

# Internet Map Services: New portal for global ecological monitoring, or geodata junkyard?

Alan Ager  
Operations Research Analyst  
USDA Forest Service  
Western Wildland Environmental  
Threat Center  
541.416.6603  
[aager@fs.fed.us](mailto:aager@fs.fed.us)

Charlie Schrader-Patton  
GIS/Remote Sensing Analyst  
USDA Forest Service  
Remote Sensing Applications  
Center  
541.312.4291  
[cschrader@fs.fed.us](mailto:cschrader@fs.fed.us)

Ken Bunzel  
GIS Applications Developer  
Kingbird Software, LLC  
208.310.0892  
[kbunzel@kingbirdsoftware.com](mailto:kbunzel@kingbirdsoftware.com)

Brett Colombe  
MicroImages  
402.477.9562  
[support@microimages.com](mailto:support@microimages.com)

## ABSTRACT

Systematic data mining of geospatial and other data available on the internet may provide a novel means for early detection and assessment of ecosystem change and impending natural disturbances. Exploring the possibilities and limitations of systematic geodata mining of the internet has just begun. Webcrawlers to locate, assess, and connect to these data are beginning to appear within experimental domains. In this project, we built a geodata webcrawler and post processor and then integrated it within a virtual earth viewer and a web interface to assess current data availability for several key topic areas for wildland ecological assessments. The work is part of a larger project at the Western Wildland Environmental Threat Assessment Center to build an early warning and monitoring system for specific wildland threats to human and ecological values.

## Keywords

USDA Forest Service, geodata webcrawler, ArcGIS Server, web mapping service, WMS, WFS, OGC

## 1. INTRODUCTION

Systematic data mining of geospatial and other information content on the internet may provide a novel means for early detection and assessment of ecosystem change and impending natural disturbances [Galez et al. 2009]. Early warnings and changes in ecological drivers may well be detectable by sophisticated web crawlers that work within cyber infrastructures [Crowl et al. 2008].

The exploration of possibilities and limitations of systematic data mining of the internet has just begun [Galez et al. 2009]. Web crawlers can be designed to collect information on rapid changes in key indicators that either directly or via cascading threats can lead to ecological declines. Operational web crawlers in the public domain are rare, but experimentation for detection of ecological and human threats is growing. For instance, a web crawler developed by the USDA Animal and Plant Health Inspection Service searches for internet sales of prohibited organisms as a means to address the threat of invasive alien species [Meyerson and Reaser, 2003]. The Global Public Health Intelligence Network [GPHIN], an early disease detection system developed by Health Canada for the World

Health Organization, gathers information about unusual disease events by monitoring internet-based global media sources

At the USDA Forest Service's Western Wildland Environmental Threat Assessment Center (WWETAC), we are exploring webcrawlers to facilitate wildland threat assessments. The Threat Center was established by Congress in 2005 to facilitate the development of tools and methods for the assessment of multiple interacting threats (wildfire, insects, disease, invasive species, climate change, land use change). Geospatial data are key to the detection, assessment, and monitoring of the wide array wildland threats common in the western USA. Threat Center products concern mapping an array of threats, and national risk maps are now being produced [Calkin et al. 2010]. While it is estimated that over 1 million spatial data sets on 30,000 internet map servers are now posted by government agencies, universities, and private organizations, identifying geospatial services pertaining to specific threats is problematic. The volume of data grows daily as new, low-cost server technologies evolve to provide widespread capability for publishing spatial data on the internet [Schrader-Patton et al. this proceedings; Bunzel et al. this proceedings]. There exists a potential to mine these data for a broad range of geospatial assessments to study natural resource management problems including the spatial distribution and co-occurrence of various wildland threats and the human and ecological values they affect. In this paper, we describe a spatial data webcrawler and post processor and its application to assess the availability of pertinent spatial data.

## 2. METHODS

### 2.1 Map Services

Map Services are published to the web using specialized server software such as MapServer, GeoServer, and ArcGIS® Server (ESRI®, Redlands CA). A client application makes a request to the server, which then returns a representation of the geographic data on the server to the client. Request statements typically include bounding coordinates of the area requested, layers of the service to deliver, and format of the delivered data.

Services are typically accessed using a URL link that is structured according to a protocol. The Open Geospatial Consortium [OGC] has developed some non-proprietary open standards for map services; developers of client applications can use these standards to build clients that can consume the services. [OGC, 2010]. Other protocols are vendor-specific and can only be ingested by client applications built by the vendor (e.g. ESRI® products). Both types of services have a specific URL format that can be identified using a webcrawler.

