

# WebGUI Search

*The Game is Afoot*



Presented by  
William McKee

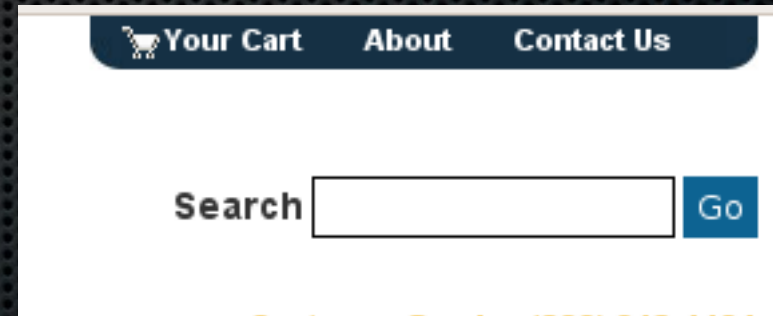


# The Importance of Search

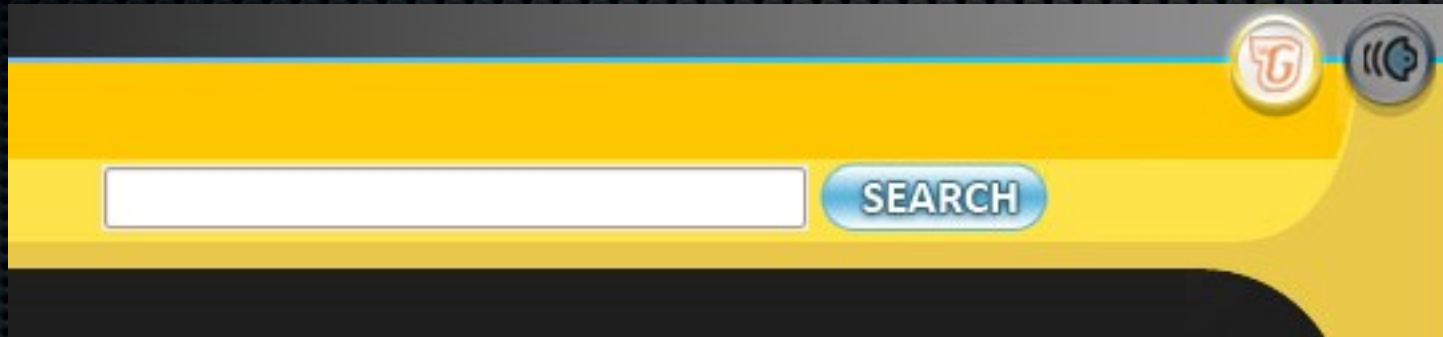
Search comprises two of “the five areas that a user is most likely to interact with a site’s information architecture”.



Louis Rosenfield, Information Architecture Consultant  
Author of Information Architecture (O'Reilly)



# WebGUI Search



- Multiple Iterations
  - external programs, self-contained wobjects, 3rd party search engines
- Current status

*Currently WebGUI's search is the most powerful, flexible, and the most integrated it's ever been.*

JT Smith, Dec 2007  
"Searching WebGUI", The Black Blog

# *Features of WebGUI Search*

- Full-site or sub-site search
- Natural language and Boolean searching
- Search assets and attachments
- Real-time indexing
- Limit results by asset type
- Permissions-aware results
- Support for multiple search forms
- Plug-in support for attachment indexing
- Return result URL's as page containers or individual assets

# Let's Get Started – Adding WebGUI Search to your site

- Adding search to your web site is easy
  - Create a page, let's call it “Search”
  - Add a Search widget, let's call it “Site Search”



The screenshot shows a web interface for a search widget. At the top, the title "Site Search" is displayed in a bold, black font. Below the title is a search input field containing the text "latest" and a "search" button. The results are listed below the input field, each starting with a red underlined heading:

