's perspectives on **Search** and Discovery
... In particular is a technique called **multifaceted search**. Rather than just show a single list of results and the total results. found number, **multifaceted search** displays a list of facets along ...
http://todd-leyba-**search**.blogspot.com/ - Cached

3. Todd Leyba's perspectives on **Search** and Discovery: Drinking From the Fire Hose
... In particular is a technique called **multifaceted search**. Rather than just show a single list of results and the total results. found number, **multifaceted search** displays a list of facets along ...
http://todd-leyba-**search**.blogspot.com/2006/09/drinking-from-fire-hose.html - Cached

1

Done

But the direct link approach described above is cumbersome to use. It forces the user to click twice to issue a search – once to get to the OmniFind search page and again to issue the search. Ideally, you would like to have the search box always present on your site so that users can type their query when needed and then click once to see the results.

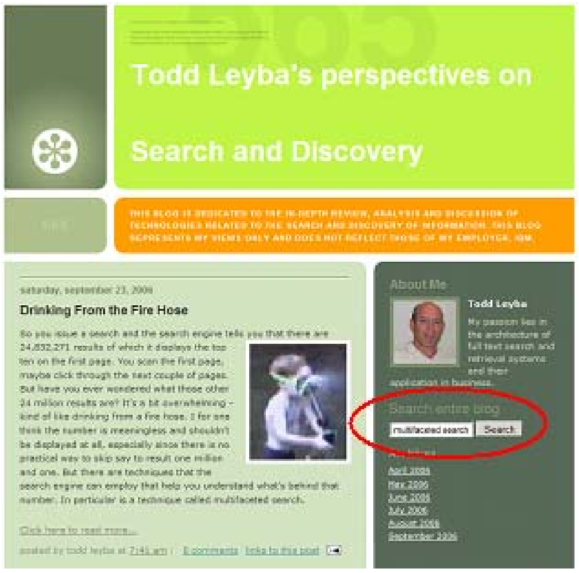
Adding Your Own Search Box

The first step is to add your own search box and button to the Web page. I will use standard HTML to add these components and a small amount of JavaScript to handle the onClick action when the button is pressed.

The following three lines add the search box and button right before the list of "Archives" in the right hand panel of my blog:

```
<h2 class="sidebar-title">Search entire blog</h2>
   <input type="text" name="Query" value="" size="25">
   <input type="button" value="Search"  onclick="runSearch()">
```

The screen shot to the right shows the search box and button added to my blog (circled in red).



Todd Leyba's perspectives on

Search and Discovery

THIS BLOG IS DEDICATED TO THE IN-DEPTH REVIEW, ANALYSIS AND DISCUSSION OF TECHNOLOGIES RELATED TO THE SEARCH AND DISCOVERY OF INFORMATION. THIS BLOG REPRESENTS MY VIEWS ONLY AND DOES NOT REFLECT THOSE OF MY EMPLOYER, IBM.

saturday, september 23, 2006

Drinking From the Fire Hose

So you issue a search and the search engine tells you that there are 24,832,271 results of which it displays the top ten on the first page. You scan the first page, maybe click through the next couple of pages. But have you ever wondered what those other 24 million results are? It's a bit overwhelming – kind of like drinking from a fire hose. I for one think the number is meaningless and shouldn't be displayed at all, especially since there is no practical way to skip say to result one million and one. But there are techniques that the search engine can employ that help you understand what's behind that number. In particular is a technique called multifaceted search.

Click here to read more...

posted by todd leyba at 7:41 am | 0 comments links to this post

About Me

Todd Leyba

My passion lies in the architecture of full text search and retrieval systems and their application in business.

