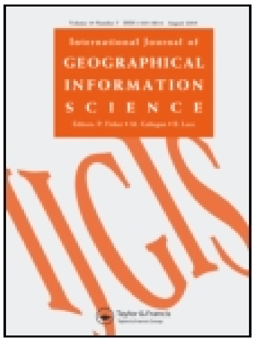


International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

A performance, semantic and service quality-enhanced distributed search engine for improving geospatial resource discovery

Zhipeng Gui , Chaowei Yang , Jizhe Xia , Kai Liu , Chen Xu , Jing Li & Peter Lostritto

To cite this article: Zhipeng Gui , Chaowei Yang , Jizhe Xia , Kai Liu , Chen Xu , Jing Li & Peter Lostritto (2013) A performance, semantic and service quality-enhanced distributed search engine for improving geospatial resource discovery, International Journal of Geographical Information Science, 27:6, 1109-1132, DOI: 10.1080/13658816.2012.739692

To link to this article: https://doi.org/10.1080/13658816.2012.739692

International Journal of Geographical Information Science, 2013
Vol. 27, No. 6, 1109–1132, http://dx.doi.org/10.1080/13658816.2012.739692

Taylor & Francis
Taylor & Francis Group

A performance, semantic and service quality-enhanced distributed search engine for improving geospatial resource discovery

Zhipeng Gui, Chaowei Yang*, Jizhe Xia, Kai Liu, Chen Xu, Jing Li and Peter Lostritto

Center for Intelligent Spatial Computing and Department of Geography, GeoInformation Sciences, George Mason University, Fairfax, VA 22030–4444, USA

(*Received 16 January 2012; final version received 30 September 2012*)

Geospatial resource discovery is a critical step for developing geographic science applications. With the increasing number of geospatial resources available online, many Spatial Data Infrastructure (SDI) components (e.g. catalogues and portals) have been developed to help manage and discover geospatial resources. However, efficient and accurate geospatial resource discovery is still a big challenge because of the heterogeneity and complexity of decentralized network environments and interdisciplinary semantics. In this article, we report a search engine framework for efficient geospatial resource discovery, which reduces integration costs by leveraging existing Geospatial Cyberinfrastructure (GCI) components. Specifically, (1) the framework provides integration capability and flexibility by adopting the brokering approach, implementing a 'plug-in'-based framework for metadata processing and proposing a dynamically configurable search workflow; (2) the asynchronous messaging and batch processing-based metadata record retrieval mode enhances the search performance and user interactivity; (3) an embedded semantic support system improves the discovery recall level and precision by providing semantic-based search rule creation and result similarity evaluation functions and (4) the engine assists user decision-making by integrating a service quality monitoring and evaluation system, data/service visualization tools, multiple views and additional information. Experiments and a search example show that the proposed engine helps both scientists and general users search for more accurate results with enhanced performance and user experience through a user-friendly interface.

Keywords: CyberGIS; spatial cloud computing; spatial web portal; data discovery; EarthCube

1. Introduction

The National Science Function (NSF) proposed EarthCube (NSF 2011) to create a knowledge management infrastructure to integrate all geoscience data in an open, transparent and inclusive manner. The capability of converging information resources from multiple domains, building bridges among scientific communities and supporting multidisciplinary scientific search, access and utilization is one of the keys to the success of the initiatives like the EarthCube and Digital Earth. With rapid development of data acquisition, processing and application technologies, geospatial data are collected at an unprecedented rate. Petabytes of geodata and thousands of geospatial web services, tools and

*Corresponding author. Email: chaowei.yang.1@gmail.com; cyang3@gmu.edu

applications are available over the Internet. To advertise, discover and share the resources, owners and contributors usually publish resource metadata via catalogues (e.g. Global Earth Observation System of Systems Clearinghouse/GEOSS CLH, Geospatial One-Stop/GOS, NASA Global Change Master Directory/GCMD and Earth Observing System Clearinghouse/ECHO, FedEO Collection Catalogue and Unidata Thematic Realtime Environmental Distributed Data Services/THREDDS catalogue) or portals (e.g. Unidata, GEON, NEON, OpenTopography and US Geoscience Information Network (GIN), where users search for needed resources). However, finding appropriate geospatial resources efficiently and accurately is still challenging in several aspects:

(1) Cross-domain barriers. Existing Spatial Data Infrastructure (SDI, Mapping Sciences Committee 1993) are usually established to enable organizations and users in specific geographic science domains (e.g. ocean, atmospheric, ecosystem and demographics) to discover resources with defined types (e.g. dataset, web services, applications and documents). To find appropriate resources, users typically must visit/search multiple portals serving different communities.
(2) Specialized technological knowledge is required. Different SDIs adopt diverse protocols/APIs (e.g. OGC Catalogue Service for the Web/CSW, Z39.50, OpenSearch, Really Simple Syndication/RSS, Atom, Web Accessible Folders/WAF, Open Archives Initiative Protocol for Metadata Harvesting/OAI-PMH) and workflows to access resources. Given the various formats, specifications (e.g. ISO 19139, FGDC's Content Standard for Digital Geospatial Metadata/CSDGM, NASA GCMD Directory Interchange Format/DIF and Service Entry Resource Format/SERF, Organization for the Advancement of Structured Information Standards/OASIS's Electronic business Registry Information Model/ebRIM), and web User Interfaces (UI), the representations of metadata also vary significantly. These issues dramatically increase the difficulty for users to query and integrate heterogeneous metadata from dispersed sources.
(3) Implementing distributed search and federation at the SDI component level is costly. Many registries adopt a periodic harvesting mechanism to retrieve metadata from other federated catalogues. These time-consuming processes lead to network and storage burdens, data redundancy and also the overhead of maintaining data consistency.
(4) Semantics is a big challenge in current SDIs. Keyword-based search technologies are widely adopted in operational SDIs. However, since keyword search uses string matching rather than semantic matching, precision and recall, the two important measurements for search relevance are hard to be guaranteed partly due to the following reasons. The terminologies/vocabularies representing the same concept may vary with domains and languages. The search content cannot be expressed using several keywords explicitly (such as the sub-concepts of a concept). The search keywords may appear in the metadata description but do not reflect the content of the metadata accurately. By contrast, semantic search offers a solution to solve these issues. However, integrating semantics into existing SDIs is challenging due to the limited expandability of metadata templates, frameworks and other issues.
(5) Current representation methods for search results lack information and functionalities for assisting users' selection. Plain-text-based content presentation is not an effective way to help users distinguish records to make a final selection. Supplemental value-added information (e.g. relevance index, service performance,

resource popularity and user feedback), enhanced results visualization through an improved user-friendly interface would enlighten the selection process.

The first two issues increase entry barriers/learning curves (Craglia *et al*. 2011, Mazzetti and Nativi 2011) for both end users (resources consumers) and providers. The third issue causes complexity and high maintenance cost for SDIs. The last two issues are critical for effective discovery and decision-making. To address these issues, we prototyped a distributed search engine – GeoSearch. GeoSearch leverages and utilizes a series of existing standards, technologies and Geospatial Cyberinfrastructure (GCI, Zhang and Tsou 2009, Yang *et al*. 2010) components (e.g. catalogues, portals, visualization tools, semantic supporting systems and quality monitoring and evaluation systems) to narrow the gap between users and geospatial resource providers/publishers and also hide the complexity of GCIs. A set of methodologies are utilized to effectively help common users and scientists discover, select and access geospatial resources in an easy fashion. In general, (1) a lightweight web engine framework to search, integrate and visualize heterogeneous geospatial resources is developed by adopting brokering middleware; (2) batch processing, asynchronous messaging and multi-threaded-based search mode (Yang *et al*. 2005) are adopted to improve search performance and interactivity; (3) a semantic-based query statement refinement is proposed to improve recall level and precision and (4) sophisticated functionalities (e.g. multiple data/information visualization tools, service quality monitoring and semantic similarity calculation) are used to improve user experience and assist decision-making.

This article is organized as follows. Section 2 reviews relevant research. Section 3 introduces the architecture of the proposed search engine. Section 4 details the components, workflow and methodologies. Section 5 describes the implementation and related experiments and demonstrations. Section 6 concludes with results. Lastly, Section 7 discusses future research.

2. Related works

2.1. Existing SDIs for discovery and metadata management

After decades of development, many domain and local/global centralized catalogues and portals (Zhang and Tsou 2009) have been deployed. For example, the intergovernmental Group on Earth Observation deployed an integrative common infrastructure of core services to facilitate discovery, access and utilization of Earth observation data, information, tools and services. CLH, Component and Service Registry (CSR) and CEOS WGISS Integrated Catalogue (CWIC) are developed as operational systems. The US National SDI portal and catalogue, known as Geospatial One-Stop (GOS) portal and catalogue, have been integrated into the official government website to provide the ability to find, download and use datasets generated and held by the federal government. OpenTopography (Krishnan *et al*. 2011) is an NSF-funded programme to provide a portal to facilitate community access to high-resolution, geospatial science-oriented topography data and related tools and resources. Unidata (Fulker *et al*. 1997) serves a diverse community of over 160 institutions with data and tools to access and visualize data, and to enhance Earth-system education and research for over 20 years. Through these SDI developments, geospatial communities have accumulated solid technologies and experience on metadata management and discovery and have maintained abundant metadata resources for different domains. A mechanism to federate these disparate SDIs collaboratively has now become critical. To address this, we designed a loosely coupled brokering framework to connect heterogeneous frameworks, protocols and applications.

