# Geospatial Search Engine Technical Description

## Introduction

The Geospatial Search Engine (GSE) is a web application created by the Western Wildland Environmental Threat Assessment Center (WWETAC). GSE provides a search engine for Geographic Information System (GIS) map servers and layers. GSE supports web map services (WMS), ArcGIS services, ArcIMS services, and shapefiles for download. The application is hosted in the Amazon Web Services cloud and is available at http://www.wwetac.us/GSE/GSE.aspx.

This system combines a searchable database of GIS map servers and layers, with a web crawler for both locating new data and updating existing data. GSE is currently a standalone application, but there are plans to build a user friendly API to allow further access to the search engine.

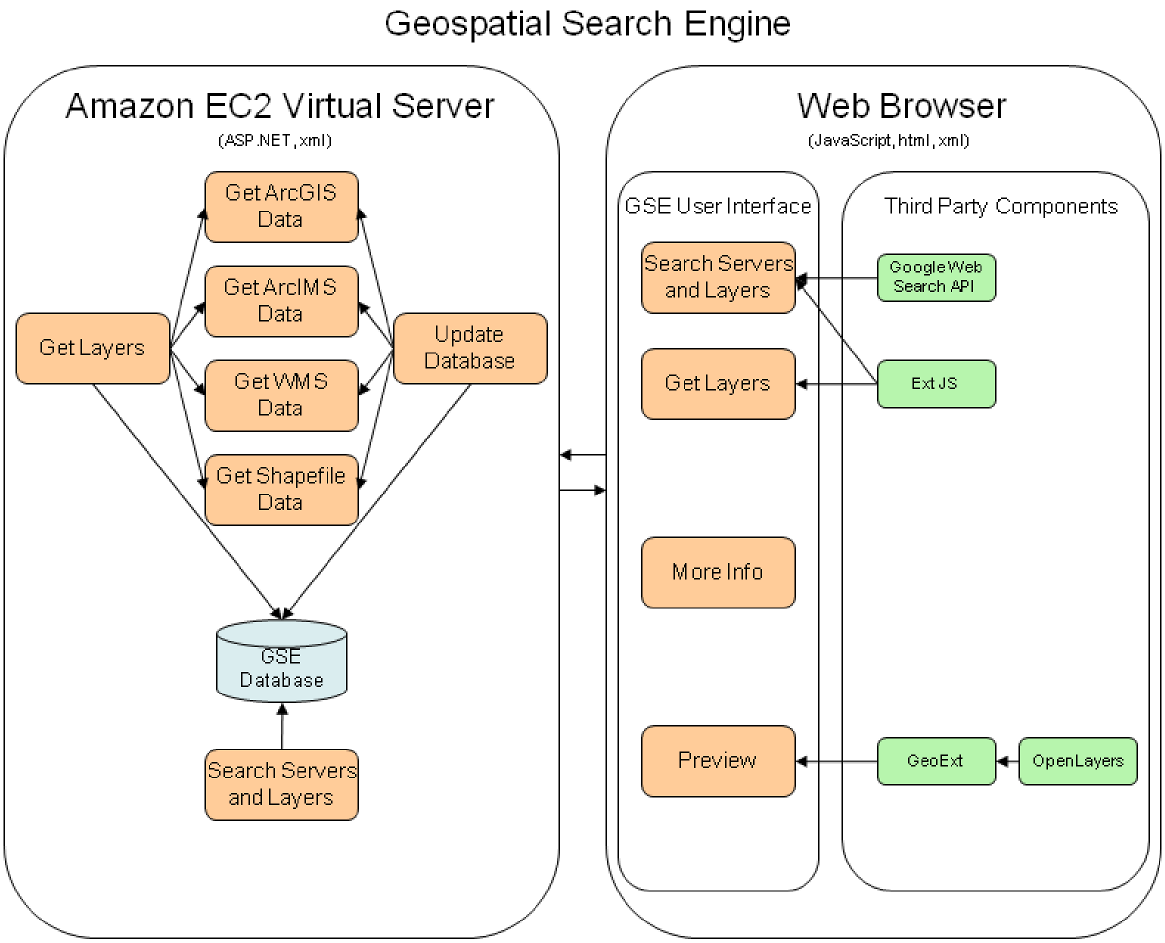


Figure 1. Gespatial Search Engine application design.

# User Interface

The GSE user interface is a browser application built using open source components such as Ext JS, GeoExt, and OpenLayers. The Google Web Search API is integrated to obtain web search results based on user specified terms. These results are used to help build the GSE database. The user interface consists of the following parts.

