International Journal of Digital Earth
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t777764757

Distributed geospatial information processing: sharing distributed geospatial resources to support Digital Earth

Chaowei Yang[a]; Wenwen Li[a]; Jibo Xie[a]; Bin Zhou[a]
[a] Joint Center for Intelligent Spatial Computing, College of Science, George Mason University, Fairfax, VA, USA

To cite this Article Yang, Chaowei , Li, Wenwen , Xie, Jibo and Zhou, Bin(2008) 'Distributed geospatial information processing: sharing distributed geospatial resources to support Digital Earth', International Journal of Digital Earth, 1: 3, 259 — 278
To link to this Article: DOI: 10.1080/17538940802037954
URL: http://dx.doi.org/10.1080/17538940802037954

International Journal of Digital Earth,
Vol. 1, No. 3, September 2008, 259–278

Distributed geospatial information processing: sharing distributed geospatial resources to support Digital Earth

Chaowei Yang*, Wenwen Li, Jibo Xie and Bin Zhou

Joint Center for Intelligent Spatial Computing, College of Science, George Mason University, Fairfax, VA, USA

(Received 14 August 2007; accepted 26 September 2007)

This paper introduces a new concept, distributed geospatial information processing (DGIP), which refers to the process of geospatial information residing on computers geographically dispersed and connected through computer networks, and the contribution of DGIP to Digital Earth (DE). The DGIP plays a critical role in integrating the widely distributed geospatial resources to support the DE envisioned to utilise a wide variety of information. This paper addresses this role from three different aspects: 1) sharing Earth data, information, and services through geospatial interoperability supported by standardisation of contents and interfaces; 2) sharing computing and software resources through a GeoCyberinfrastructure supported by DGIP middleware; and 3) sharing knowledge within and across domains through ontology and semantic searches. Observing the long-term process for the research and development of an operational DE, we discuss and expect some practical contributions of the DGIP to the DE.

Keywords: Digital Earth; DGIP; interoperability; cyberinfrastructure; ontology; semantic search

Introduction

The past half century has witnessed rapid advancements in various computing and information technologies (Joseph and Shmuel 2007). The advancements enable us to digitise information and data about the Earth to develop services for facilitating our daily life. For example, the online driving routing systems and geographical positioning system (GPS)-integrated car navigation systems provide us convenient tools (Rae-Dupree 2006). Selecting, registering, or booking merchandise online relieves us from the old fashion of doing business in person. The emergence and flourishing of the internet provides us with vast amounts of information and tools to improve almost every aspect of our life (Patrice 2007). Observing the fast-paced technology evolution, Vice President of the United States Al Gore proposed to develop a Digital Earth (DE) to leverage and advance those technologies to facilitate our life with better services (Gore 1998). In 1999, the International Symposium on DE (ISDE) was held in Beijing, China to discuss the technology, science, policy, and other needs of building such a DE (Guo and Yang 1999). Thereafter, four bi-annual symposiums were held in New Brunswick, Canada (http://www.agu.org/cgi-bin/gycal?details = 1&file = 2001062400

*Corresponding author. Email: cyang3@gmu.edu

ISSN 1753-8947 print/ISSN 1753-8955 online

DOI: 10.1080/17538940802037954
http://www.informaworld.com

320182800), the Czech Republic (http://digitalearth03.geogr.muni.cz/), Japan (http://de.gsec.keio.ac.jp/digitalearth/past/sympo2005/index.html), and the United States (http://www.isde5.org/) to advance the vision of and to implement the DE. Through a decade of research and development, especially in China, the DE has been greatly advanced and was taken as a flagship for informationisation at different levels, from the DE (Cheng *et al.* 2000), to digital China (Chen and Wu 2003), to digital province (Li 2001), and to digital counties and cities (Li and Lin 2001). These divisions not only help with informationisation, but also help to build different applications for citizens of Earth. Even though geospatial information are better at providing reliable services to citizens than ever before, we still confront many global issues, such as global warming, sea level change, public health, air quality, and other challenges (http://www.liveearth.org/).