Now we need to provide a JavaScript function to handle the onClick action when the button is pressed. Right before the body tag in my Blogger template, I insert the following JavaScript:

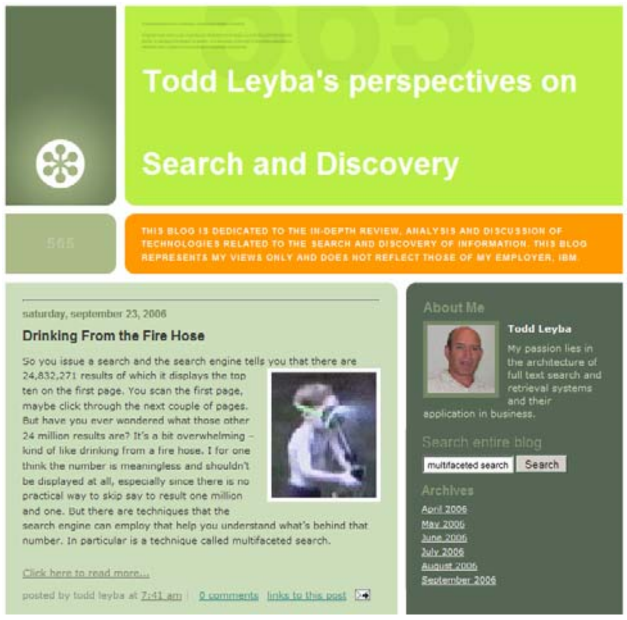
```
<SCRIPT LANGUAGE="JavaScript">

<!—Begin

  function runSearch()
  {
    var dest    = http://OmniFindHost:8080/search?;
    var params  = "index=Default&start=0&results=10&query=";
    var request = dest + params + escape(document.forms[0].Query.value);

    window.open(request,                        // complete search url
                "OmniFind Search Results",  //  Title of the window
                toolbar=1,                      // toolbar provides back/fwd
                resizable=1,                    // allow them to resize window
                scrollbars=1,                   // and to scroll as well
                height=500,                     // and I like smaller windows
                width=400,                      // of this size and position
                left=80,top=80");
  }
// End -->

</SCRIPT>
```

Invoking Search

The primary job of the JavaScript function is to collect the keywords entered in the search box and include them in an OmniFind search request. In this first example, I invoke the OmniFind search page directly with no modifications. The results page will appear as shown in Figure 1 above and is the same search results page presented as if you had entered the search directly in OmniFind. The only difference is that we have used our own search box to accept the search expression.

The URL is similar to the direct link described above but is further qualified with a few additional parameters. In the JavaScript, I broke up the URL into its constituent parts for readability. There are four parameters used. The index that I created to search is named "Default". The number of results to return is 10 starting with result zero. If you wanted to return the second page of results for the same query the "start" parameter would be set to 10.

The “request” variable contains the concatenation of the URL parts and appends the query terms provided by the user to the end of the “query” parameter. Note that I used the escape function to convert blanks and other special characters to their escaped representation. The “request” variable containing the fully built OmniFind URL is then passed as the first parameter to the window.open() function call. The window.open call will submit the request and cause a new window to be opened with the results of the search. I added a few parameters to the window.open call to control the size, location, and options of the window. Below is an example search from my blog.



Todd Leyba's perspectives on Search and Discovery

THIS BLOG IS DEDICATED TO THE IN-DEPTH REVIEW, ANALYSIS AND DISCUSSION OF TECHNOLOGIES RELATED TO THE SEARCH AND DISCOVERY OF INFORMATION. THIS BLOG REPRESENTS MY VIEWS ONLY AND DOES NOT REFLECT THOSE OF MY EMPLOYER, IBM

saturday, september 23, 2006

Drinking From the Fire Hose

So you issue a search and the search engine tells you that there are 24,832,271 results of which it displays the top ten on the first page. You scan the first page, maybe click through the next couple of pages. But have you ever wondered what those other 24 million results are? It's a bit overwhelming – kind of like drinking from a fire hose. I for one think the number is meaningless and shouldn't be displayed at all, especially since there is no practical way to skip say to result one million and one. But there are techniques that the search engine can employ that help you understand what's behind that number. In particular is a technique called multifaceted search.