## Search Panel

The left side of the interface is used for specifying search terms and retrieving a list of map servers and layers. For shapefiles, a web page link is provided so that the user may review the web page containing shapefiles available to download. The specified terms are combined with additional predefined terms in a Google search that enable the system to find map servers, and websites with shapefiles to download. This process occurs in the background seamless to the user interface. In addition, the GSE database is searched using the specified terms, and results are returned to the user in a tree view. There is a separate node in the tree view for each of the 4 server types (WMS, ArcGIS Server, ArcIMS, and Shapefile). Below the server type node is a node for each map server, and each map server has nodes for each layer that meets the search criteria.  It is possible to have a map server node with no layer sub nodes.  This occurs when the map server abstract or keywords contain the search terms but none of the layer abstracts or keywords contain the terms.  As the user select nodes in the tree view, the total number of layers and the number of layers that meet the search criteria are displayed in the lower right corner of the search panel.

## View

The View button below the search panel will either display a preview map of the selected layer, or if a shapefile server is selected, open the web page.

## Get All Layers

The Get All Layers button will retrieve a list of all layers for the selected map server, rather than just the layers that meet the search criteria.  The layers will be listed in the layer list panel on the right side of the application interface.

## Layer List Panel

The right side of the interface allows the user to specify a known URL and retrieve available map layers.  When a URL is specified that is not found in the database, the GSE application will contact the map server to retrieve metadata and add this server to the database.  This allows for a convenient method of manually adding map servers to the database.  The list of layers is returned to the user and displayed in a table with Title and Description columns.  By clicking on the column headers you can sort the layers by Title or Description.

## Preview

The Preview button at the bottom of the layer list panel displays a preview map for the selected layer.  This map contains a background map for reference purposes in addition to a map of the selected layer.  On the right side of the map is a plus symbol.  Clicking this symbol opens a small window allowing the user to turn layers on and off.  On the left side of the map is a navigation bar, allowing the user to pan and zoom the map.  Panning can also be accomplished by clicking and dragging the mouse.  Holding the shift key down while clicking and dragging on the map will draw a rectangle defining an area to zoom to.  The mouse wheel can also be used for zooming in and out.  The user can also double-click a layer in the list to display the preview map.  Note that the GSE application currently only displays preview maps for layers that support the following spatial reference systems:

- EPSG 4326 (latitude/longitude based on the WGS84 datum)
- EPSG 900913/3857/102113/3785/102100 (Spherical Mercator)