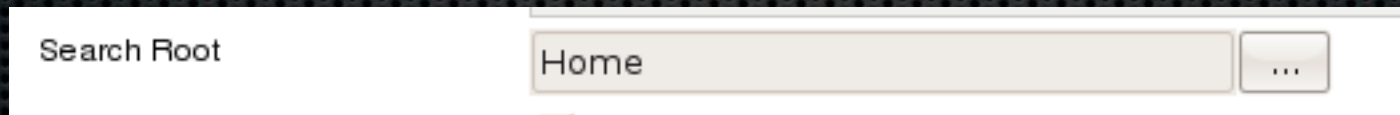
- The Latest News**  
This is the latest news from Plain Black and WebGUI pulled directly from the site every hour.
- The Latest News**  
The Latest News The Latest News the latest news
- Welcome**  
The WebGUI Content Engine® is a powerful and easy to use system for managing web sites, you can extend it to match your needs. It's easy eno

# *Caveats and Gotchas*

- Page Layouts that have a matching title but no content will show up in search results
  - Workaround: Exclude WebGUI::Asset::Layout
- Stopwords are not indexed
  - These are common words that do not add strength to the search and return too many matches, e.g., “a”, “the”, “getting”, “from”, “all”, “sure”, “via”
  - See MySQL Manual for the complete list – 11.8.4. Full-Text Stopwords
- Limited support for substring matching
  - “int” matches “integer” but not “flint”

# Advanced Options – Search Root

The “Search Root” property allows you to limit the results of your search to a portion of your web site.

A screenshot of a configuration interface for the 'Search Root' property. The label 'Search Root' is on the left. To its right is a text input field containing the word 'Home'. To the right of the input field is a small square button with three dots inside, indicating a dropdown menu.

Limiting the area being searched (aka, the lineage) will improve performance.

# Advanced Options – Limit Assets

The “Limit Asset classes to” property allows you to limit the results of your search to specific types of assets (e.g., files, images, wiki entries).

Limit Asset classes to:

- WebGUI::Asset
- WebGUI::Asset::File
- WebGUI::Asset::File::Image
- WebGUI::Asset::RichEdit
- WebGUI::Asset::Snippet
- WebGUI::Asset::Template
- WebGUI::Asset::Wobject::Article