Click here to read more...

posted by todd leyba at 7:41 am | 0 comments links to this post

About Me

Todd Leyba

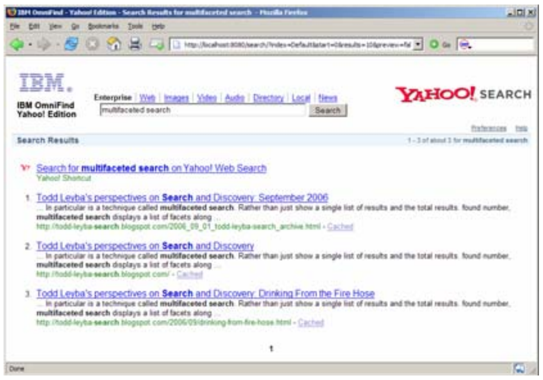
My passion lies in the architecture of full text search and retrieval systems and their application in business.

Search entire blog

multifaceted search | Search

Archives

April 2006
May 2006
June 2006
July 2006
August 2006
September 2006



IBM OmniFind - Yahoo! Edition - Search Results for multifaceted search - Mozilla Firefox

IBM OmniFind Yahoo! Edition

Enterprise | Web | Images | Video | Audio | Directory | Local | News

multifaceted search | Search

Search Results

Search for **multifaceted search** on Yahoo! Web Search
Yahoo! Shortcut

1. Todd Leyba's perspectives on **Search** and Discovery: September 2006
... In particular is a technique called **multifaceted search**. Rather than just show a single list of results and the total results found number, **multifaceted search** displays a list of facets along ...
http://todd-leyba-**search**.blogspot.com/2006_09_01_todd-leyba-search_archive.html - Cached

2. Todd Leyba's perspectives on **Search** and Discovery
... In particular is a technique called **multifaceted search**. Rather than just show a single list of results and the total results found number, **multifaceted search** displays a list of facets along ...
http://todd-leyba-**search**.blogspot.com/ - Cached

3. Todd Leyba's perspectives on **Search** and Discovery: Drinking From the Fire Hose
... In particular is a technique called **multifaceted search**. Rather than just show a single list of results and the total results found number, **multifaceted search** displays a list of facets along ...
http://todd-leyba-**search**.blogspot.com/2006/09/drinking-from-fire-hose.html - Cached

1

Using the OmniFind REST API

Up to this point we have successfully used our own search box to submit a search rather than the search box that appears on the OmniFind search page. We now would like to change the appearance of the search results more so than what is offered in the OmniFind layout editor. This can be accomplished with the help of the OmniFind search API.

OmniFind’s API is REST based which means that you use a standard HTTP GET request with parameters to submit the search. The search results are returned as XML, which we then transform into our custom HTML using XSLT.