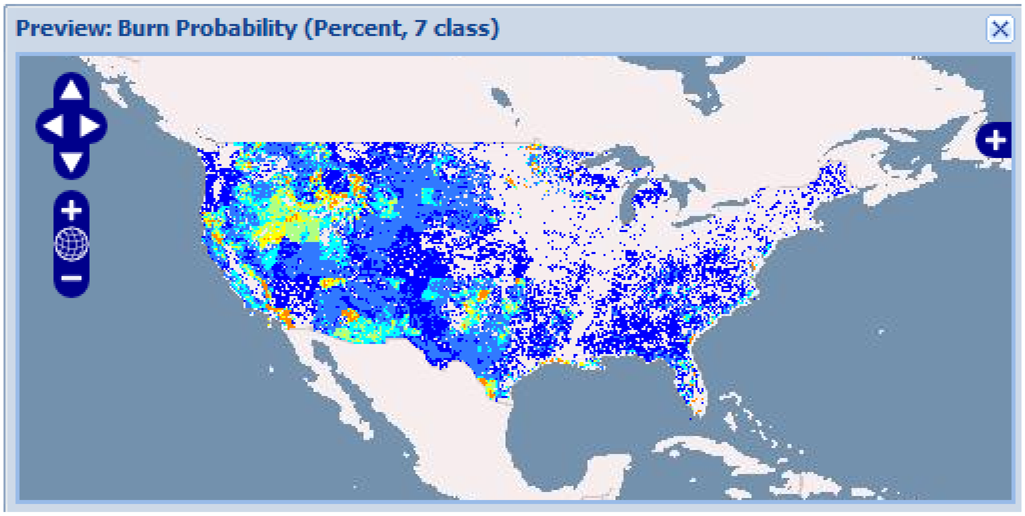
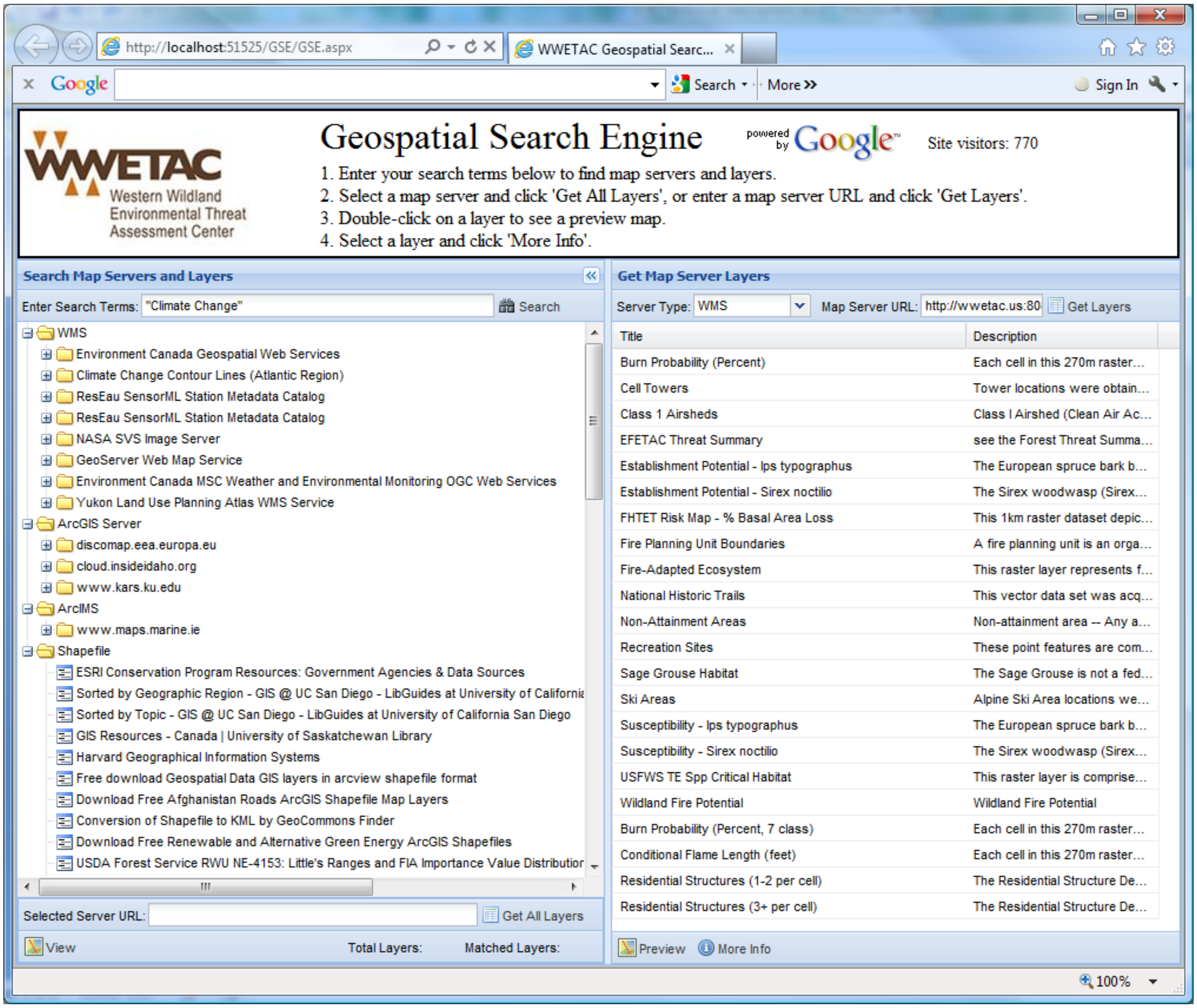


Figure 2. Preview map.

## More Info

The More Info button at the bottom of the layer list panel will open a small window displaying the following additional information for the selected layer:

- Name.  This is the unique name for the layer and is required for outside access to the layer.
- Abstract.  A description of the layer.
- Keywords.  Keywords specified for the layer.
- Lat Lon Bounding Box.  The extent of the layer in latitude/longitude units.
- Supported Spatial Reference Systems.  A comma delimited list of EPSG codes for spatial reference systems supported by the layer.


http://localhost:51525/GSE/GSE.aspx
WWETAC Geospatial Searc...
Google
Search
More
Sign In
WWETAC
Western Wildland Environmental Threat Assessment Center
powered by Google
100%

Geospatial Search Engine

Site visitors: 770

1. Enter your search terms below to find map servers and layers.
2. Select a map server and click 'Get All Layers', or enter a map server URL and click 'Get Layers'.
3. Double-click on a layer to see a preview map.
4. Select a layer and click 'More Info'.