**Limiting the assets being searched will improve performance.**

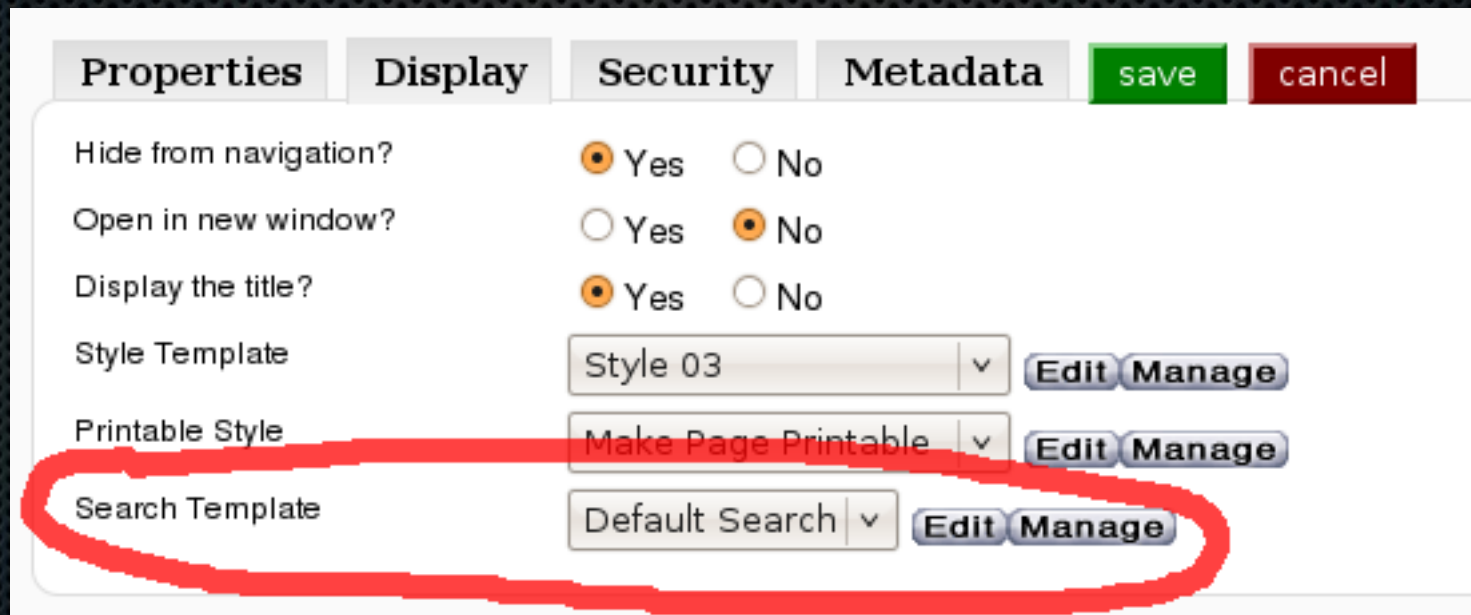
# *Advanced Options – Container URL's*

The newly added “User container URLs in results” property allows you to force results links to go to the container asset (e.g., Page Layout) instead of to the individual asset which often resulted in unexpected display problems.

Use container URLs in results?  Yes  No

# Controlling the Results

- The Search Template (Display tab) is a template that contains both the form and results
- WebGUI ships with only one default template



The screenshot shows a configuration window with four tabs: Properties, Display, Security, and Metadata. The Display tab is active. At the top right of the window are 'save' and 'cancel' buttons. The configuration options are as follows:

Property	Value	Buttons
Hide from navigation?	<input checked="" type="radio"/> Yes <input type="radio"/> No	
Open in new window?	<input type="radio"/> Yes <input checked="" type="radio"/> No	
Display the title?	<input checked="" type="radio"/> Yes <input type="radio"/> No	
Style Template	Style 03	Edit Manage
Printable Style	Make Page Printable	Edit Manage
Search Template	Default Search	Edit Manage

A red hand-drawn circle highlights the 'Search Template' row, which is set to 'Default Search'.

# Moving Right Along – Understanding How WebGUI Search Works



MySQL Database



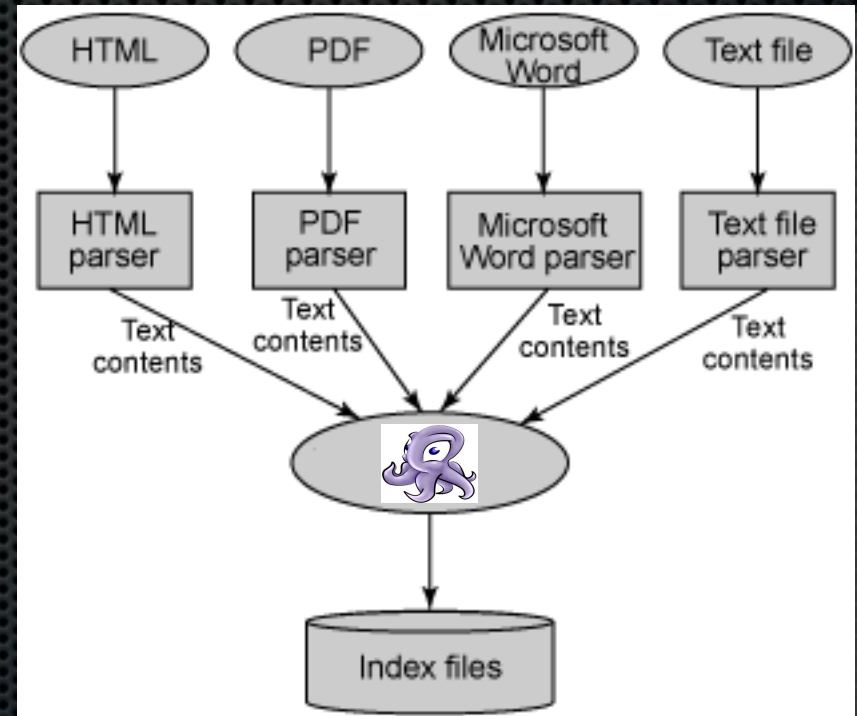
WebGUI Content Engine

Your Web Site



# Components of a Search System

- Indexing Subsystem
  - collect
  - parse
  - store
- Retrieval Subsystem
  - query
  - filter
  - display