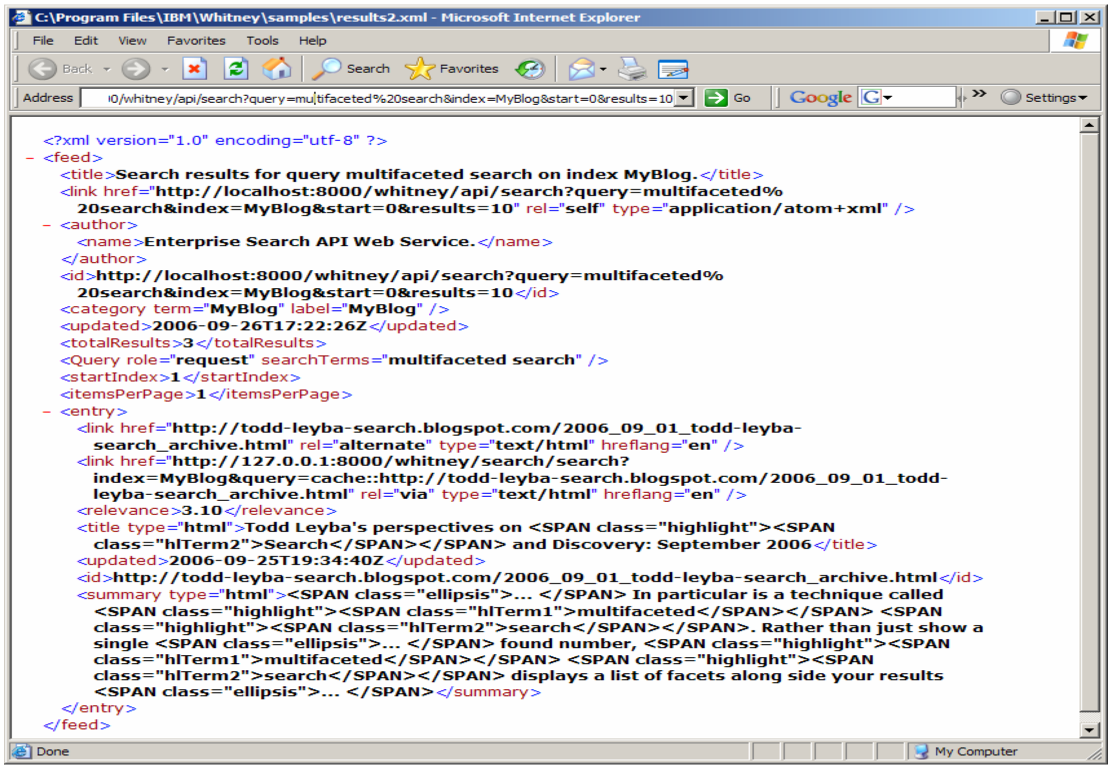
Below is an example OmniFind search request:

```
http://OmnifindHost:8080/api/search?index=Default&results=10&star
t=0&query=conferences
```

You may have noticed that the above URL is nearly identical to the URL issued in the previous example with one exception. The sub domain "/api/search" is used instead of "/search". This instructs OmniFind to return the results as XML instead of the fully formatted HTML page shown in Figure 1.

The XML that is returned conforms to an ATOM 1.0 feed. Consequently, you can test your API search requests by using any conventional RSS feed reader that supports ATOM 1.0 (I personally use the open source FeedReader program). The feed reader program automatically issues the search and formats the results for you.

You can also test your API searches with a standard browser which will display the returned XML natively as shown below.