**Search Map Servers and Layers**

Enter Search Terms: "Climate Change" Search

- WMS
  - Environment Canada Geospatial Web Services
  - Climate Change Contour Lines (Atlantic Region)
  - ResEau SensorML Station Metadata Catalog
  - ResEau SensorML Station Metadata Catalog
  - NASA SVS Image Server
  - GeoServer Web Map Service
  - Environment Canada MSC Weather and Environmental Monitoring OGC Web Services
  - Yukon Land Use Planning Atlas WMS Service
- ArcGIS Server
  - discomap.eea.europa.eu
  - cloud.insideidaho.org
  - www.kars.ku.edu
- ArcIMS
  - www.maps.marine.ie
- Shapefile
  - ESRI Conservation Program Resources: Government Agencies & Data Sources
  - Sorted by Geographic Region - GIS @ UC San Diego - LibGuides at University of Californi
  - Sorted by Topic - GIS @ UC San Diego - LibGuides at University of California San Diego
  - GIS Resources - Canada | University of Saskatchewan Library
  - Harvard Geographical Information Systems
  - Free download Geospatial Data GIS layers in arcview shapefile format
  - Download Free Afghanistan Roads ArcGIS Shapefile Map Layers
  - Conversion of Shapefile to KML by GeoCommons Finder
  - Download Free Renewable and Alternative Green Energy ArcGIS Shapefiles
  - USDA Forest Service RWU NE-4153: Little's Ranges and FIA Importance Value Distributior

Selected Server URL: Get All Layers

View Total Layers: Matched Layers:

**Get Map Server Layers**

Server Type: WMS Map Server URL: http://wwetac.us:80 Get Layers

| Title | Description |
|---|---|
| Burn Probability (Percent) | Each cell in this 270m raster... |
| Cell Towers | Tower locations were obtain... |
| Class 1 Airsheds | Class I Airshed (Clean Air Ac... |
| EFETAC Threat Summary | see the Forest Threat Summa... |
| Establishment Potential - Ips typographus | The European spruce bark b... |
| Establishment Potential - Sirex noctilio | The Sirex woodwasp (Sirex... |
| FHTET Risk Map - % Basal Area Loss | This 1km raster dataset depic... |
| Fire Planning Unit Boundaries | A fire planning unit is an orga... |
| Fire-Adapted Ecosystem | This raster layer represents f... |
| National Historic Trails | This vector data set was acq... |
| Non-Attainment Areas | Non-attainment area -- Any a... |
| Recreation Sites | These point features are com... |
| Sage Grouse Habitat | The Sage Grouse is not a fed... |
| Ski Areas | Alpine Ski Area locations we... |
| Susceptibility - Ips typographus | The European spruce bark b... |
| Susceptibility - Sirex noctilio | The Sirex woodwasp (Sirex... |
| USFWS TE Spp Critical Habitat | This raster layer is comprise... |
| Wildland Fire Potential | Wildland Fire Potential |
| Burn Probability (Percent, 7 class) | Each cell in this 270m raster... |
| Conditional Flame Length (feet) | Each cell in this 270m raster... |
| Residential Structures (1-2 per cell) | The Residential Structure De... |
| Residential Structures (3+ per cell) | The Residential Structure De... |

Preview More Info

Figure 3. Gespatial Search Engine application interface.

## GSE Search Engine

The GSE application contains a database of map servers (WMS, ArcIMS, and ArcGIS Server) and websites with shapefiles available for download. Users can search this database by entering terms in the user interface. As in most search engines, common words are eliminated from the search terms unless they are enclosed in quotes. Users will also need to enclose terms in quotes in order to search for a specific phrase. For example, if the terms “Climate Change” are not enclosed in quotes, the search engine will return all items that include both the words “climate” and “change” anywhere in the text. The application will search for these terms in the title, abstract and keywords at both the server level and the map layer level. For shapefiles, the application will search for the terms in the web page title, description, and main text.

Servers and layers that meet the search criteria will be ranked and returned to the client browser in xml format. The servers will be ranked higher if the search criteria were found in the server abstract, keywords, URL, or title. The rank is increased for each layer that contains the search criteria in the layer abstract, keywords, or title. For shapefiles, the websites that contain the search criteria in the URL, title, or page description are ranked higher and placed at the top of the list. The browser application will process the xml data and display it to the user in a tree view window sorted by rank.

## GSE Web Crawler

A web crawler is incorporated into the GSE application to both update server data and find new servers. When a user searches for map servers and layers, the user specified terms are combined along with predefined terms and sent to the Google search engine in order to retrieve web pages related to both the user terms and potential GIS data. The Google search engine results are added to a list of URLs stored on the server. The list of URLs is analyzed daily to find map servers mentioned in the web sites. Analyzed URLs are added to a list and are only analyzed again after a predefined time period to improve overall performance.