### 2.2 Web Crawler

Searching the servers and mining the metadata are accomplished in a two-step process. To search for servers, we used a webcrawler written by MicroImages®, Inc. For more information about Microimages see <http://www.microimages.com/>. The webcrawler uses a series of filtered Google searches on the *GetCapabilities* string to identify servers with WMS and WFS services. After each search, the server URL is stored in a database and the content of the XML file is parsed into a catalog containing 42 metadata attributes at the server level. The attributes include the layer names, coordinates, and other information. The database is searchable based on these attributes with URL search strings.

### 2.3 Mining Layer Level XML

To probe the catalog and acquire information on specific data layers and respective metadata, we built a web interface that queried the catalog based on user keywords and then located the server, acquired the metadata, and then mined the metadata for information of interest. The web interface also lists all the layers present on the server and reports the number of layers (Figure 1). This secondary utility essentially cycles through the servers returned from a search in the catalog. For each unique server, the utility requests the capabilities document from the server. This document stores all metadata for the map server in xml format. The xml document uses a standard document format described by the Open Geospatial Consortium. The xml document is read and parsed to obtain detail information on the map server such as the server abstract and layer-level metadata. The parsed data is summarized and stored in a table for each keyword search. Outputs from separate searches were combined in a spreadsheet for further summary statistics.

We used this system described above to perform keyword queries and examine the number of unique servers, matched layers, the presence of an abstract that described the layer, and the mean abstract length. We chose a number of keywords, but report here on keyword searches for: wildfire, land use, climate change, climate, forest, bark beetles, *Dendroctonus*, and invasive plants.

### 2.3 Integration

The data mining system was integrated into an internet map viewing system hosted on an Amazon EC2 cloud (Bunzel et al., this proceedings) along with the postprocessor scripts described above. Once a server and layers of interest are identified from search parameters, data can then be previewed in a window within the browser window and added as an overlay or base map in the mapping system. Base maps stored on the system include an array of spatial data on wildland threats and the values they affect that have been generated by the Forest Service and other federal agencies (see left panel Figure 1).

## 3. Results

Our search for OGC-WMS spatial data for specific wildland threats revealed that extensive data are available for some features of interest, such as wildfire (10 servers, 128 layers), but few existed for more specific threats like bark beetles (1 layer, Table 1). However, even in the case of features where a large number of layers were found [e.g. wildfire], abstracts were only available for four layers, or about 3 percent of the total layers. Of the abstracts found for wildfire, the average word count was 995, which is more than sufficient to evaluate the usefulness of the data for a specific assessment. A large number of layers were found by searching on the climate keyword (28 servers, 79 layers), but only five abstracts were found, with an average word count of 982. Other keywords describing ecological threats or their manifestation, such as invasive plants and forest, turned up little or no data.

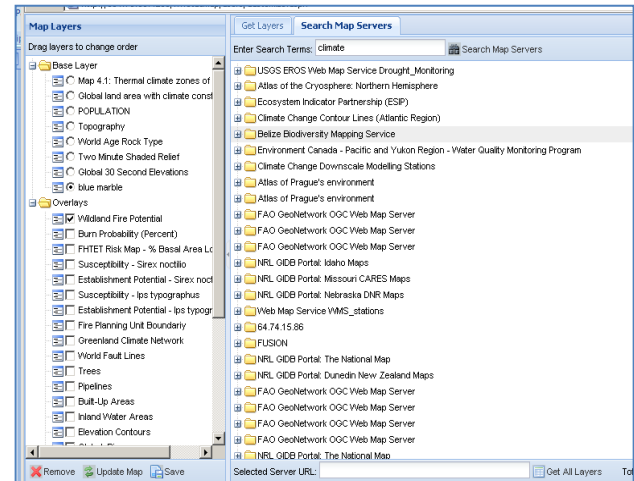


Figure 1. The query interface and a partial listing of server folders returned from a search using the term “climate”.

Applying the system for manual searches of WMS data revealed several insights about the process of geospatial data mining as the data currently exist. First, the map previewing system was equally useful for examining layers that turned up in data searches. Specifically, the primary searches on layer name keywords identified servers that contained similar data of general potential interest that were subsequently interrogated for a complete listing of layers. Browsing these layers in the map previewer allowed additional information gathering about these

specific spatial data. Server level queries were informative largely because a relatively few number of servers have a large number of layers.