To deal with these global challenges, numerous sensors are built to observe the Earth surface, atmosphere, solid Earth, and the ocean in different dimensions to help us understand the planet in a scientific and comprehensive fashion. As termed by the Global Earth Observation (GEO 2005), we are constantly taking the pulse of the Earth to diagnose the potential problems and issues. For example, 1) GOES (http://www.goes.noaa.gov), GOES-R (http://www.osd.noaa.gov/goes_R/), CLOUDSAT (http://www.nasa.gov/cloudsat/), and CALIPSO (http://www-calipso.larc.nasa.gov/) are helping us understand the formation of rain, clouds, and other atmospheric phenomena; 2) TRMM (http://trmm.gsfc.nasa.gov/), GPM (http://gpm.gsfc.nasa.gov/), and ground-based sensors, such as gauges, are observing the precipitation process by combing space-based, airborne-based, ground-based, and water-based sensors. Precipitation observations further contribute to the research of the Earth's water cycle (Huffman *et al.* 2007); 3) ice observing satellites and airborne sensors are detecting the ice thickness in the polar regions to conduct research on how ice melted in the past year, decade, and century has increased sea level by about 50 cm to 1 m (Mörner 2004), which, as a consequence, increases the destructive power of hurricanes and causes many ocean or coastal disasters, such as Hurricane Katrina (*IT Professional* 2005).

To better understand the linkage of these phenomena and to make better decision-making information available to respond to these disasters and other global issues, NASA and its partner agencies, NOAA, EPA, and USDA, have worked together and envisioned 12 national application areas (agricultural efficiency, air quality, aviation, carbon management, coastal management, disaster management, ecological forecasting, energy management, homeland security, invasive species, public health, and water management (Birk *et al.* 2006) that can benefit research and development addressing these global issues. Similarly, the GEO was established in 2003 as an international organisation with its secretariat in Geneva to address global issues from 9 societal benefit areas (GEO 2005), which share some areas with the 12 national applications. These Earth observation systems directly contribute to the DE (Guo 1999) and provide envision-level incentives to the advancement of the DE for dealing with the challenges related to our daily life and the sustainable development of our home planet.

However, these visions can only be achieved by sharing geospatial resources located at different geographical locations across jurisdiction boundaries. Such sharing includes at least 3 levels (Yang 2007): a) sharing the earth observation data and information collected by a variety of sensors and generated by a variety of models and simulations, b) sharing software and computing resources to improve the response time for demanding computing processing requirements, and c) sharing the knowledge gained from different domains to benefit all levels of users by formalising our understanding of the problem domains and utilising them for application development.

To respond to the sharing requirements, the GEO proposed the Global Earth Observation System of Systems (GEOSS) architecture to integrate the Earth observation systems and Earth system models to support the decision support tools for policy and management decisions (GEO 2005). The systems or models/tools are geographically distributed. The integration of both distributed data and processing components fit into the concept of Distributed Geospatial Information Processing (DGIP), a term we coined with our partners to capture the recent developments in geospatial information processing that is characterised by distributed datasets and distributed processing but connected by computer networks (such as Yang and Raskin 2007). DGIP is defined to include research and development spans across five different areas (http://www.cisc.gmu.edu/): a) DGIP architecture and algorithms, b) DGIP interoperability issues, c) GeoCyberinfrastructure, a high performance geospatial information computing platform, d) DGIP intelligent geospatial processing, and e) DGIP applications of national and international interests. While relevant GIS terms (Coleman 1999), such as Network GIS, deal with geographical information, DGIP deals with all different types of geospatial information including, for example, atmospheric modelling information.

The evolution of DGIP has been accompanied by the advancement of computer networks. For example, the early distributed computing of client/server (C/S) mode (Foote and Kirvan 1998) has evolved into the browser-server (B/S) or web-based GIS (Xu *et al.* 2003; Yang 2000). The recent developments of ubiquitous computing expand the architecture to multi tiers (Yang and Tao 2005) with a promising network computing vision of 'any time, anywhere'. This also fits into the architecture of the DE (Li *et al.* 1999). The developments will provide ideal platforms to share data, information, computing, and knowledge across different scientific and user domains. The popularisation of the DGIP applications, such as Google earth (http://earth.google.com/), help advance the field by both providing vast amounts of online information and cultivating a large number of users for future applications of national and international interests.