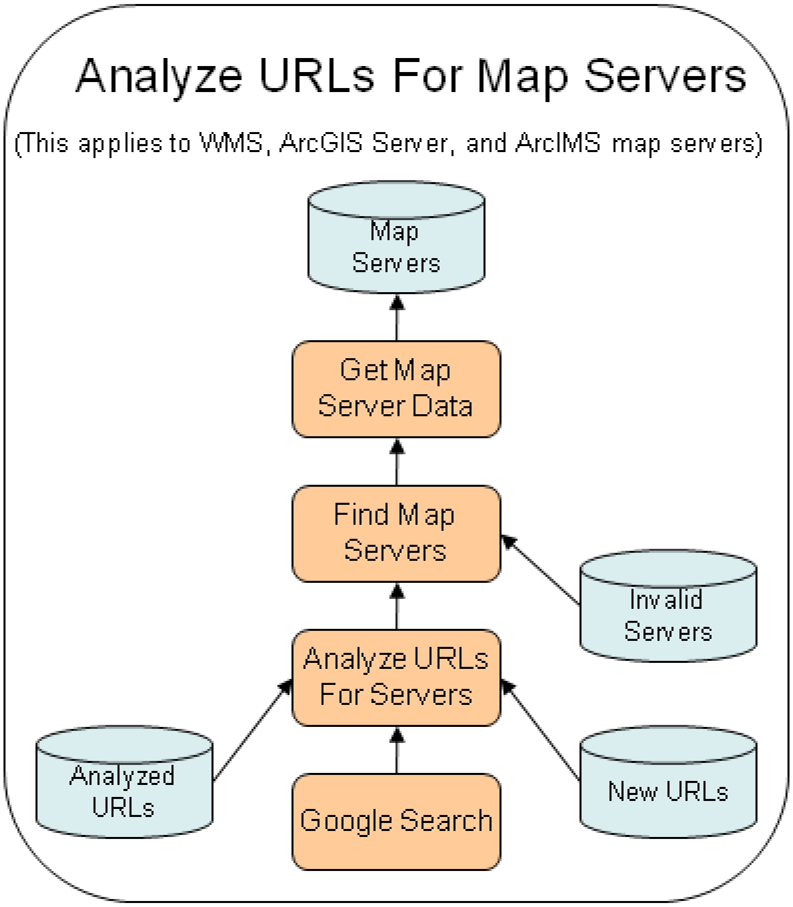


Figure 4. URLs are analyzed for potential references to map servers.

Map servers mentioned in the websites that are not already in the database are contacted to retrieve metadata. If the metadata is successfully returned, this data is added to the database.