2.2. Search protocols, formats and models

Diverse specifications and protocols have been proposed and utilized to facilitate metadata description and acquisition. OASIS (Organization for the Advancement of Structure Information Standards) specifications UDDI (Universal Description, Discovery and Integration) (Clement *et al*. 2004) and ebRIM/ebRS (Najmi *et al*. 2011a, b) are promoted as e-business industrial standards. OpenSearch (Clinton 2011) is a broadly used open specification, which defines a collection of simple conventions for sharing search results. Following that, Walsh *et al*. (2009) tested how to support geospatial discovery. Meanwhile, OGC Web Services (OWS) also defines a series of web-service specifications. OGC Catalogue Service for the Web (CSW) (Nebert 2007) defines the ability to publish, discover and retrieve spatial data and service metadata, whereas other specifications (such as Web Map Service/WMS, Web Coverage Service/WCS, Web Feature Service/WFS, Web Process Service/WPS) also define significant information for geospatial resources. THREDDS (Domenico *et al*. 2002) based catalogues are widely adopted among communities for various applications and research, such as ocean observations[1] and climate change.[2] WAF is a Hypertext Transfer Protocol (HTTP) based accessible directory of files for organizing and accessing metadata content, while OAI-PMH (Van *et al*. 2004) is a protocol used to collect metadata descriptions from archives. NASA also funds Datacasting[3] and Service Casting (scast).[4] The former utilizes RSS-based technology to access geospatial science information, while the latter uses atom syndication feeds to advertise services. OpenSearch, DataCasting and ServiceCasting adopt inconsistent formats, the ESIP Discovery Cluster[5] proposes an Atom response (named ESIP Discovery Cast Atom Response Format) to unify these discovery formats. However, the protocols and specifications have their unique applicability in different domains. It is unrealistic to establish a single unified protocol/specification. For example, assuming one resource publisher changes its protocol and format, all connected systems and clients should change together and such changes will result in a high maintenance cost. So, search engines and middleware frameworks should be able to tackle these heterogeneous problems. In this research, we use a flexible plug-and-play metadata processing framework to address the problem.

2.3. Discovery technologies and strategies

Much research has explored semantic assisted geospatial information discovery and utilization. Various semantic metadata description methods (Klien *et al*. 2006, Lutz and Klien 2006, Lutz 2007, Alam *et al*. 2011) have been introduced to improve discovery and composition capabilities based on different semantic specifications, such as OWL-S (OWL-based Web Service Ontology), SAWSDL (Semantic Annotations for WSDL), and WSMO (Web Service Modeling Ontology). Di *et al.* (2006) and Yue *et al.* (2007) proposed ontology-based semi-automatic geospatial service chaining methods. But many existing SDIs lack metadata semantic annotations and extension mechanisms for semantic search. Applying these methods into those SDIs is difficult. Therefore, a semantic engine is proposed as a complementary component to support semantic search. Noesis (Ramachandran *et al.* 2006) is a semantic search engine, which uses atmospheric science domain ontologies to improve search accuracy. Li and Yang (2008) proposed a servlet-based semantic search engine frameworks to assist distributed search. However, these frameworks are not flexible enough to tackle heterogeneous system integration and performance problems.

For discovering geospatial web services, Sample *et al.*(2006) and Li *et al*. (2010) investigated active crawler-based search mechanisms for discovering WMS, since traditional centralized registry-based mode cannot guarantee that the registered metadata are up to date. Zhang *et al.* (2010) developed a semantic specification and geospatial semantic web-enabled search engine to discover geospatial features from WFS. Li *et al.* (2011) proposed a performance optimized framework to search OGC web services through CSW catalogues by using AJAX, but the research only focused on the discovery of certain OGC services.

While the research discussed above has contributed to geospatial resources discovery in certain aspects, they do not provide a solution to improve the discovery through integrating existing GCIs. To enhance the heterogeneous integration capabilities, the EuroGEOSS project[6] investigated a brokering approach to facilitate geospatial resources discovery, access and integration. Four broker middleware are proposed, including a discovery broker, an access broker, a semantic broker and a web 2.0 broker. The discovery broker clusters heterogeneous results from different data providers and caches them in local catalogue mediation (GI-cat) (Nativi and Bigagli 2009). The access broker downloads and retrieves datasets on demand. The semantic broker works as a discovery augmentation component to federate both semantics and ISO-compliant catalogues (Santoro *et al*. 2011). The web 2.0 broker retrieves web 2.0 resources (i.e. RSS, ATOM, OpenSearch, etc.). EuroGEOSS also plans to integrate the GeoViQua project for enabling quality-aware visualization and advanced geo-search capabilities. Using these four brokers, EuroGEOSS has developed a powerful search platform. The local cached catalogue mediation can improve search efficiency and lessen real-time network communication, but a heavy local catalogue has to be established (although GI-cat supports distributed search, most remote catalogues are locally cached, except GEOSS CLH and GCMD currently) and therefore maintenance cost increases accordingly. Furthermore, EuroGEOSS components may work tightly to some extent (e.g. GeoViQua needs to add quality tag Geo-Label to the metadata in catalogues mediation, which may impact the independency of the components) and related studies do not show how to loosely integrate the third-party functionalities into its framework and discovery workflows.

The main objective of this research is to explore a scalable solution to facilitate the discovery of distributed heterogeneous geospatial resources from public catalogues and portals through a lightweight brokering middleware framework. Comparing with EuroGEOSS, which maintains local catalogues at the broker side, we explore how to search heterogeneous sources efficiently on the fly and provide a value-added discovery process to assist user decision-making. Especially, our engine enhances the interactivity and visualization functionalities by flexibly integrating third-party controls and GCIs through a uniform web client.

3. Architecture

Our search engine, GeoSearch, which searches against the dispersed heterogeneous SDIs (catalogues and portals), consists of four loosely coupled components: a search broker, visualization tools, resource monitoring and evaluation systems and semantic support systems (Figure 1). The search broker processes and responds to users' search requests, and also plots search results, acting as the brokering middleware between GCIs and end users. The semantic support system translates search input into query rules and calculates relevance indexes between the search subject and the retrieved records. The visualization tools

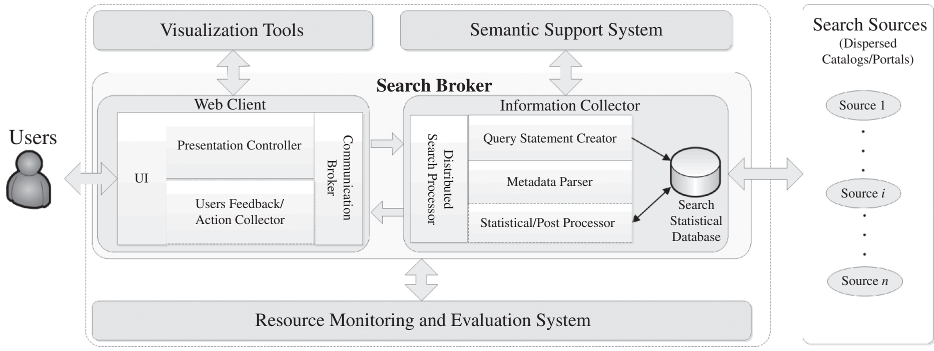


Figure 1. Architecture of GeoSearch.

integrated with web browsers are used to visualize data and data services. The resource monitoring and evaluation system monitors service performance and collects subjective evaluation from users and calculates comprehensive (overall) scores.

The search broker, the core component of our search engine, has two tiers – a web client and an information collector at the server side. The Web client provides a web-based Graphical User Interface (GUI) and all interactive functions. (1) A *UI* includes a series of web components or pages, including basic/advance search pages, multiple records views (Paged list viewer, Bing Maps Viewer, Pivot Viewer), a quality viewer, a metadata detail viewer and others. (2) A *Presentation Controller* initializes web components or invokes external tools to show search results, data or quality information and controls user interactions through the GUI components that change with user context. (3) A *Communication Broker* manages the communication between client and server. (4) A *User Feedback/Action Collector* collects users' feedback about resource evaluation and captures user behaviours (e.g. which records are searched and viewed frequently and what viewers/tools are used).

The Information Collector retrieves metadata records and additional information and responds to search requests from clients. (1) A *Distributed Search Processor* (DSP) handles search requests from web clients and collaborates with other components as a coordinator. (2) A *Query Statement Creator* constructs query statement based on query rules. (3) A *Metadata parser* retrieves metadata abstract/detail from search sources. (4) a *Statistical/Post Processor* acquires, calculates and associates additional information (e.g. overall quality, relevance index and feedbacks) with retrieved records. (5) A *Search Statistical Database* stores search activities related to statistical information.

4. Methodologies

The methods described in Sections 4.1–4.3 provide a flexible framework to tackle heterogeneity on search sources that can be easily integrated with other GCIs. Section 4.4 discusses possible solutions to improve search performance and user interactivity. Section 4.5 addresses enhancing search accuracy. Section 4.6 describes methods for assisting user selection using quality information.

4.1. Loosely coupled brokering approach

GeoSearch adopts brokering middleware in its design, which is suitable for multi-source information integration applications, especially in heterogeneous decentralized environments. The GeoSearch search broker works as a bridge between users and GCIs.

In a typical enterprise or domain-based service-oriented architectures (SOAs), a single registry works well with limited users and providers. However, when these systems are integrated for larger applications in wider environments, it becomes difficult to establish a single centralized registry, because of cost, policy and security issues. In such a circumstance, the entire ecosystem becomes complex and ill-organized if multiple registries coexist without the centralized coordinator (discovery broker). As illustrated in Figure 2a, users have to go through various search sources for more resources. To obtain the best benefit and influence, resource providers have to publish resource metadata information into multiple portals/catalogues. Besides, complicated federation and harvest strategies may be adopted in catalogues/portals for implementing federated search.