# *Asset Indexing*

- Assets are indexed in real-time as they are created or updated
- The keywords of an asset are derived from the following fields:
  - title
  - menuTitle
  - synopsis
  - url
  - description
  - keywords
- TODO: The search synopsis is derived from either the synopsis of the asset, the first 256 characters of the description or first 256 characters the keywords

# *Attachment Indexing*

- In addition to the Asset indexing steps, File assets also attempt to index the attachment
- Default supported attachments
  - MS Word
  - Rich Text Format (RTF)
  - Excel
  - Powerpoint
  - PDF
  - Text
  - HTML
- Configured via WebGUI conf file (see Bonus Slide for more details)

# *Index Storage*

- The indexed keywords are stored in MySQL (assetIndex table)
- Stores the Title & Synopsis for each asset that is displayed on the results page
- Synopsis is one of the following:
  - Asset synopsis
  - Asset description (first 255)
  - Derived from joining title, menuTitle, url, keywords
- Stores the keywords and other meta data for each asset

# *Data Retrieval – Query*

- Searches include all words using AND matching
- Searches are case-insensitive
- Templates and system internals are filtered from the results
- Results are ordered by a “best effort” scoring mechanism using an MySQL query
- There is no relevancy weighting based on title, keywords, keyword location, or other complex algorithms
- Returns 25 results per page\*

# Boolean Search Primer – Filter

- WebGUI supports all the Boolean search operators
- Common Operators
  - The double-quote operator ("")
    - “"latest news"”
    - “"Plain Black"”
  - The asterisk, or wildcard, operator (\*)
    - “Bl\*” - finds Bloody, Black, Blog
  - The +/- operators
    - “+news +latest” - requires 'news' and 'latest'
    - “+news -latest” - requires 'news', filters 'latest'

# *Customizing the Search Results*

- Form modifications
  - My favorite hack is to submit back to layout page, not the search asset
  - Provide feedback on no matches with `no_results`
- Results modifications
  - For each match, you have the following fields available: `url`, `title`, `synopsis`, `assetId`
  - Results can be paged in increments of 10 or 20
  - See “Search Template” in help system for complete list of variables available

# Custom Search Template

## Search

### Site Search

[Search Help](#)

Search Results for **news**

Page 1 of 1

#### [The Latest News](#)

This is the latest news from Plain Black and WebGUI pulled directly from the site every hour.

#### [The Latest News](#)

A bus is not an integer.

#### [What should you do next?](#)

Here's the news!

#### [Welcome](#)

The WebGUI Content Engine® is a powerful and easy to use system for managing web sites, and building web applications. It provides thousands of features out of the box, and lots of plug-in points so you can extend it to match your needs. It's easy eno

# Complex Searches

Find a Doctor   Careers   Contact Us   Referring Physician Portal   Staff Portal

Midwest Heart Specialists  
SINCE 1975

Search

Home   Programs & Services   Patient Resources   Request An Appointment   News   Locations   About Us

Home   Wednesday, July 30, 2008

### Find A Doctor


by location  
Please Select

by last name  
Please Select

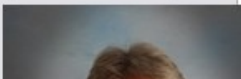
by specialty  
Please Select

### Physicians

[Vincent Bufalino, M.D.](#)



[Thomas Discher, M.D.](#)



### Search

cardiology   search

[Interventional Cardiology](#)  
Interventional Cardiology Angiography is a special type of x-ray that allows your coronary arteries to be viewed and recorded on film. Your doctor can see if the blood vessels to your heart are clogged. (Also see Heart Procedure - Angiography).

[Interventional Cardiology](#)  
Interventional Cardiology Interventional Cardiology resources education patient education interventional cardiology

[Interventional Cardiology](#)  
Interventional Cardiology Interventional Cardiology services programs services interventional cardiology3

[Interventional Cardiology](#)  
Interventional Cardiology Advanced cardiac technology in the hands of our board-certified cardiologists can potentially stop a heart attack in progress, minimize damage to the heart muscle, or even prevent a heart attack. We have extensive experience in [March 18, 2002 - Midwest Heart Specialist Offers Expanded Services to Northwest Suburbs](#)