Five research aspects of DGIP directly contribute to the development of the DE by providing a) an infrastructure and platform built on the principles of DGIP architecture and algorithms, b) interoperable approaches to share the widely distributed and heterogeneous geospatial resources, c) computing infrastructure to share the widely distributed computing resources, d) intelligent approaches to share knowledge using ontology and distributed resources, and e) best-practices of the GEOSS societal benefit areas (including Disasters, Health, Energy, Climate, Water, Weather, Ecosystems, Agriculture, and Biodiversity, GEO 2005). We expect that the advancements in these aspects help to build an integral and interoperable infrastructure to support the seamless integration of broadly distributed geospatial resources (Yang 2007) for the DE. The following sections detail three areas: Section 2 discusses and provides examples of the DGIP architecture and interoperable platform for sharing data, information, and services. Section 3 discusses GeoCyber-infrastructure for supporting the operation of significant applications of importance to national and international interests. Section 4 uses Air Quality as an example to demonstrate the sharing and utilisation of knowledge, and the last section summarises and discusses further research.

Interoperability: sharing geospatial data, information, and services

Earth observation data and Earth simulation results have been produced and archived at widely distributed locations. Other geospatial data and information are also widely collected and archived to record historical Earth phenomena and would be very useful in providing services in the DE and solving challenging global issues. These data and information have been utilised in developing applications to provide services for specific domains. Most developments in the current resources took a stovepipe approach and are not easily interoperable with others. Therefore, heterogeneous problems exist widely within the data, information, and service resources. To fully utilise the resources and apply them to global issues, we need to share the resources in an open fashion among vast amounts of users. In other words, we need to ensure that these resources are interoperable through open, community-consensus standards among many users.

In 1994, the Open Geospatial Consortium (OGC) and the Federal Geographic Data Committee (FGDC) as well as the International Standards Organization Technical Committee 211 (ISO/TC211) were formed to address the interoperability issue. The past decade has witnessed considerable progress in at least three aspects (Goodchild *et al.* 1999, Peng and Tsou 2003): 1) Standards and Specifications were developed with the participation of different organisations led by the main three. 2) Test beds (such as the OGC Web Services Initiative) and systems (such as the Earth Science Gateway, Evans and Bambacus 2005, and the Geospatial One-Stop, Nebert 2004) were developed to implement the standards and provide interoperable solutions (Maguire and Longley 2005, Goodchild *et al.* 2007). 3) Companies built open interfaces into their software using the developed standards and governments made interoperability a request in their procurements. Using the principles of DGIP architecture, algorithms and interoperability, The Joint Center for Intelligent Spatial Computing (CISC) works with its funding agencies (such as NASA and EPA) and other partners (such as ESIP) to develop the Earth Information Exchange (EIE, http://eie.cos.gmu.edu/) for sharing heterogeneous resources. Figure 1 illustrates the architecture of the EIE: a) the data sources including the developed applications, archived data, and other information are built with interoperable interfaces, such as Web Mapping Service (WMS; de La Beaujardiere 2004), Web Feature Service (WFS; Vretanos 2002), and Web Coverage Service (WCS; Evans 2003), for them to be invoked by clients. b) The back-end application provides an integration function to ingest interoperable services and supports the server system's portlet functions, which produce web fragments to be browsed within the web browser. While only four layers including *web browser*, *portal system*, *back-end application*, and *data sources* are illustrated, the EIE was modeled on the DGIP multi-tier architecture and each of the four layers could have its own tiers.