The brokering approach provides a centrally controlled mechanism to integrate systems in divergent operating environments. Figure 2b shows the discovery environment with a broker middleware between users and distributed search sources. The brokering mode is an inevitable evolution for the traditional-three-roles-based SOA architecture in large-scale interoperable environments. It also provides a feasible solution for geospatial resource discovery from heterogeneous and decentralized sources.

4.2. Dynamically configurable and fault-tolerant discovery workflow

The capabilities to customize the search workflow against user requirements, preferences and network conditions dynamically are crucial to the distributed search system. To address this, we adopted the following workflow (Figure 4):

(1) When a user triggers a search, *the Communication broker* sends a group of asynchronous search request messages to the DSP.
(2) The DSP starts semantic-aided discovery and invokes the *semantic support system*, *query statement creator*, *search sources* and *metadata parser* in sequence.
(3) Once the results are obtained from the *metadata parser*, the DSP conducts post-processes to retrieve additional information or returns the results to the client directly without post-processes depending on workflow configuration.
(4) When the *communication broker* gets the search responses, *the presentation controller* initializes UI components to sort and present results.

The following describes the previous workflow. (1) For a request group (created for a user's single search request), a semantic pre-processing is triggered only once since the search

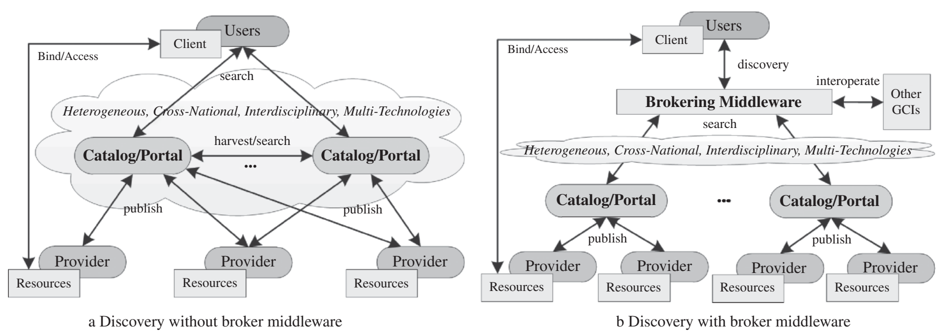


Figure 2. Discovery without or with broker middleware.

contents are the same. (2) By default, post-processing is done before sending results back to the client. Alternatively, the DSP can send results back to the client without post-processing (showing as grey arrows in Figure 3). (3) After result presentations are initialized at the client side, the *communication broker* will communicate with the *post processor* asynchronously to retrieve additional information in the background. This is suitable for the scenarios when users want quick response, but it will cause a higher communication cost between client and server. (4) The *post processor* interacts simultaneously with GCIs to minimize processing time.

Such a workflow with the collaboration of multi-layered components has the following advantages:

(1) Robust fault-tolerant capabilities. The DSP works as a centralized communication handler; so any single GCI's interoperation error will not affect the entire workflow. For example, if the semantic support system fails, the DSP can connect to substituted systems or escape semantic pre-processing and invoke the query statement creator directly. When the search fails on a source, there will be no impact on either the searches on other sources or the following user interactions.
(2) Configurable capabilities. Since all iterative discovery requests start from the communication broker at the client side, users can easily start and stop a search at any time. Furthermore, the end user can customize all search parameters and behaviours (e.g. search sources, maximum number of results, records number in each response, when and where to trigger what post-processing).

4.3. Plug-and-play framework for heterogeneous sources

In order to deal with heterogeneities (e.g. protocol, format and description) among various search sources, a plug-and-play extension framework for query statement creator and

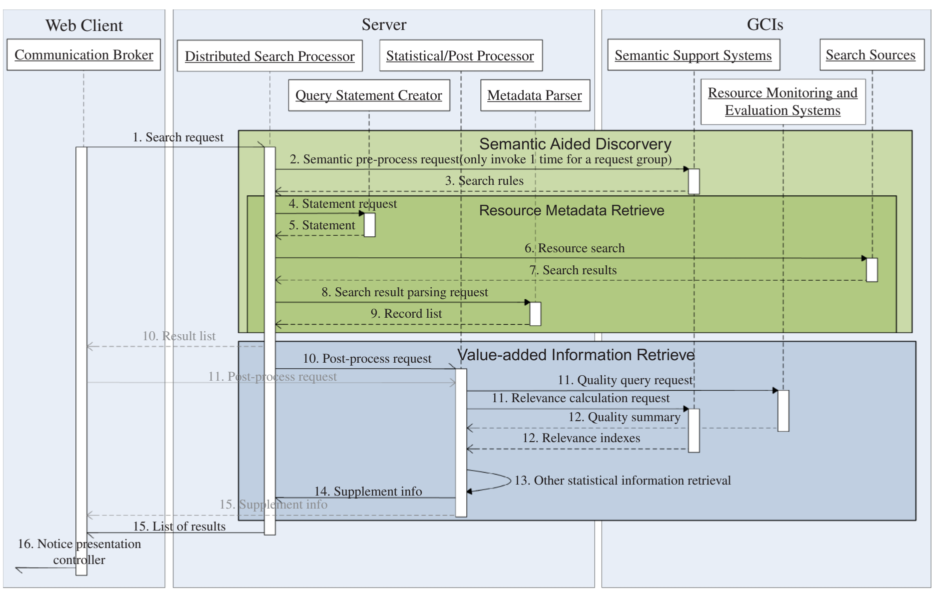


Figure 3. Resource search workflow.

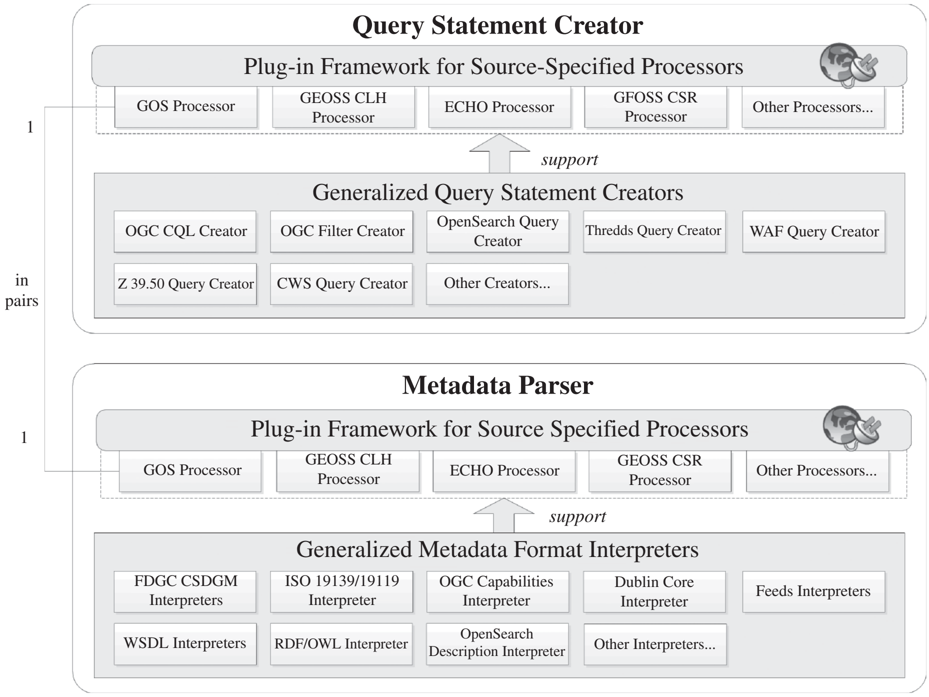


Figure 4. Plug-in frameworks for query statement creator and metadata parser.

metadata parser is designed by using plug-in mode (Figure 4). Using this framework, the capability to process certain search sources can be added/eliminated dynamically. As the backbone, the generalized query creators and metadata format interpreters provide basic process functions for specified resource types. When a new search source needs to be integrated, developers only need to develop a new pair of function modules which implement interfaces defined for statement creating and metadata parsing, respectively, and embed them into the information collector. Then, the new search source will be added to sources list and loaded by clients automatically.

4.4. An asynchronous progressive prefetching communication strategy

GeoSearch is a typical concurrent and communication-intensive application due to the loosely coupled architecture and discovery workflows. The intensive concurrency is due to the potentially large number of concurrent accesses to the engine from end users. Intensive communication is caused by frequent interaction with other GCIs. So, how to efficiently find massive numbers of resources without yielding poor user experience and interactivity is a big issue.

The shortest response time is one of the critical user expectations in any web-based applications (Yang *et al*. 2007, 2011b). Although many kinds of search sources provide functions or APIs to leverage the retrieval performance and the amount/detail level of retrieved content, the different access behaviours and retrieval strategies of different searching engines may also incur significantly different access performance and user experiences. Figure 5 summarizes four types of abstract search modes and their response time formulas

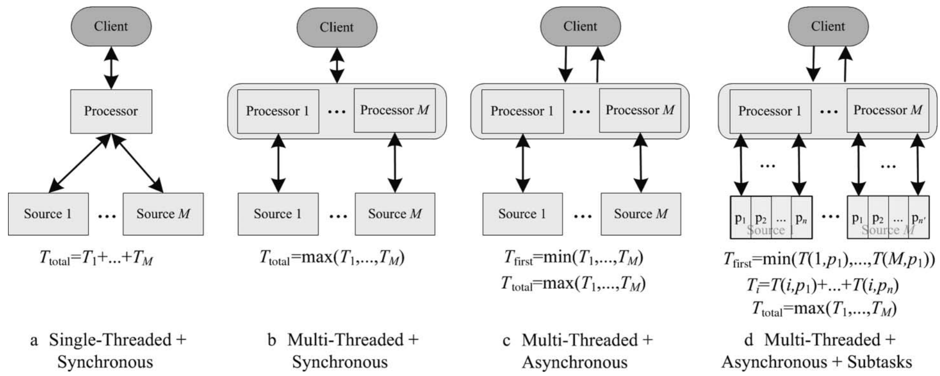


Figure 5. Abstract search mode.

from the searcher's perspective. Since there is no need to synchronously wait for all results from different sources and then display the results, Li *et al*. (2010) introduced a multi-threaded and AJAX-based asynchronous and parallel communication strategy (Figure 5c) where the first response time will dramatically drop to the minimum response time in M sources ($\leq 1/M$ of the sum of individual response time for each sources, which is the response time of the traditional mode without multi-threaded and asynchronous messaging as shown in Figure 5a), and the total search time depends on the maximum response time in M sources.