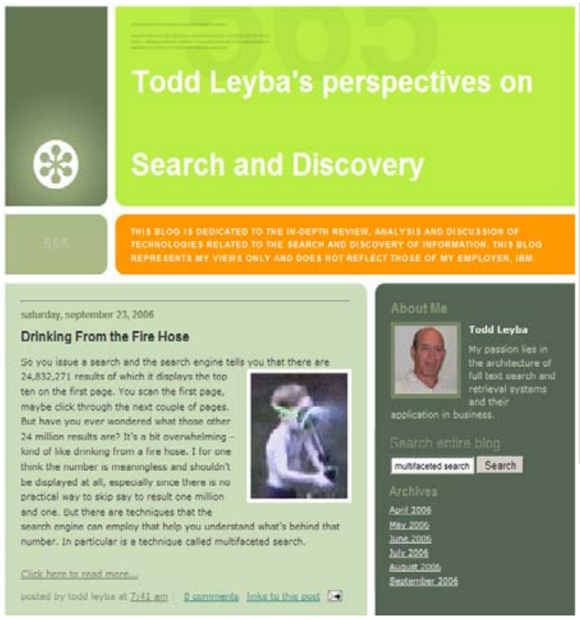

C:\Program Files\IBM\Whitney\samples\results2.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address 0/whitney/api/search?query=multifaceted%20search&index=MyBlog&start=0&results=10 Go
Google Settings
Done
My Computer

```
  <?xml version="1.0" encoding="utf-8" ?>
- <feed>
    <title>Search results for query multifaceted search on index MyBlog.</title>
    <link href="http://localhost:8000/whitney/api/search?query=multifaceted%
      20search&index=MyBlog&start=0&results=10" rel="self" type="application/atom+xml" />
  - <author>
      <name>Enterprise Search API Web Service.</name>
    </author>
    <id>http://localhost:8000/whitney/api/search?query=multifaceted%
      20search&index=MyBlog&start=0&results=10</id>
    <category term="MyBlog" label="MyBlog" />
    <updated>2006-09-26T17:22:26Z</updated>
    <totalResults>3</totalResults>
    <Query role="request" searchTerms="multifaceted search" />
    <startIndex>1</startIndex>
    <itemsPerPage>1</itemsPerPage>
  - <entry>
      <link href="http://todd-leyba-search.blogspot.com/2006_09_01_todd-leyba-
        search_archive.html" rel="alternate" type="text/html" hreflang="en" />
      <link href="http://127.0.0.1:8000/whitney/search/search?
        index=MyBlog&query=cache::http://todd-leyba-search.blogspot.com/2006_09_01_todd-
        leyba-search_archive.html" rel="via" type="text/html" hreflang="en" />
      <relevance>3.10</relevance>
      <title type="html">Todd Leyba's perspectives on <SPAN class="highlight"><SPAN
        class="hlTerm2">Search</SPAN></SPAN> and Discovery: September 2006</title>
      <updated>2006-09-25T19:34:40Z</updated>
      <id>http://todd-leyba-search.blogspot.com/2006_09_01_todd-leyba-search_archive.html</id>
      <summary type="html"><SPAN class="ellipsis">... </SPAN> In particular is a technique called
        <SPAN class="highlight"><SPAN class="hlTerm1">multifaceted</SPAN></SPAN> <SPAN
        class="highlight"><SPAN class="hlTerm2">search</SPAN></SPAN>. Rather than just show a
        single <SPAN class="ellipsis">... </SPAN> found number, <SPAN class="highlight"><SPAN
        class="hlTerm1">multifaceted</SPAN></SPAN> <SPAN class="highlight"><SPAN
        class="hlTerm2">search</SPAN></SPAN> displays a list of facets along side your results
        <SPAN class="ellipsis">... </SPAN></summary>
    </entry>
  </feed>
```

Results Returned as HTML Snippets

Before we further discuss the XML that is normally returned it is important to note that the output of the search API can also return the results as a snippet of standard HTML. I refer to the output as a snippet because it is not a complete HTML page (no <HTML><BODY> tags). HTML output is indicated with the "output=snippet" parameter on the search request, which is shown with its effect below:

```
http://OmnifindHost:8080/api/search?index=Default&results=10&start=0&query=conferences&output=snippet
```



Todd Leyba's perspectives on

Search and Discovery

THIS BLOG IS DEDICATED TO THE IN-DEPTH REVIEW, ANALYSIS AND DISCUSSION OF TECHNOLOGIES RELATED TO THE SEARCH AND DISCOVERY OF INFORMATION. THIS BLOG REPRESENTS MY VIEWS ONLY AND DOES NOT REFLECT THOSE OF MY EMPLOYER, IBM.

saturday, september 23, 2006

Drinking From the Fire Hose

So you issue a search and the search engine tells you that there are 24,832,271 results of which it displays the top ten on the first page. You scan the first page, maybe click through the next couple of pages. But have you ever wondered what those other 24 million results are? It's a bit overwhelming – kind of like drinking from a fire hose. I for one think the number is meaningless and shouldn't be displayed at all, especially since there is no practical way to skip say to result one million and one. But there are techniques that the search engine can employ that help you understand what's behind that number. In particular is a technique called multifaceted search.