Midwest Heart Specialist Offers Expanded Services to Northwest Suburbs March 18, 2002 Downers Grove, IL Midwest Heart Specialists, one of the largest private cardiology practices in Illinois has merged with

[Swan Ganz](#)  
Swan Ganz Swan Ganz resources education patient education interventional cardiology swan ganz


### Other Resources

[American Heart Association](#)

[Heart Authority](#)

[National Heart & Blood Institute](#)

[MedlinePlus](#)

Select your text size: 

# Complex Searches (pt 2)

- Requirements
  - Search entire web site including site content, Health Library, and News
  - Return list of doctors whose specialty matches any of the search term

The screenshot shows the website for Midwest Heart Specialists, established in 1973. The search results for the term "cardiology" are displayed. The page includes a navigation menu with links for Home, Programs & Services, Patient Resources, Request An Appointment, News, Locations, and About Us. A search bar at the top right contains the text "Search". The search results are organized into several sections:

- Find A Doctor:** A sidebar with dropdown menus for "by location", "by last name", and "by specialty", each with a "Please Select" option.
- Physicians:** A list of doctors with their names and photos. Visible names include Vincent Bufalino, M.D. and Thomas Discher, M.D.
- Search Results:** The main content area shows the search term "cardiology" in a search box. Below it, there are multiple entries for "Interventional Cardiology" with detailed descriptions of the procedure and its benefits. The text describes it as a special type of x-ray that allows coronary arteries to be viewed and recorded on film, and notes that it can help identify clogged arteries. It also mentions that the procedure is performed by board-certified cardiologists and that the hospital has extensive experience in this field.
- Other Resources:** A sidebar with links to external resources such as the American Heart Association, Heart Authority, National Heart & Blood Institute, and MedlinePlus.

# Complex Searches (pt 3)

## Implementation

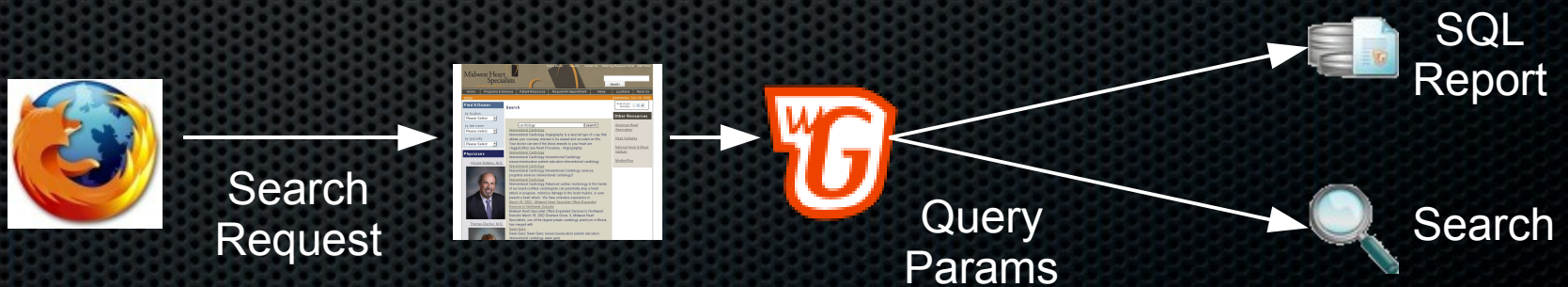
Search

SQL Report

The screenshot displays the Midwest Heart Specialists website interface. At the top, there is a navigation bar with links for 'Home', 'Programs & Services', 'Patient Resources', 'Request An Appointment', 'News', 'Locations', and 'About Us'. A search bar is located in the top right corner. Below the navigation bar, the date 'Wednesday, July 30, 2008' is displayed. The main content area is divided into several sections. On the left, there is a 'Find A Doctor' section with three dropdown menus for 'by location', 'by last name', and 'by specialty'. Below this is a 'Physicians' section featuring a profile for Vincent Bufalino, M.D., with a photo and a profile picture for Thomas Discher, M.D. On the right, there is a 'Search' section with a search box containing the text 'cardiology' and a 'search' button. Below the search box, there is a list of search results for 'Interventional Cardiology', including a detailed description of the procedure and a link to a news article titled 'March 18, 2002 - Midwest Heart Specialist Offers Expanded Services to Northwest Suburbs'. The search results are enclosed in a red rectangular box. To the right of the search results, there is an 'Other Resources' section with links to 'American Heart Association', 'Heart Authority', 'National Heart & Blood Institute', and 'MedlinePlus'. A text size selector is also visible in the top right corner.