However, when the size of matched records in a source is large, retrieving all of them to client with a single request is inefficient due to the long response time and large volume of transmission. Considering the previous scenario, if we can divide one single request into a series of logically connected sub-requests and execute them in an optimized manner, each process unit should have improved response time since the time costs in retrieving, parsing and transmitting resources are reduced. To achieve this, many geospatial portals use passive click-and-retrieve mode (i.e. organize and retrieve records by page, and the records in a page are retrieved only when the user clicks the hyperlink of this page). This mode retrieves records exactly on demand, but may reduce interactivity and worsen user experience. Therefore, we propose a batch processing, multi-threaded and asynchronous messaging combined search strategy (Figure 5d) and temporarily cache all retrieved records in the client's memory for rapid processing and presentation. When a user starts a new search or when more search results are likely to be needed (e.g. user start to browse the last part of the result pages), the communication broker will send groups of asynchronous batch processing requests simultaneously to the DSP in the background. Each requests group searches against a certain source using independent threads and is composed of multiple requests which retrieve different fractions from the same source. When records corresponding to a request are retrieved, the DSP will send them back to the client. So, results are retrieved in an increasing and dynamic way, and the entire iterative search process has no influence on user interactivity.

This retrieval mode can be applied on various access protocols. For example, for CSW, the parameters *startPosition* and *maxRecords* of the *GetRecords* operation can be utilized to specify the retrieval fraction of the matched results. For OpenSearch v1.1, the parameters, *count*, *startIndex* and *startPage* can be used. Since CSW is a widely adopted catalogue service specification (e.g. GEOSS CLH, GOS and EuroGEOSS), this article uses it to illustrates how to implement the retrieval mode. Two kinds of Unit Retrieval Capacity (URCs) (*maxRecords*, maximum records number should be returned from matched results set in a request) are specified. One is *initial* (n_i) for the first request in the requests group, which should be small to improve first response time. Another is *regular* (n_r) for each request after the first request, which should be moderate to leverage performance and retrieval ratio. Supposing requests in a group are sent in sequential order, the iterative request number (I) and response time for a single search source will follow the following formulas:

$$I = \begin{cases} \lfloor (N_{\text{total}} - n_i) / n_r \rfloor + 2 \ \text{if}\, (N_{\text{total}} - n_i) \,\% n_r > 0 \\ \lfloor (N_{\text{total}} - n_i) / n_r \rfloor + 1 \ \text{if}\, (N_{\text{total}} - n_i) \,\% n_r = 0 \end{cases} \tag{1}$$

$$T_{\text{first}} \begin{cases} < T_{\text{single}} < T_{\text{total}} \approx T_{\text{first}} + (I - 1) \times T_{\text{regular}} & \text{if } N_{\text{total}} > n_i \\ = T_{\text{single}} = T_{\text{total}} = T_{\text{first}} & \text{if } N_{\text{total}} \le n_i \end{cases} \tag{2}$$

where N_{total} is the maximum records retrieval number, T_{first}, T_{total} and T_{regular} are the first response time, the total time and the average of regular response time, respectively, while T_{single} is the response time in the un-batched mode. The performance experiments are described in Section 5.

However, this retrieval mode may induce an overhead on network communication and an unnecessary fetching burden. How to allocate URCs and waiting intervals in each group is critical to balance performance and processing overhead. Network conditions (bandwidth and concurrency), the characteristics of search sources (access protocol, template format and size of metadata record), and user requirements should be considered. This will be part of our future research. To simply alleviate the above drawbacks, the following strategies are currently adopted: (1) retrieve abstract instead of all metadata detail, (2) configurable total maximum records retrieval number and URCs constraint and (3) suspend/stop/resume controls on fetching activity.

4.5. Semantic-aided search

In most operational catalogues and portals, searching is based on keyword matching between the search input and the content of metadata fields. But to some extent, both user inputs and metadata descriptions include ambiguity, inaccuracy and domain/culture limitations. It is hard to find all the resources that exactly fit the user's requirement, that is, recall and precision. Semantic reasoning gives us a solution. Semantic reasoning can help improve the recall via enriching the search query to match the entire cluster (e.g. different expression/terms and sub-concepts of the same concepts) rather than one or several terms. Additionally, semantic similarity evaluation provides a way to sort and filter search results by relevance and can improve the precision (Nasraoui and Zhuhadar 2010). However, to implement semantic technology in an existing SDI, one might need to construct a structured ontology and processing tools for the domain(s) of interest where such did not already exist. Our objective is to implement semantic-aided search with minimum reconstruction costs by utilizing existing GCI components. To achieve that, we developed a semantic support system which is independent from any catalogues and is loosely coupled with our search broker by expanding the ESIP semantic Testbed.[7] This system translates users' search input into a set of search rules that consist of a series of constructs and related

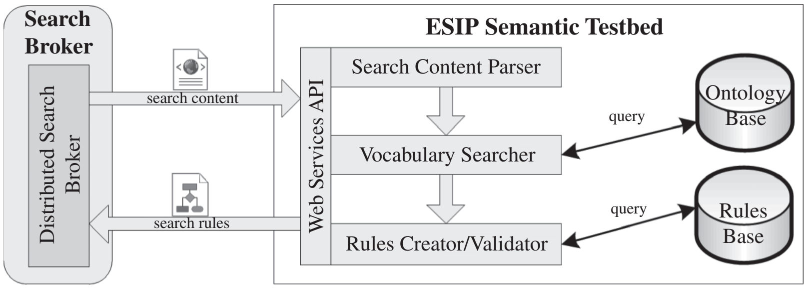


Figure 6. Interoperation between search broker and ESIP semantic Testbed.

vocabularies. Using the rules, the service broker constructs different query statements for different search sources. Figure 6 shows the interoperation process.

The following example shows how the system works. A user, for example, inputs '*USGS ecosystem WMS*' as a search content. The DSP sends it to the semantic support system where the search content is compared with the terminologies about the providers, resource types in established ontology base sequentially and decomposed into different segments. If the segments are neither provider nor resource type, they are considered as search subjects. Here, 'USGS', 'ecosystem' and 'WMS' are recognized as provider, search subject and resource-type constraints, respectively. Then, the subject '*ecosystem*'-related terminologies are queried. Based on that, the semantic support system will wrap the following rules in an XML payload and send it back to the DSP:

- Resource provider: *USGS* | *U.S. Geological Survey* | {. . . sub-sectors . . .}
- Resource Type: *OGC WMS*
- Subject: ecosystem | {*biome, ecosys,* terminologies (sub-domain/ multilingual) . . .}

Using the above query rules, the query statement creator will construct statements for different search sources. For CLH, the following OGC Filter fragment will be created to construct CSW *GetRecord* request (Table 1).

Since we still use keywords to search against each source, irrelevant records retrieval is inevitable. To help users identify them, we introduced relevance index from semantic similarity evaluation to measure the proximities between the search results and users' requirement. We adopted an edge-based similarity evaluation approach (Jiang and Conrath 1997, Liu *et al*. 2011): (1) pre-defined features set $F = \{f_i\}$ (metadata fields which participate calculation) and their contribution (weight set $W = \{w_i\}$) to the relevance; (2) calculate the similarity between each feature f_i and corresponding query rule r_i in rule set R by measuring the distance in the semantic ontology (Equation (3)); (3) summarize the similarity of every feature with its weight w_i into a single value *Re* (relevance index, Equation (4)).

$$\mathrm{sim}(f_i, r_i) = \frac{e}{\mathrm{dis}\,(f_i, r_i) + e} \tag{3}$$

$$\mathrm{Re} = \sum w_i \times \mathrm{sim}(f_i, r_i) \tag{4}$$

Table 1. OGC filter fragment sample for CSW GetRecords request.

```
<ogc:Or> <!--provider constraints-->
    <ogc:PropertyIsLike escapeChar="\" singleChar="?" wildCard="*">
        <ogc:PropertyName>OrganizationName</ogc:PropertyName>
        <ogc:Literal>USGS</ogc:Literal>
    </ogc:PropertyIsLike>
    <ogc:PropertyIsLike>...</ogc:PropertyIsLike>
</ogc:Or>
<ogc:Or> <!--resourceType constraints-->
   <ogc:PropertyIsEqualTo matchAction="One">
       <ogc:PropertyName>ServiceType</ogc:PropertyName>
       <ogc:Literal>OGC:WMS</ogc:Literal>
   </ogc:PropertyIsEqualTo>
</ogc:Or>
<ogc:Or> <!--keywords or full text constraints-->
   <ogc:PropertyIsLike escapeChar="\" singleChar="?" wildCard="*">
       <ogc:PropertyName>Subject</ogc:PropertyName>
       <ogc:Literal>ecosystem</ogc:Literal>
   </ogc:PropertyIsLike>
   <ogc:PropertyIsLike>...</ogc:PropertyIsLike>
</ogc:Or>
```

where e is an empirical constant correction value and $\mathrm{dis} = (f_i, r_i)$ is the measure distance. The similarity will get lower if the distance between f_i and r_i gets longer. When $\mathrm{dis} = (f_i, r_i)$ is equal to zero or f_i related rule does not exist, the similarity is 1.