Click here to read more...

posted by todd leyba at 7:41 am | 0 comments links to this post

About Me

Todd Leyba

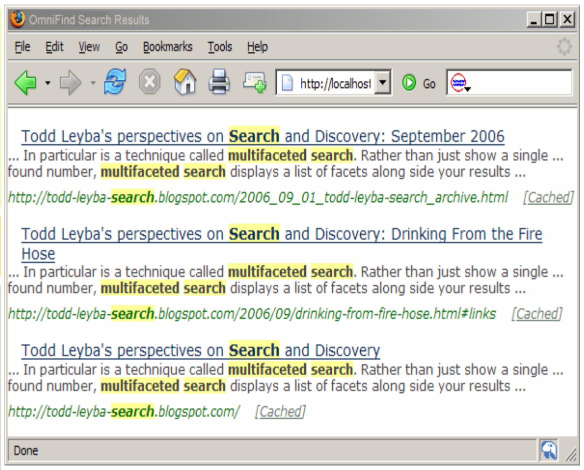
My passion lies in the architecture of full text search and retrieval systems and their application in business.

Search entire blog

multifaceted search Search

Archives

April 2006
May 2006
June 2006
July 2006
August 2006
September 2006



Todd Leyba's perspectives on Search and Discovery: September 2006
... In particular is a technique called multifaceted search. Rather than just show a single ... found number, multifaceted search displays a list of facets along side your results ...
http://todd-leyba-search.blogspot.com/2006_09_01_todd-leyba-search_archive.html [Cached]

Todd Leyba's perspectives on Search and Discovery: Drinking From the Fire Hose
... In particular is a technique called multifaceted search. Rather than just show a single ... found number, multifaceted search displays a list of facets along side your results ...
http://todd-leyba-search.blogspot.com/2006/09/drinking-from-fire-hose.html#links [Cached]

Todd Leyba's perspectives on Search and Discovery
... In particular is a technique called multifaceted search. Rather than just show a single ... found number, multifaceted search displays a list of facets along side your results ...
http://todd-leyba-search.blogspot.com/ [Cached]

Done

Notice that the format of the results are somewhat similar to those in the OmniFind search results page with the exception of the missing search box and page controls. This approach has value in certain applications but is somewhat inflexible. If you want to change the HTML formatting you would need to parse the HTML yourself. Not an easy task to do.

Using XSLT to Format the Search Results

Because results are normally returned as XML, you have the ability to use an XSLT style sheet to transform the XML into HTML and format the results as desired. The XSLT style sheet would be prepared by you and would contain the appropriate XSL and XPath directives to process the XML ATOM feed elements.

In this case I would like the motif of the results page to match that of my blog using the same color schemes and fonts. Below is the XSLT stylesheet that I used, which I stored in a file named “myStyleSheet.xsl”. Each line is numbered for easy reference.

```
1   <?xml version="1.0" encoding="utf-8"?>
2   <xsl:stylesheet version="1.0"
3      xmlns:atom="http://www.w3.org/2005/Atom"
4      xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
5      xmlns:dc="http://purl.org/dc/elements/1.1/">
6   <xsl:output method="html"/>
7   <xsl:template match="/">
8      <style>
9          <xsl:comment>
10          .content-area {background-color: #dcedcb;}
11         .description {font-size: .9em; margin: 0 0 10px 0;}
12         // Remaining styles omitted for readability
13          </xsl:comment>
14     </style>
15     <xsl:apply-templates select="/atom:feed"/>
16  </xsl:template>
17  <xsl:template match="/atom:feed">
18     <div class="content-area">
19       <div class="title"><xsl:value-of select="atom:title"/></div>
20       <ol class="list"><xsl:apply-templates select="atom:entry"/></ol>
21     </div>
22  </xsl:template>
23  <xsl:template match="atom:entry">
24     <li>
25        <a href="{atom:link/@href}">
26           <xsl:value-of select="atom:title" disable-output-escaping="yes"/>
27        </a>
28        <div class="list-item-description">
29         <xsl:value-of select="atom:summary" disable-output-escaping="yes"/>
30        </div>
31     </li>
32  </xsl:template>
33  </xsl:stylesheet>
```

XSLT can be used to transform XML into a variety of formats. Line 6 indicates that the output of this transformation is to be HTML. I use XSL templates to match the various elements in the XML stream. Line 7 is the main template and matches on all elements in the XML file. It is within this template that you specify your styles inserted between the xsl:comment directives (Lines 9 & 13). For readability I omitted the majority of my style directives.