# Complex Searches (pt 4)

- How it works
  - Search terms are submitted to the Page Layout
  - WebGUI passes the terms to all assets on the page
  - Page is rendered
    - SQL Report is generated
    - Search object returns matching results



# *SysAdmin Tasks*

- Customize MySQL Configuration
  - Customize stopwords, minimum and maximum word lengths
- Re-Indexing Your Site
  - Added new plug-ins and want to re-index existing content
  - Changes to the search system (including bug fixes, custom assets)
  - Database changes
    - configuration changes
    - external content import
    - site splits or merges



# Search.pl

- WebGUI/sbin/search.pl
- Supports reindexing one or all sites, updating the index, and searching the index\*

```
$ /data/wre/sbin/setenvironment.sh
```

```
$ cd /data/WebGUI/sbin
```

```
$ perl search.pl --configFile config.conf --indexsite
```

```
$ perl search.pl --configFile config.conf --search news
```

\* Returns all asset types including public and non-public assets (e.g., templates)

# Looking into the Future

- JT's Thoughts on the Future of Search\*  
(“Searching WebGUI”, TBB)
  - Pattern filters
  - Term highlighting
  - Relevancy rules
  - Narrowable results
  - Keyword tags
  - “Buildable” asset manager searches



\* Subject to change!

# *Additional Resources*

- WebGUI
  - *WebGUI Content Manager's Guide*
  - WebGUI.org Forums & Wiki
  - IRC – #webgui on freenode
- MySQL
  - MySQL Manual
    - Section 11.8 Full-Text Search Functions
  - MySQL Developer Zone
    - "The Full-Text Stuff That We Didn't Put In The Manual"



***Thank You***

Questions? Comments?

**Contact Me:**

William McKee

*william@knowmad.com*

<http://www.knowmad.com>



# *Extending the Attachment Indexer*

- Attachments are indexed using an external program
- The default indexers are included with the WRE
- WebGUI uses the file extension determine what application to use for indexing
- You can add as many additional indexers or file extensions as you would like to the site configuration file (SearchIndexerPlugins)

# Configuring MySQL Full-text Search

- Several configuration variables affect full-text search
  - `ft_min_word_len` – defaults to 4 (2 for WRE v0.8.3)
  - `ft_max_word_len` – default is version dependent (84 for WRE v0.8.3)
  - if you change either of these, you must rebuild your FULLTEXT indexes
  - use “`SHOW VARIABLES LIKE 'ft%'`” to view them all
- Currently, full-text searches are supported for MyISAM tables only

# *Programmer's Primer – MySQL*

- The “search” table contains the actual search assets that have been added to your site
- The “assetIndex” table contains all indexed assets including
  - title
  - synopsis
  - url
  - keywords
  - lineage

# *Programmer's Primer – WebGUI*

- WebGUI::Search contains methods for searching content
- WebGUI::Search::Index contains methods for indexing content
- WebGUI::Asset::Wobject::Search contains the wobject interface
- The indexContent method of WebGUI::Asset provides basic indexing but can be overridden to index files or collateral data (e.g., WebGUI::Asset::Wobject::Article)

## *Programmer's Primer (pt 2)*

- Additional search criteria are available via the API that are not part of the Search wobject
  - creationDate – set the start and/or end
  - revisionDate – set the start and/or end
  - where – add a custom WHERE clause to the query
  - columns – specify the columns to be returned
  - lineage – support for multiple lineages
- Internally, WebGUI does Boolean searches on keywords and Natural Language searches for relevancy scoring

## *My RFE's*

- Provide more details about returned data (e.g., date, relevance score)
- Control the length of the synopsis
- Better control of pagination (similar to SQL Report)
- Output ordering (e.g., by relevance, by date)
- Search query logging & reporting
- User preferences like Google or Yahoo to control number of results per page, open in new window, etc.