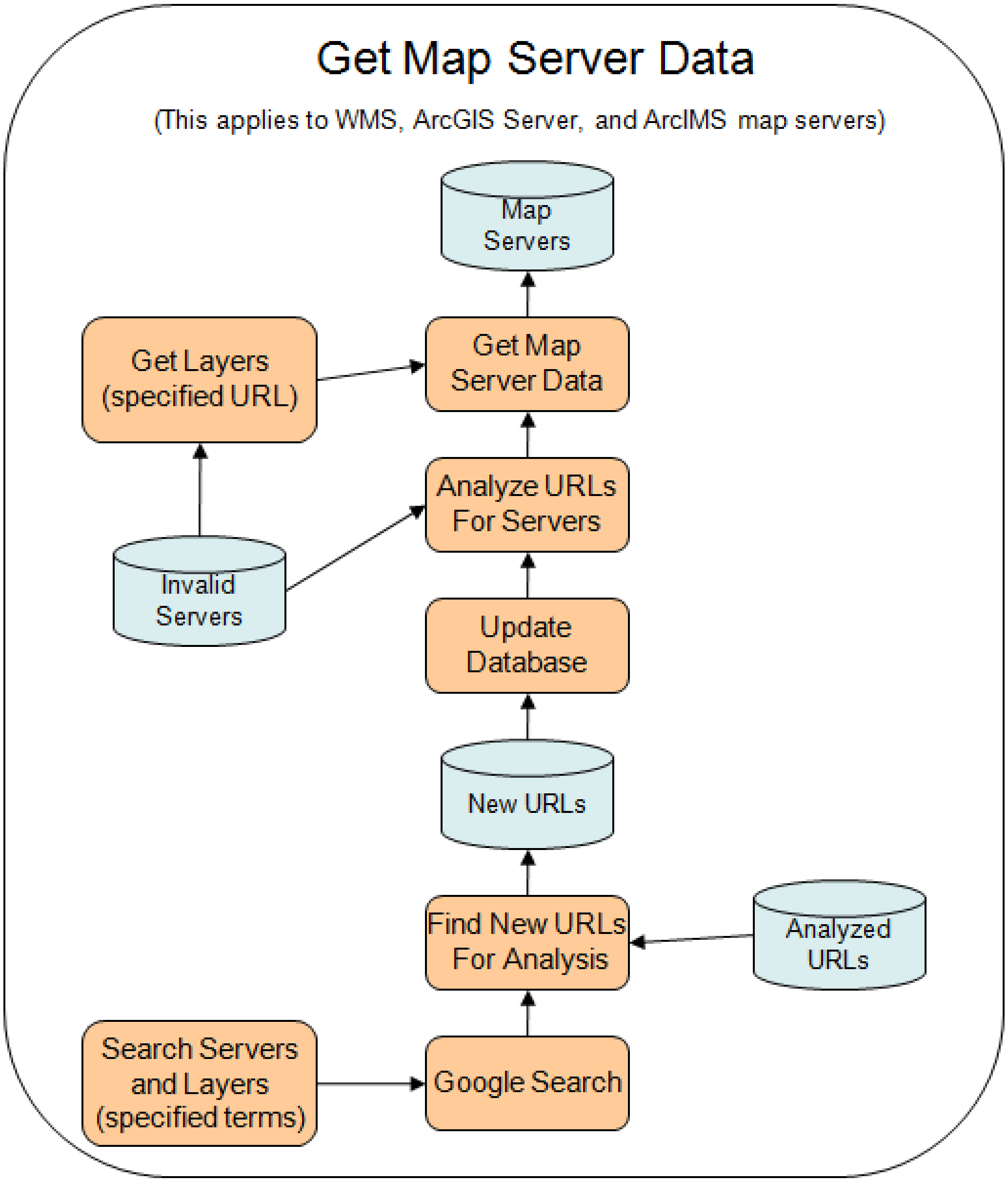

Get Map Server Data
(This applies to WMS, ArcGIS Server, and ArcIMS map servers)
Map Servers
Get Layers (specified URL)
Get Map Server Data
Analyze URLs For Servers
Invalid Servers
Update Database
New URLs
Find New URLs For Analysis
Analyzed URLs
Search Servers and Layers (specified terms)
Google Search

Figure 5. Retrieving map server data.

For shapefiles, Google returns a list of websites that mention the user specified terms, and also mention shapefiles available for download. The list of URLs is analyzed daily to find shapefile websites that are not already in the database. These websites are contacted to retrieve and further analyze the website text. The application has a list of keywords that are likely to be found in websites that have shapefiles available for download. The system will search for these keywords in the website text and rank the website according to how many keywords were found. Keywords found in the web page title, URL, or description, hold a greater weight in the ranking algorithm. If the website is ranked high enough, it is added to the database. Shapefiles are returned to the user interface sorted by rank.