4.6. *QoS-aware discovery/selection*

To provide quality and user feedback information for geospatial resources, we developed a resource monitoring and evaluation system – QoS Checker (Figure 7), which contains four components. (1) A *Query/Feedback broker* responds query/feedback requests. (2) A

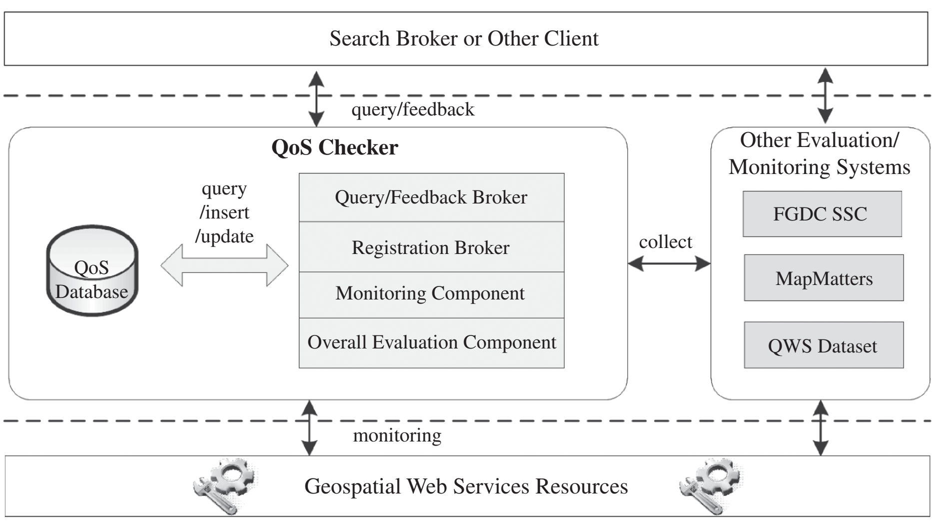


Figure 7. Architecture and workflow of QoS checker.

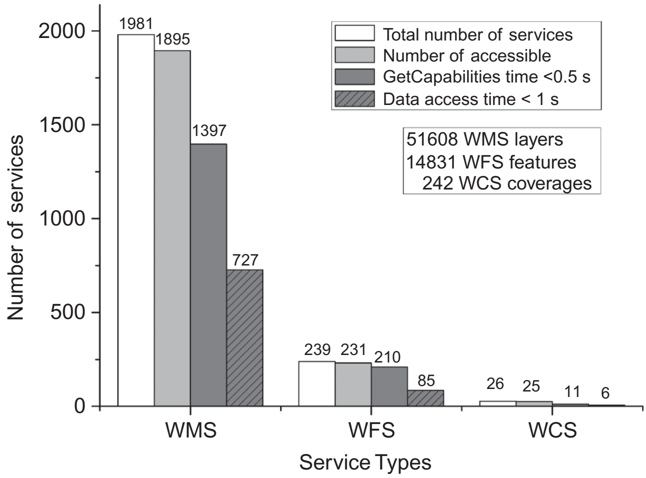


Figure 8 Accessibilities and performances of monitored OGC data services.

Registration broker inserts new services needed to be monitored. (3) A *Monitoring component* monitors the performance of registered services daily. (4) An *Overall evaluation component* evaluates the overall quality of services. The major characteristics of this system are as follows:

(1) Quality information from multiple channels (active monitoring, end users feedback and other monitoring systems[8]);
(2) Quality query capabilities in different granularities (overall quality grade from long-time statistics and service/operations/dataset levels performance in a certain period of time).

Figure 8 shows the accessibilities and performance statistics of registered OGC data and portrayal services. Since the data access response time is different upon the data size and measurement method, we summarize the average response time of retrieving 1 layer/coverage (250 $\times$ 250 pixels) for WMS/WCS, and 20 features for WFS as our measurement. We found that the performance varies. Only about one in three services provide good data access performance. The performance information can be used as a selectable sorting criterion to help end users distinguish services, which provide the same data/functionalities, and also help service providers find a potential performance bottleneck. With an increasing number of services being monitored (Wu *et al.* 2011), user ranking, feedback and relevance being assessed, such quality information will improve resource selection.

5. Implementation and experiments

5.1. Implementation

The search broker is developed using Microsoft.Net Framework 4. Its web client adopts Silverlight to provide cross-browser uniform GUI and rich interactive capacity. The information collector is based on ASP.NET. The communications between GeoSearch and

GCIs components and also GeoSearch internal components use standard SOAP/RESTful web-service interfaces. Most importantly, 'Silverlight-enabled Windows Communication Foundation' (WCF) services are utilized to implement asynchronous communication between the web client and the information collector at the server side. These technologies will help users develop Rich Internet Application (RIA) with user-friendly interface and scalable processing capacities.

In the current stage, GeoSearch can support GOS, GEOSS CLH, GEOSS CSR, CIWC, USGIN AASG Geothermal Data Catalog and other CSW search sources. Sources with other protocols and metadata formats are under development (e.g. ECHO, THREDDS). BingMaps, Openlayers, World Wind (web version) and Spatial Web Portal (SWP) (Xu *et al.* 2011) have been integrated as visualization tools for OGC data services. Tools to visualize and process datasets are under consideration. *User Feedback/Action Collector* and *Statistical/Post Processor* are still under development.

5.2. Search performance experiment

Two experiments were performed to test the search performance of our batched progressive prefetching mode.

(1) We compared our batched progressive prefetching mode with un-batched asynchronous multi-threaded mode by searching against two CSW catalogues (i.e. GEOSS CLH and GOS) in the same environment with 100 Mbps internet connection. In the experiment, we set the initial and regular URCs as 50 and 300, respectively, and assume that both sources have a sufficiently large amount of matched records. Figure 9 shows the response time comparison in different retrieval levels (total retrieved records number). The response time for the

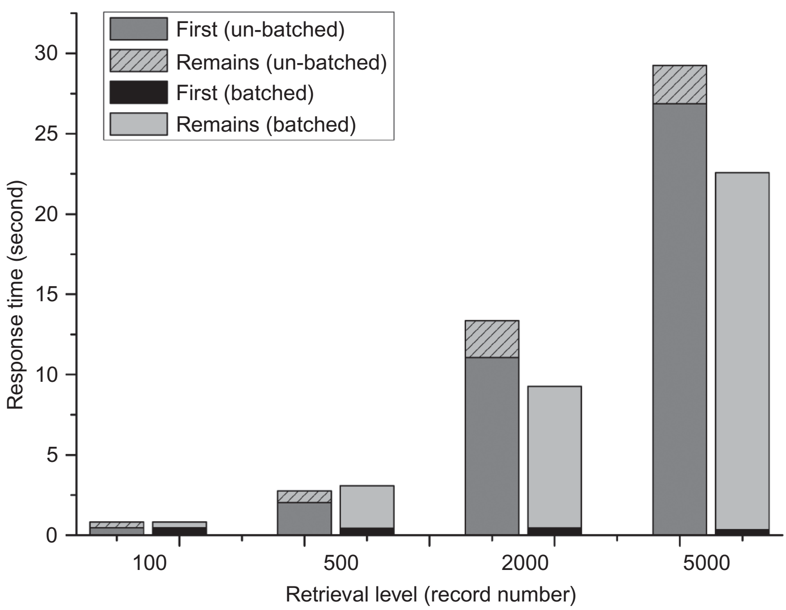


Figure 9. Response time comparison for batched and un-batched mode.

un-batched mode depends on the retrieval level distinctly, but the average first response time and regular response interval in the batched mode are stable and short (0.5 and 1.3 s, respectively). At a small retrieval level (e.g. 500 records), since the batched mode uses more requests to fetch the same amount of records, the total response time is a little higher than in the un-batched mode which is in compliance with Equation (2). However, if there are multiple search sources, as the retrieval number increases, clients will fetch more records from the sources with good performance and spend less time than in the un-batched mode (as shown in retrieval level 2000 and 5000). The more the search sources, the quicker the retrieval speed in the batch mode than in the un-batched one.

(2) We compared the response time on switching web pages with passive click-and-retrieve mode (GOS and GEOSS GEOportal[9]). Although records are retrieved exactly on demand in the click-and-retrieve mode, the response time increases accordingly (Figure 10). One of the reasons the apparent response time of GEOPortal is longer than GOS was due to network latency between the United States and Europe. In contrast, since all metadata in existing pages have been retrieved by the client, the response time is smaller and more stable. Well-configured progressive prefetching mode could provide better interactivity to end users (especially, when they want to quickly browse the search result set) without leading to communications overhead.

5.3. Search accuracy improvement

The simple text-based search example ‘water wms’ in Figure 11 shows search accuracy improvement with semantic support. Since ‘ice’, ‘ocean’, ‘lake’, ‘river’ and other water-related terminologies will be included as search keywords. A total of 42 WMS were found by a search against GEOSS CLH. In contrast, without semantic assistance, the result number will be 985 but only 22 WMS, and 944 dataset and 19 other records (e.g. WCS and WFS) are included incorrectly. So, by collaborating with the semantic supporting system, users can get more of what they really need and eliminate those irrelevant results.