Within this main template I reference a subordinate template for each atom:feed element XSLT encounters. In this case there is only one atom:feed element specified in the OmniFind XML results.

The template for an atom:feed element begins on line 17. It creates an outer HTML “div” tag whose style class is “content-area”. Note that a style for each specified “div” class attribute must be defined above in the main template. Again I purposely omitted most of the style definitions for readability. The atom:feed template creates an inner “div” tag for the title of the results (line 19). The title is actually pulled from the XML using the xsl:value-of statement with a select on the element named “atom:title”. If you want to provide a different hard coded title, just replace line 19 with your own HTML statement (for example, <h2>My Title</h2>). Line 20 inserts an HTML ordered list and applies another sub template for each “atom:entry” element found in the XML.

The last template definition (starting on line 23) provides the HTML transformations to be applied to each search result. In this template I create an HTML line item tag for the ordered list, a link to the document, and a brief description of the document. For the link URL I use the XPath directive `(atom:link/@href)` to extract that value of the href attribute in the atom:link element (line 25). For the anchor text itself I use the xsl:value-of directive to extract the contents of the atom:title element within the entry element (line 26). The same technique is used for the result description as well.

OmniFind conveniently highlights any search terms contained in the title and summary of each result. It does this by bracketing the encountered search terms with HTML <SPAN> tags to indicate the style to be used for the highlighting. These HTML tags are embedded within the original XML and are normally escaped by the XSLT processor. This causes the HTML tags to be shown as is when displayed in the browser (an effect we do not want). The XSLT processor does not know that these are valid HTML tags and dutifully escapes any special characters it encounters during the processing of the xsl:value-of directive. Under these circumstances we can instruct the XSLT processor not to escape any special characters with the `disable-output-escaping="yes"` attribute on lines 26 and 29.

It is important to note that the `disable-output-escaping` attribute is not honored by all browsers. Microsoft’s Internet Explorer disables the output escaping with the desired effect but not according to the W3 XSLT specification. Mozilla on the other hand ignores the attribute so as to be in compliance with the XSLT specification. For Mozilla browsers you can achieve the same effect with different XSLT commands (which are not shown in the example).

Using the Stylesheet Parameter

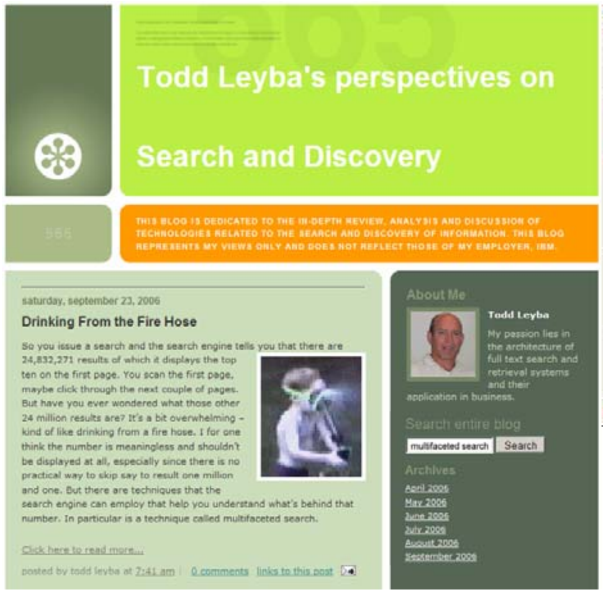
The “stylesheet” parameter is used on the search request to indicate which XSLT style sheet is to be applied as shown below.

```
http://OmnifindHost:8080/api/search?query=digital%20cameras&index=Default&resul
ts=10&start=0&stylesheet=”http://myserver.com/myStyleSheet.xsl”
```

The use of the “stylesheet” parameter causes OmniFind to insert an xml-stylesheet entry as the second line of the XML search results (see below in red).