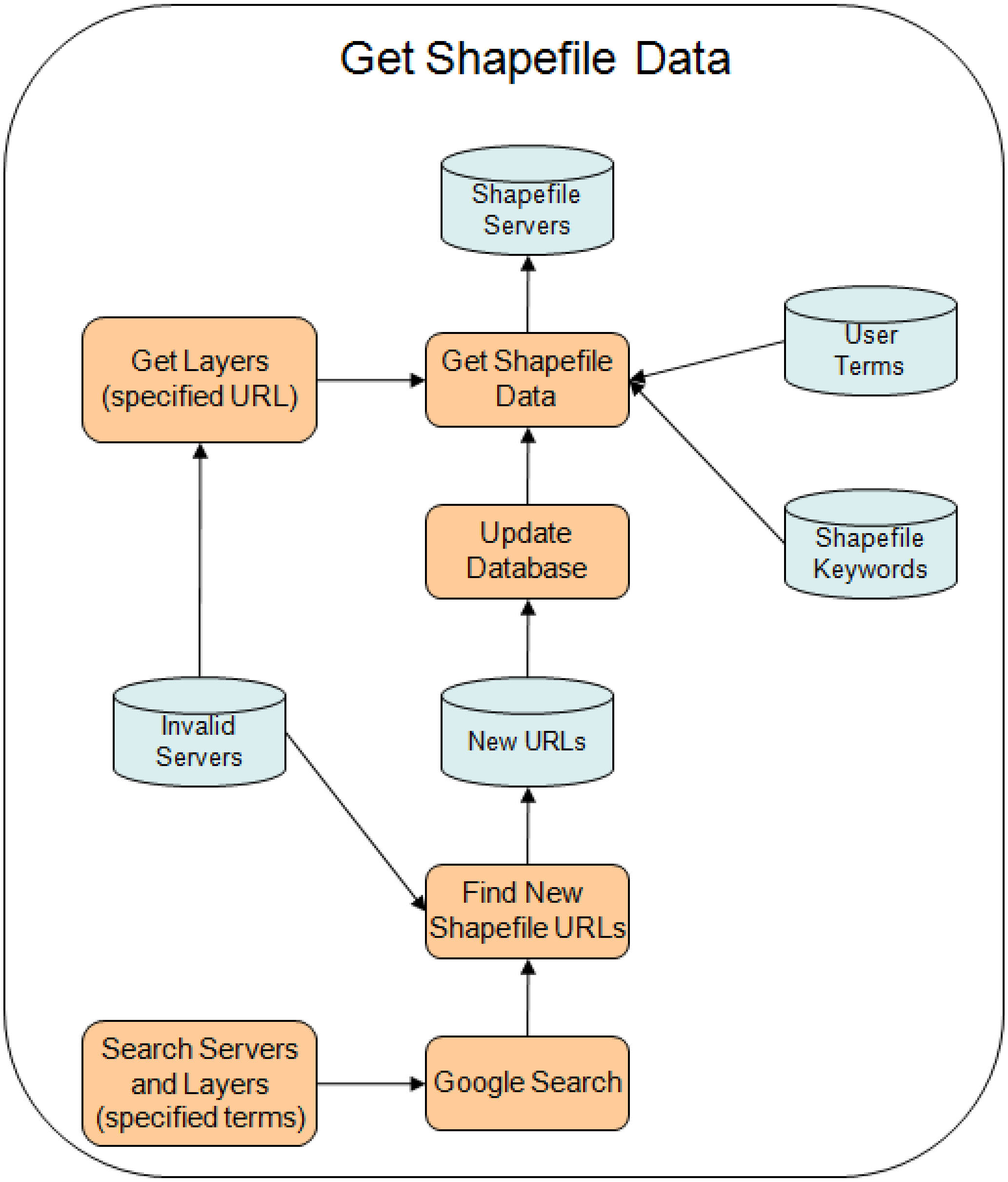


Figure 6. Retrieving shapefile data.

## Database Updates

The database is updated on a daily basis. This process will contact all map servers in the database and reload metadata to ensure the latest changes are included in the system. Servers that don't respond or have an invalid response are added to a list of invalid servers. These servers are ignored in the future to improve performance.  Invalid servers are tested for validity periodically.  If an invalid server is later found to be valid, it is removed from the invalid servers list and added into the main database.  If an invalid server continues to be invalid after a predefined time period, it will be removed from the invalid servers list.  This process will eventually remove old inactive servers from the system.

During database updates, the system will also process the list of new URLs that were collected from Google search results.  After the database is updated, reports in CSV file format will be created for each map server type.  The report calculates the following fields.

- Server Title
- Server URL
- Server Abstract Length
- Server Keywords Length
- Total Layers
- Percent Layers With Abstract
- Mean Layer Abstract Length
- Percent Layers With Keywords
- Mean Layer Keywords Length

The bottom of the report calculates and displays the following summary fields.

- Total Servers
- Total Server Keyword Length
- Total Layers
- Mean Server Keyword Length
- Mean Number of Layers
- Mean layer abstract length for all servers
- Mean layer keyword length for all servers
- Percent layers with abstract for all servers
- Percent layers with keywords for all servers
- Percent layers with EPSG 4326