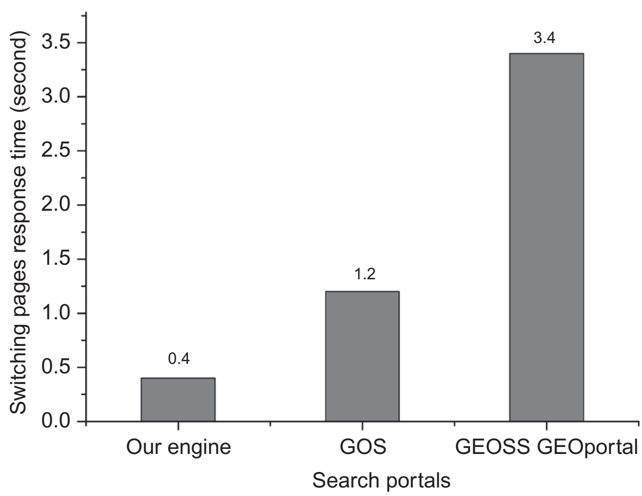


Figure 10. Switching pages response time comparison.

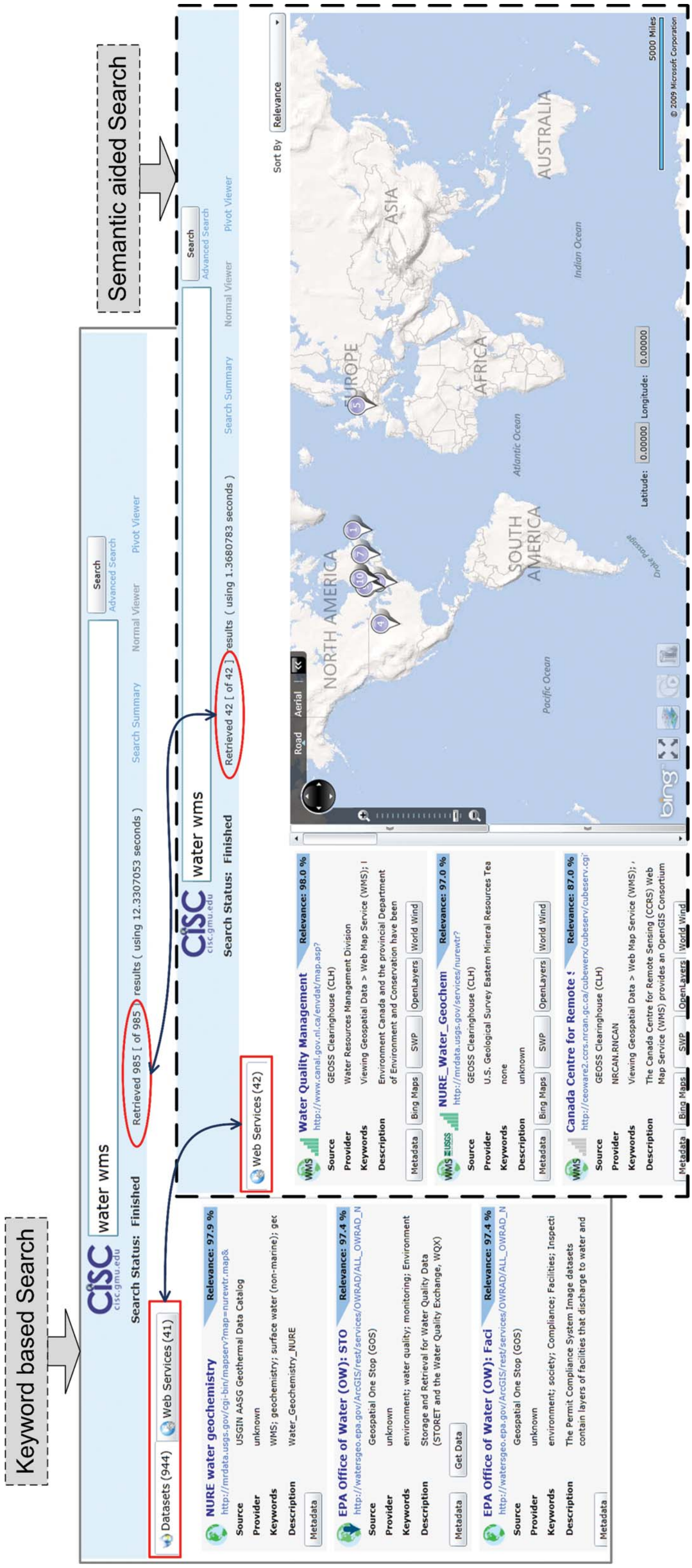
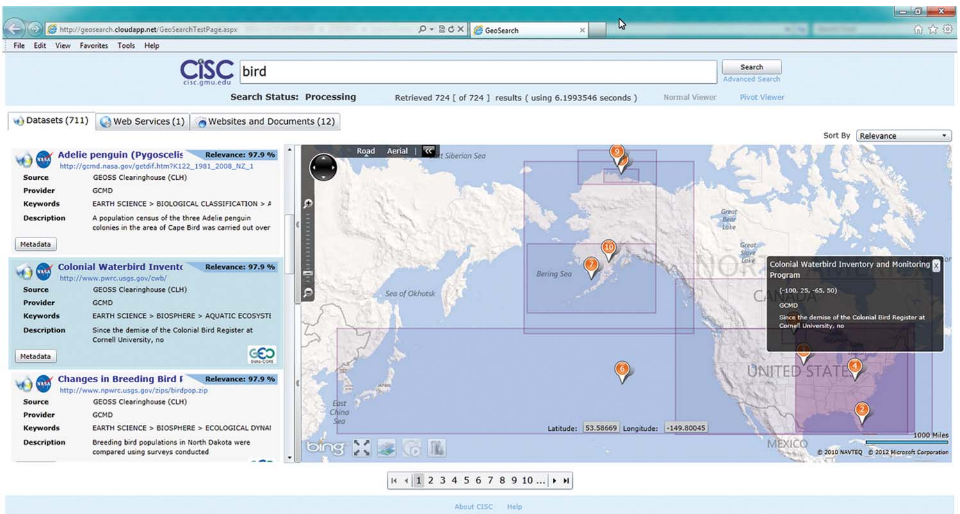


Figure 11. Keyword and semantic aided search comparison.

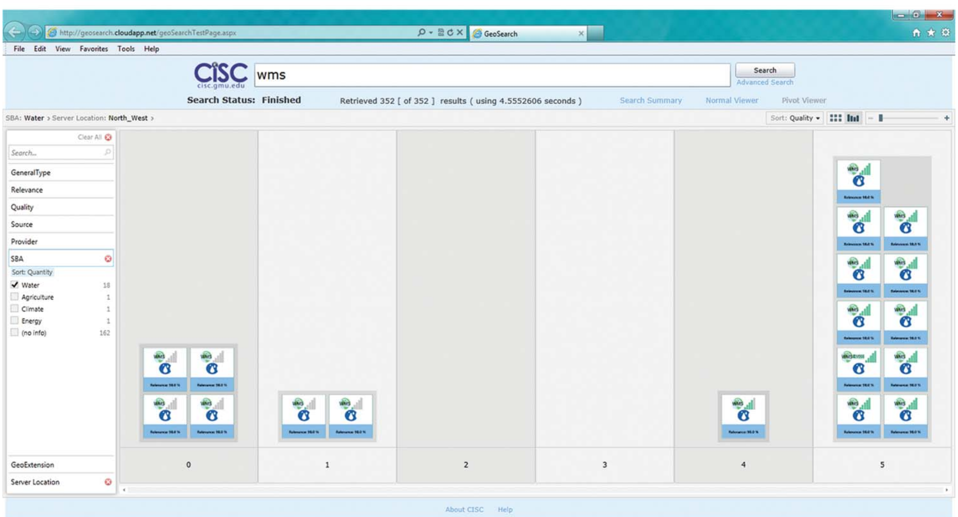
5.4. Integrated GUI for assisting in decision-making

Interactivity and intuitive information visualization methods are critical factors of any framework that needs widespread and voluntary adoption. Following this rule, we developed a series of interactive viewers (e.g. Paged List Viewer, Bing Maps Viewer and Pivot Viewer, Service Quality Viewer) and functions, and integrated them through a single client UI. Users can easily browse and interact with search results in any of the above viewers or their combination.

(1) A Paged List Viewer (Figure 12a) provides traditional, paged, list-based view to cluster records with categorized resource types (e.g. dataset, services, documents, web sites, systems and models) and multiple sorting rules (e.g. relevance, quality, provider and search source).



a Paged list viewer & bing maps viewer (Bird_Searched)



b Pivot viewer (WMS_Searched)

Figure 12. GUI of search result page.

(2) Bing Maps Viewer (Figure 12a) shows records and their location information (e.g. bounding box of data, server location of services/website/applications) in map context.
(3) A Pivot Viewer filters and sorts records through various rules and displays records with deep-zoomable column diagrams (Figure 12b) where found WMSs are filtered by Societal Benefit Area and server location and then clustered by service quality, which can help users find the potential patterns and correlations hidden in the result collection.
(4) A Service Quality Viewer (Figure 13) shows the response time of the *getCapabilities* operation, FGDC Scores and each layer's performance for a specified WMS in a certain period with diagrams. These quality information reveals the non-functional properties and can help users distinguish the services with the same function and data.
(5) The integrated visualization tools (i.e. Bing Maps Viewer, World Wind, SWP and Openlayers) (Figure 14) help users interact with the retrieved OGC data services. The download tools facilitate resources access.

Furthermore, the relevance index, which measures the similarities between the user's requirement and the retrieved records, can be utilized as sorting and filtering rules for the result set. In short, though integrating these additional information/tools, user's experience and decision-making capacity would be improved.