**Table 1. Keyword WMS service search results.**

<b>Keywords</b>	<i>Servers</i>	<i>Layers</i>	<i>Abstracts</i>	<i>Word Count</i>
Wildfire	10	128	4	995
Land use	45	156	1	4
Climate change	3	6	1	1338
Bark Beetles	1	0	0	0
<i>Dendroctonus</i>	1	0	0	0
Invasive plants	1	0	0	0
Forest decline	1	0	0	0
Climate	28	79	5	982
Forest Threat	1	1	0	0
Forest	26	262	3	26

#### 4. Future Work and Conclusions

Our initial work to evaluate the information content of internet map services focused on the development of a processing system to filter the results of keyword searches into meaningful metrics. To that end, we created an efficient data mining system to report on the quantity and quality of global spatial data provided by map services. The assessment here was limited to WMS servers, and future work will include WFS, ArcIMS, and ArcGIS servers as well. However, much work is needed to complete the assessment of the data for applications in ecological threat assessments. First, by querying based on the layer name we are missing relevant layers that do not contain the keyword in the name. Second, our assessment also revealed that much of the online data is stored in large institutional data warehouses (Natureserve, Geodata.gov, etc.) that have their own catalog and searching systems and are not open to webcrawlers like ours. In fact, most federal land management agencies deliver their data as map services, but allow downloading and in-house viewers (i.e. FHTET 2006). This policy does not simplify the problem of integrated threat assessments for federal land management agencies. While the government data portals might be useful in long-term monitoring work, their use for early warning systems may be limited due to the lag time for ingesting new data. Both of these situations could help explain our finding of zero layers returned for such common keywords as ‘Forest decline’, and ‘Invasive plants’.

While the results of our query point to large amounts of information in abstracts, that could help decipher the geodata content, abstracts were available for only a small number layers. Without abstracts, data are of limited use for specific ecological or other environmental studies.

In any event, the results show that information like abstracts is rarely found on internet map services, and this level of detail is critical for scientific assessments. It is possible that the scattered web mapping services that exist outside of the large data warehouses, are slowly creating “information junkyards” [McDermott 1999] as discussed by Galez et al. [2009]. Leveraging these fragmented data as part of a webcrawler will require ancillary knowledge systems to scan for key information pertinent to specific ecological problems

We will continue to refine our webcrawler and post-processing system and integrate it into a larger package of geospatial products being developed at WWETAC [Schrader-Patton et al. this proceedings, Bunzel et al. this proceedings] for integrated wildland threat assessment and mapping. We are integrating search capability for ArcIMS and ArcGIS servers. In addition, we are working on interactive metadata retrieval for individual layers, and on search engines for printed media, kml files, and other information sources.

#### 5. REFERENCES

- [1] Amato-Gauci, A. and Ammon, A. 2008. The surveillance of communicable diseases in the European Union – a long-term strategy. *Eurosurveillance* 13: 26.
- [2] Bunzel, K., Ager, A.A., Schrader-Patton, C.. (this proceedings). Open Source Map Server Deployment on a Cloud Server: A Case Study.
- [3] Calkin, D.E.; Ager, A.A.; Gilbertson-Day, J., eds. 2010. Wildfire risk and hazard: Procedures for the first approximation. Gen Tech. Rep. RMRS-GTR-235. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 62p.
- [4] Crowl, T.A., Crist, T.O., Parmenter, R.R., *et al.* 2008. The spread of invasive species and infectious disease as drivers of ecosystem change. *Front Ecol Environ* 6: 238–46.
- [5] FHTET – US Forest Service Forest Health Technology Enterprise Team National Insect and Disease Risk Map. 2006. <http://www.fs.fed.us/foresthealth/technology/nidrm.shtml>. Accessed 18 February 2010
- [6] Galaz, V., Crona, B., Daw, T., Bodin, O., Nyström, M., Olsson, P. Can web crawlers revolutionize ecological monitoring *Front Ecol Environ* 2010; 8(2) 99–104, doi:10.1890/070204 (published online 19 Mar 2009)
- [7] Geodata.gov. <http://gos2.geodata.gov/wps/portal/gos> Accessed 3/18/2010.
- [8] McDermott, R. 1999. Why information technology inspired but cannot deliver knowledge management. In: Lesser, E.L., Fontaine, M.A., and Slusher, J.A. (Eds). *Knowledge and communities*. Boston, MA: Butterworth-Heinemann. OGC. 2010.
- [9] Meyerson, L.A. and Reaser, J.K.. 2003. Bioinvasions, bioterrorism, and biosecurity. *Front Ecol Environ* 1: 307–14.
- [10] Open Geospatial Consortium – Open GIS Standards. <http://www.opengeospatial.org/standards>. Accessed 18 February 2010.
- [11] Schrader-Patton, C., Ager, A.A., Bunzel K. GeoBrowser Deployment in the USDA Forest Service: A Case Study. (This proceedings).

[12] WWETAC. 2003. US Forest Service Western Wildland  
Environmental Threat Center Charter Statement.

[http://www.fs.fed.us/wwetac/wwetac\\_charter.html](http://www.fs.fed.us/wwetac/wwetac_charter.html).  
Accessed 18 February 2010.