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/xsl" href=" http://myserver.com/myStyleSheet.xsl"?>
<feed xmlns="http://www.w3.org/2005/Atom" xmlns:opensearch=http://a9.com/-/spec/opensearch/1.1/
    •
    •
    •
</feed>
```

This will cause the browser to retrieve the stylesheet from the location specified in the HREF and apply it by using XSLT. The results are shown below:



Todd Leyba's perspectives on

Search and Discovery

THIS BLOG IS DEDICATED TO THE IN-DEPTH REVIEW, ANALYSIS AND DISCUSSION OF TECHNOLOGIES RELATED TO THE SEARCH AND DISCOVERY OF INFORMATION. THIS BLOG REPRESENTS MY VIEWS ONLY AND DOES NOT REFLECT THOSE OF MY EMPLOYER, IBM.

saturday, september 23, 2006

Drinking From the Fire Hose

So you issue a search and the search engine tells you that there are 24,832,271 results of which it displays the top ten on the first page. You scan the first page, maybe click through the next couple of pages. But have you ever wondered what those other 24 million results are? It's a bit overwhelming - kind of like drinking from a fire hose. I for one think the number is meaningless and shouldn't be displayed at all, especially since there is no practical way to skip say to result one million and one. But there are techniques that the search engine can employ that help you understand what's behind that number. In particular is a technique called multifaceted search.

Click here to read more...

posted by todd leyba at 7:41 am | 0 comments links to this post

About Me

Todd Leyba

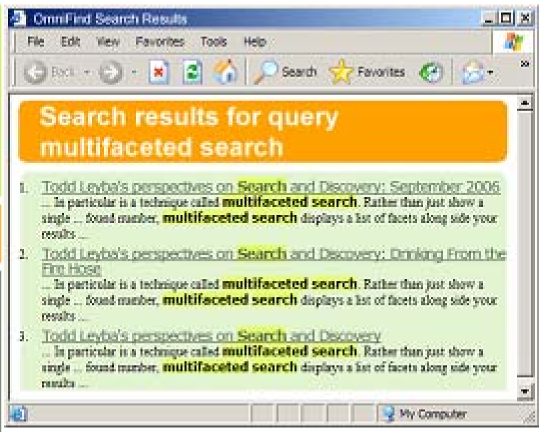
My passion lies in the architecture of full text search and retrieval systems and their application in business.

Search entire blog

multifaceted search

Archives

April 2006
May 2006
June 2006
July 2006
August 2006
September 2006



Search results for query multifaceted search

1. Todd Leyba's perspectives on Search and Discovery: September 2006
... In particular is a technique called **multifaceted search**. Rather than just show a single ... found number, **multifaceted search** displays a list of facets along side your results ...
2. Todd Leyba's perspectives on Search and Discovery: Drinking From the Fire Hose
... In particular is a technique called **multifaceted search**. Rather than just show a single ... found number, **multifaceted search** displays a list of facets along side your results ...
3. Todd Leyba's perspectives on Search and Discovery
... In particular is a technique called **multifaceted search**. Rather than just show a single ... found number, **multifaceted search** displays a list of facets along side your results ...

Summary

In this article I presented four methods for integrating OmniFind search functionality into your Web site. The first and simplest approach was to insert links to the OmniFind search page directly. Next was to replace the direct links with my own search box and button but to keep the OmniFind search results page unchanged. I then switched over to using the OmniFind search API so as to better control the formatting of the returned search results. I first showed how the API can return the search results as a snippet of HTML and then ultimately XML. I finally demonstrated how an XSLT Stylesheet can be applied to the XML to create completely customized search results.