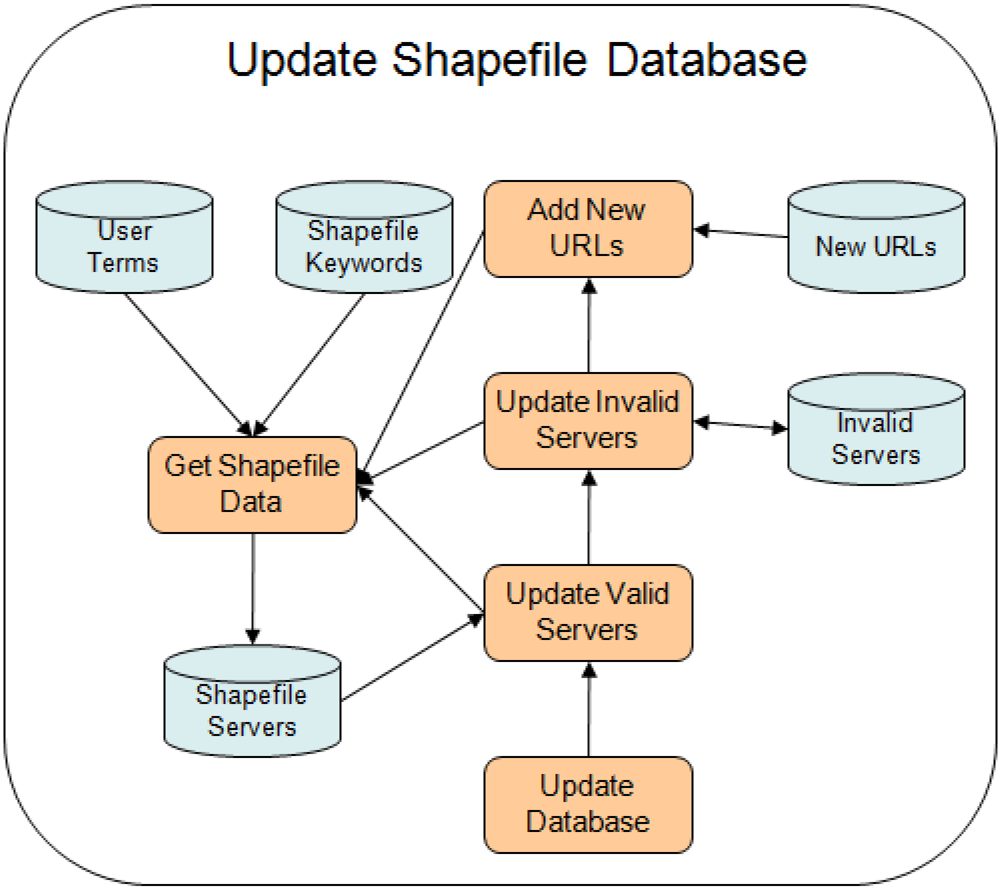


Figure 7. Design for updating the shapefile database.

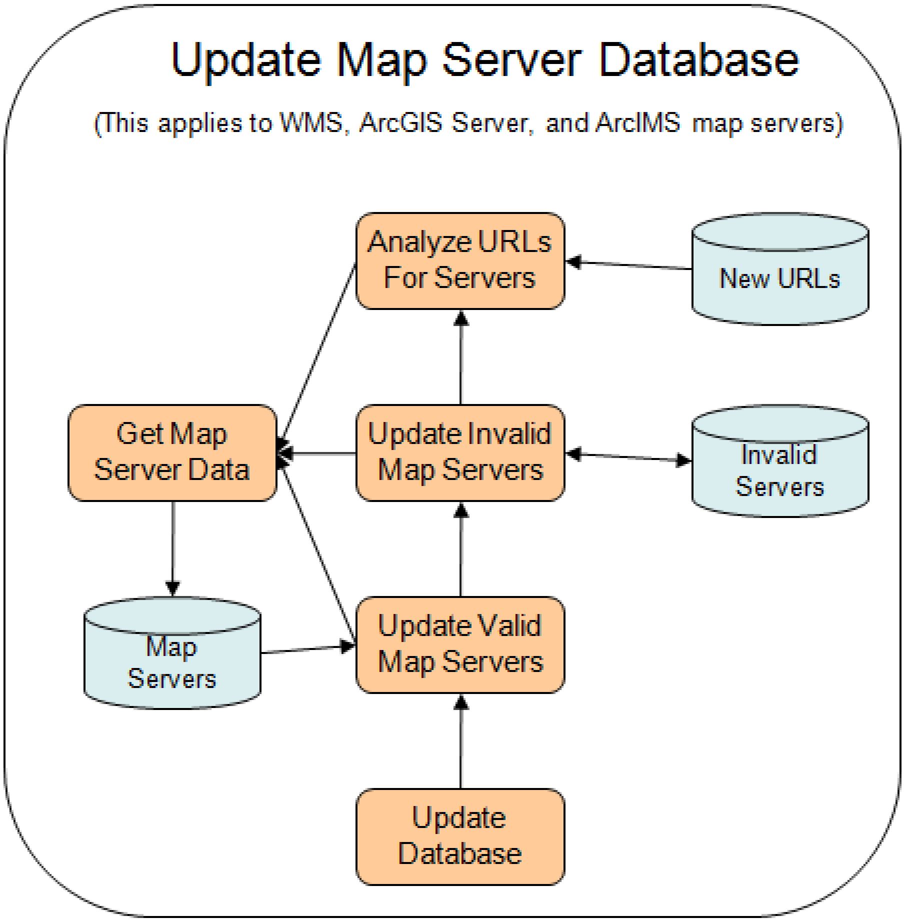


Figure 8. Design for updating the map server database.

## GSE Database

The GSE database is stored in xml format and is comprised of the following data:

- ArcIMS
  - Server
    - Type
    - URL
    - Title
    - NumLayers
    - DateUpdated
    - Layer
      - Name
      - Title

            - Abstract
            - Keywords
            - LatLonBBox
            - SRS
- InvalidServers
    - Server
        - Type
        - URL
        - DateAdded
        - DateUpdated
- WMS
    - Server
        - Type
        - URL
        - Title
        - Abstract
        - Keywords
        - NumLayers
        - DateUpdated
        - Layer
            - Name
            - Title
            - Abstract
            - Keywords
            - LatLonBBox
            - SRS
- NewURLs
    - URL
        - Type
        - Name
- Terms
    - Term
        - Name
        - Count
        - LastDate
        - Included
- AnalyzedURLs
    - URL
        - Type
        - Name
        - lastDateAnalyzed

- Shapefile
    - Server
        - Type
        - URL
        - Title
        - Abstract
        - Keywords
        - Rank
        - SiteFilename
- Keywords
    - Keyword
        - Name
        - Points
- Stopwords

A simple text file list of search engine stop words