6. Conclusions

This article introduces the framework and methodologies of GeoSearch, a system aiming to improve geospatial resource discovery by leveraging and utilizing existing GCIs with a low-integration cost.

(1) To hide cross-domain barriers and heterogeneities to users and providers, a loosely coupled plug-in search framework is proposed based on brokering middleware mode and configurable search workflow. By adopting the framework, communication and maintenance costs between federated catalogues can also be reduced.
(2) To improve search performance and user interactivity, a batch processing-based asynchronous progressive prefetching mode is introduced to prefetch a moderate number of metadata abstracts. This strategy retrieves results on demand strictly but with a high performance and interactivity cost in that finer granularity and frequent retrievals will incur concurrency access burden to GCI components.
(3) To improve recall level and precision, a semantic support system is developed by expanding the ESIP Semantic testbed with a new methodology. However, further research is needed to improve the search accuracy and recall level.
(4) To assist in decision-making, a series of viewers, external visualization tools and functions are integrated through a sophisticated GUI. All those would enrich the user's experience and help users make selection intuitively and wisely.

7. Future research

For additional improvements in geospatial resource discovery, access and usage, future research would include:

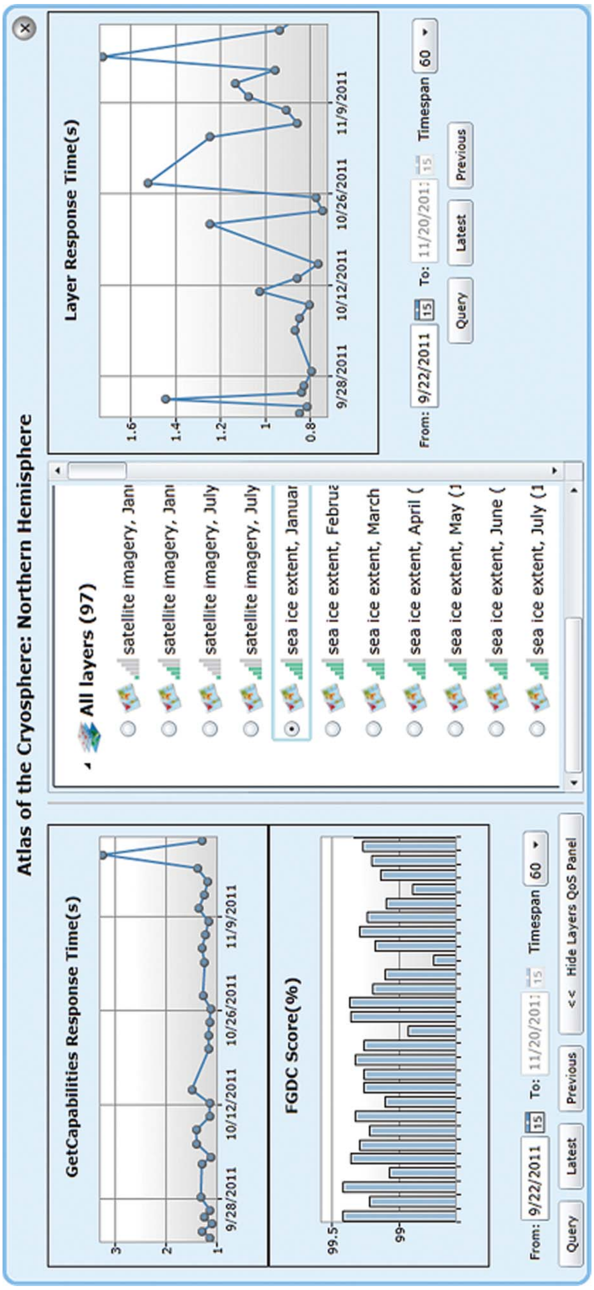
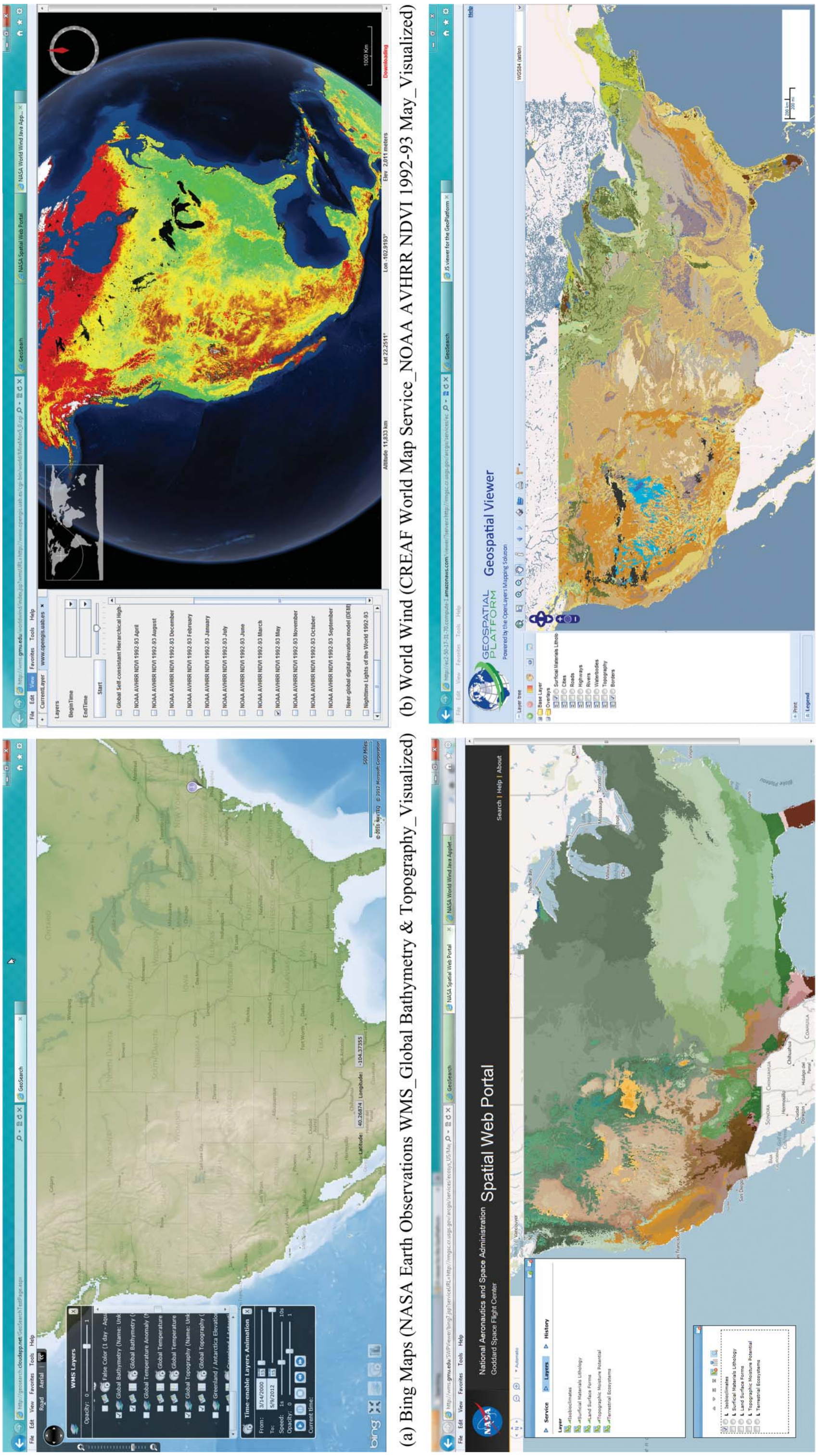


Figure 13. GUI of service quality viewer.



(a) Bing Maps (NASA Earth Observations WMS_Global Bathymetry & Topography_Visualized)

(b) World Wind (CREAF World Map Service_NOAA AVHRR NDVI 1992-93 May_Visualized)

(c) Spatial Web Portal (USGS Ecosystem Service_Isobioclimates_Visualized)

(d) OpenLayers (USGS Ecosystem Service_Surficial Materials Lithology_Visualized)

Figure 14. GUI of visualization tools.

(1) Improve performance and broad network access capabilities. We intend to deploy our engine onto cloud platforms and investigate spatiotemporal principles (Yang *et al*. 2011a) to further optimize the search performance and reduce costs. We will design an infrastructure to dynamically allocate search engine instances at the right place to meet the dynamic demand and make it more accessible and faster in delivering responses to end users from different regions around the world by investigating and deploying geographically distributed load balancing and availability and minimizing GCI components communication cost. Since any discovery operation will trigger distributed queries on the fly and incur network communications, we will refine our progressive prefetching algorithm and make the retrieval parameters (retrieval unit capacities and waiting time intervals) self-adaptable to real-time network condition and user's search request. To alleviate the communication burden and deal with the temporary off-line problems of search sources, we will investigate the efficient cache/store and index mechanism.

(2) Integrate more useful information and functionalities to assist in resource selection. We plan to extend statistics and analytical functions on resource popularity and user feedbacks to help users identify resources through other users' selection and evaluation by collecting access frequency of metadata, records/data, data downloads, data visualization and quality information. Meanwhile, we will enhance the semantic search capabilities, such as multilingual and advanced word segmentation and matching methods.

(3) Enhance value-added search capabilities. Associating data with services can help users discover tools to work with the retrieved data and vice versa. These information can be inter-workable both horizontally (data with data, result with result) and vertically (data with tool, with workflow, with result, with experiment) (Duerr *et al*. 2011). We will provide the glue function to integrate the components of data, processing, applications and infrastructure (Yang *et al*. 2010). Furthermore, we plan to develop a user registration and subscription service to help users find potential resources from profile and search context and history and also establish a professional networking system to help researchers and scientists find potential collaborators.