Using the EIE, we implemented an Earth information sharing system supporting the DE and the GEOSS architecture at the levels of data, information, and services. 1) Data interoperability is achieved by using standardised data content and formats. For example, the FGDC and ISO metadata is utilised to record the uniform

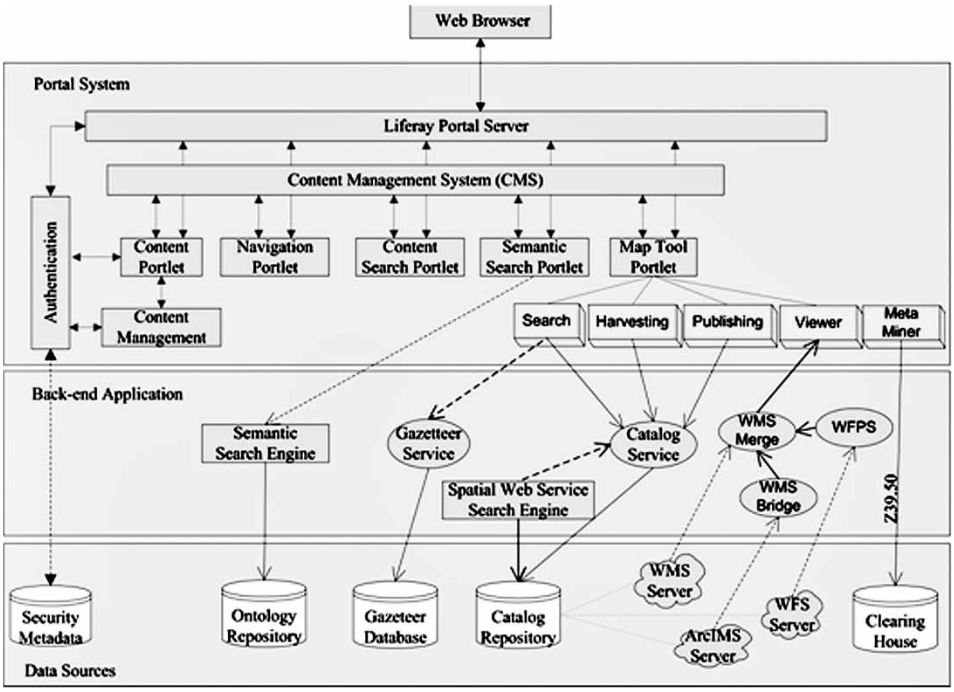

Figure 1. EIE architecture for interoperability (Li 2007 and Yang *et al.* 2007): The top layer is the graphic user interface (GUI), where end users can simply use a web browser to view and operate geospatial data and information. The second layer is the portal server layer, which manages several functional portlets (such as content portlet, navigation portlet, semantic search portlet, etc.) and at the same time provides a run-time environment for the portlets. The map tool portlet supports searching, harvesting, publishing and viewing spatial web services, which are registered in the catalog service at the third layer. In this layer, the back-end applications, which provide server-side service, are maintained. The bottom layer is the data layer, where the ontologies are used to support semantic searches, and the gazetteer database supports gazetteer query and service as well as other data are stored.

meta-information about datasets so that the data can be discovered from different catalogues. The data content are formed into de facto or certified standards, such as the Graphics Interchange (GIF) format and the Joint Photographic Experts Group (JPEG) format, and Geographical Markup Language (GML). Therefore, many resources can be served through a fast hook-up between the provider and the consumer (Figure 2(a)). 2) The information interoperability is achieved through the JSR 168/268 information standards for sharing developed portlets so that the fragments containing information can be shared within and across portlet servers (Figure 2(b), Yang *et al*. 2007). 3) The service interoperability is achieved by utilising the open geospatial standards, such as WMS, WFS, WCS, Web Gazetteer Service (WFS-G; Rob Atkinson and Jens Fitzke 2002), Web Catalogue Service (CSW; Nebert and Whiteside 2005), and other Distributed Geospatial Information Services (DGIS, Yang and Tao 2005). This service interoperability enables even the simplest browser application to leverage the capabilities of immense data repositories and processing facilities deployed across the globe (Lieberman *et al.* 2005). Service Oriented

Architecture (SOA) is adopted to register, find, and bind the DGIS to enable the interoperability among services. The SOA is implemented as harvest/publish, search, view and Meta miner as illustrated in Figure 2(c) and 2(d) (Yang *et al.* 2007), where wind services registered are found by the search and added/bound to the application.