(4) Provide and consume standard web 2.0 search and casting interfaces (such as OpenSearch and RSS) to facilitate integration and interoperation with other systems.

Acknowledgements

This work was supported by NSF (IIP-1160979 and CNS-1117300), NASA (NNX12AF89G), FGDC GeoCloud and GEOSS Clearinghouse grants and Microsoft Research.

Notes

1. Examples are National Oceanic and Atmospheric Administration (NOAA) Coastwatch and SWFSC/Environmental Research Division, National Aeronautics and Space Administration (NASA) Physical Oceanography Distributed Active Archive Center.
2. An example is the Center for Ocean-Atmospheric Prediction Studies, Florida State University.
3. http://datacasting.jpl.nasa.gov/
4. http://sciflo.jpl.nasa.gov/scast/
5. http://wiki.esipfed.org/index.php/Discovery_Cluster
6. http://www.eurogeoss.eu

7. http://wiki.esipfed.org/index.php/Semantic_Web_Testbed_Implementation
8. Examples are Service Status Checker (SSC, http://registry.fgdc.gov/statuschecker/index.php) provided by the Federal Geographic Data Committee (FGDC), MapMatters (http://www.mapmatters.org/) and QWS Dataset (http://www.uoguelph.ca/~qmahmoud/qws/index.html).
9. http://devos.sapienzaconsulting.com/

References

Alam, A., Khan, L., and Thuraisingham, B., 2011. Geospatial Resource Description Framework (GRDF) and security constructs. *Computer Standards and Interfaces*, 33 (1), 35–41.

Clement, L., *et al*., 2004. OASIS UDDI, Version 3.0.2, Available from: http://www.oasis-open.org/committees/uddi-spec/doc/spec/v3/uddi-v3.0.2-20041019.htm [Accessed 10 October 2011].

Clinton, D., 2011. OpenSearch Specification [Online], Version 1.1, Draft 5. Available from: http://www.opensearch.org/Specifications/OpenSearch/1.1 [Accessed 10 October 2011].

Craglia, M., *et al*., 2011. Inter-disciplinary interoperability for global sustainability research. *In*: C. Claramunt, S. Levashkin, and M. Bertolotto, eds. *GeoS 2011, lecture notes in computer science for proceedings of Geos 2011 conference, LNCS 6631*. Heidelberg: Springer-Verlag, 1–15.

Di, L., *et al*., 2006. Ontology-driven automatic geospatial-processing modeling based on web-service chaining. *In*: *Proceedings of the sixth annual NASA earth science technology conference*, 7 pp.

Domenico, B., *et al*., 2002. Thematic real-time environmental distributed data services (THREDDS): incorporating interactive analysis tools into NSDL. *Journal of Digital Information*, 2 (4), Article No. 114.

Duerr, R., *et al*., 2011. Discovery white paper short [Online], ESIP Federation Wiki. Available from: http://wiki.esipfed.org/index.php/Discovery_White_Paper_Short [Accessed 10 October 2011].

Fulker, D., Bates, S., and Jacobs, C., 1997. Unidata: a virtual community sharing resources via technological infrastructure. *Bulletin of the American Meteorological Society*, 78, 457–468, ISSN 0003-0007.

Jiang, J. and Conrath, D., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *In*: *Proceedings of international conference research on computational linguistics (ROCLING X)*, Taiwan.

Klien, E., Lutz, M., and Kuhn, W., 2006. Ontology-based discovery of geographic information services: an application in disaster management. *Computer, Environment and Urban Systems*, 30, 102–123.

Krishnan, S., *et al*., 2011. OpenTopography: a services oriented architecture for community access to LIDAR topography. *In*: *Proceedings of the 2nd international conference on computing for geospatial research & applications (COM.Geo '11)*. New York, NY, USA: ACM.

Li, W. and Yang, C., 2008. A semantic search engine for spatial web portals. *In*: *Proceedings of IEEE IGRASS'08*, 6–11 July 2008. Boston, MA, USA, II-1278–1281.

Li, W., Yang, C., and Yang, C.J., 2010. An active crawler for discovering geospatial web services and their distribution pattern – a case study of OGC web map service. *International Journal of Geographic Information Science*, 24 (8), 1127–1147.

Li, Z., *et al*., 2011. An optimized framework for seamlessly integrating OGC Web services to support geospatial sciences. *International Journal of Geographical Information Science*, 25 (4), 595–613.

Liu, K., *et al*., 2011. The GEOSS clearinghouse high performance search engine. *In*: *The 19th international conference on geoinformatics*, 24–26 June 2011, Shanghai, China.

Lutz, M., 2007. Ontology-based descriptions for semantic discovery and composition of geoprocessing services. *Geoinformatica*, 11, 1–36.

Lutz, M. and Klien, E., 2006. Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science*, 20, 233–260.

Mapping Sciences Committee, 1993. *Towards a coordinated spatial data infrastructure for the nation*. Washington, DC: National Academy Press.

Mazzetti, P. and Nativi, S., 2011. Multi-disciplinary interoperability for Earth observation sensor web. Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International, 978-1-4577-1005-6/11.

Najmi, F., Stojanovic, N., and Breininger, K., 2011a. OASIS ebXML RegRep, Part 1: Registry Information Model (ebRIM) [Online], Version 4.0. Available from: http://docs.oasis-open.org/regrep/regrep-core/v4.0/cs01/regrep-core-rim-v4.0-cs01.pdf [Accessed 10 October 2011].

Najmi, F., Stojanovic, N., and Breininger, K., 2011b. OASIS ebXML RegRep, Part 2: Services and Protocols (ebRS) [Online], Version 4.0. Available from: http://docs.oasis-open.org/regrep/regrep-core/v4.0/cs01/regrep-core-rs-v4.0-cs01.pdf [Accessed 10 October 2011].

Nasraoui, O. and Zhuhadar, L., 2010. Improving recall and precision of a personalized semantic search engine for e-learning. *In: 2010 Fourth international conference on digital society*.

Nativi, S. and Bigagli, L., 2009. Discovery, mediation, and access services for earth observation data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2 (4), 233–240.

Nebert, D., 2007. OpenGIS Catalogue Services Specification [Online], Version 2.0.2, OGC 07-006r1. Available from: https://portal.opengeospatial.org/modules/admin/license_agreement.php?SuppressHeaders=0&access_license_id=3&target=http://portal.opengeospatial.org/files/%3fartifact_id=20555 [Accessed 10 October 2011].

NSF, 2011. Earth Cube Guidance for the Community [Online]. Available from: http://www.nsf.gov/pubs/2011/nsf11085/nsf11085.pdf [Accessed 13 October 2011].

Ramachandran, R., *et al*., 2006. Ontology-based semantic search tool for atmospheric science. *In*: *22nd International conference on interactive information processing systems for meteorology, oceanography, and hydrology*.

Sample, J.T., *et al*., 2006. Enhancing the US Navy's GIDB portal with web services. *Internet Computing, IEEE*, 10, 53–60.

Santoro, M., *et al*., 2011. Enhancing discovery capabilities of SDIs with semantic queries. *In*: *Proceedings of EGU general assembly 2011, geophysical research abstracts*, 13, EGU2011-5963-4.

Van, H., *et al*., 2004. Resource harvesting within the OAI-PMH framework. *D-Lib Magazine*, 10 (12), ISSN 1082-9873.

Walsh, J., Gonçalves, P., and Turner, A., 2009. [Online], Version 0.0.1, OGC 09-084rl. Available from: https://portal.opengeospatial.org/modules/admin/license_agreement.php?suppressHeaders=0&access_license_id=3&target=http://portal.opengeospatial.org/files/%3fartifact_id=35983 [Accessed 10 October 2011].

Wu, H., *et al*., 2011. Monitoring and evaluating the quality of web map service resources for optimizing map composition over the internet to support decision making. *Computers and Geosciences*, 37, 485–494.

Xu, C., *et al*., 2011. A service visualization tool for spatial web portal. *In*: *Proceedings of the 2nd international conference on computing for geospatial research & applications (COM.Geo '11)*. New York, NY, USA: ACM.

Yang, C., *et al*., 2005. Performance improving techniques in WebGIS. *International Journal of Geographical Information Science*, 19 (3), 319–342.

Yang, C., *et al*., 2010. Geospatial Cyberinfrastructure: past, present and future. *Computers, Environment, and Urban Systems*, 34 (4), 264–277.

Yang, C., *et al*., 2011a. Using spatial principles to optimize distributed computing for enabling the physical science discoveries. *Proceedings of National Academy of Sciences*, 108 (14), 5498–5503.

Yang, C., *et al*., 2011b. WebGIS performance issues and solutions. *In*: S. Li, S. Dragicevic, and B. Veenendaal, eds. *Advances in Web-based GIS, Mapping Services and Applications*. London: Taylor and Francis, 121–138.

Yang, P., Cao, Y., and Evans, J., 2007. WMS performance and client design principles. *Journal of GeoInformation Science and Remote Sensing*, 44 (4), 320–333.

Yue, P., *et al*., 2007. Semantics-based automatic composition of geospatial web service chains. *Computers & Geosciences*, 33 (5), 649–665.

Zhang, C., *et al*., 2010. Towards logic-based geospatial feature discovery and integration using web feature service and geospatial semantic web. *International Journal of Geographical Information Science*, 23 (6), 903–923.

Zhang, T. and Tsou, M.-H., 2009. Developing a grid-enabled spatial web portal for internet GIServices and geospatial cyberinfrastructure. *International Journal of Geographical Information Science*, 23 (5), 605–630.