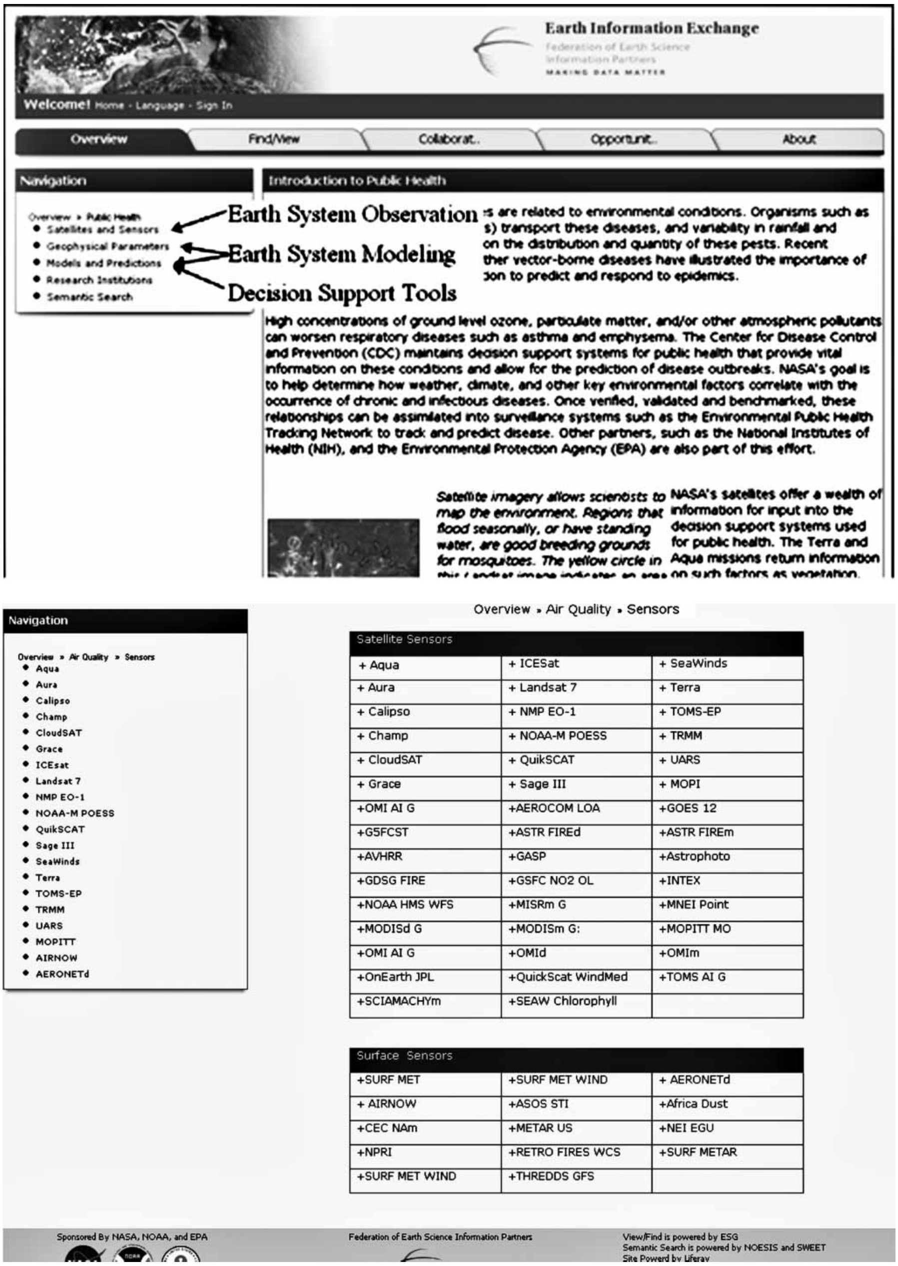

Figure 2 (*Continued*)

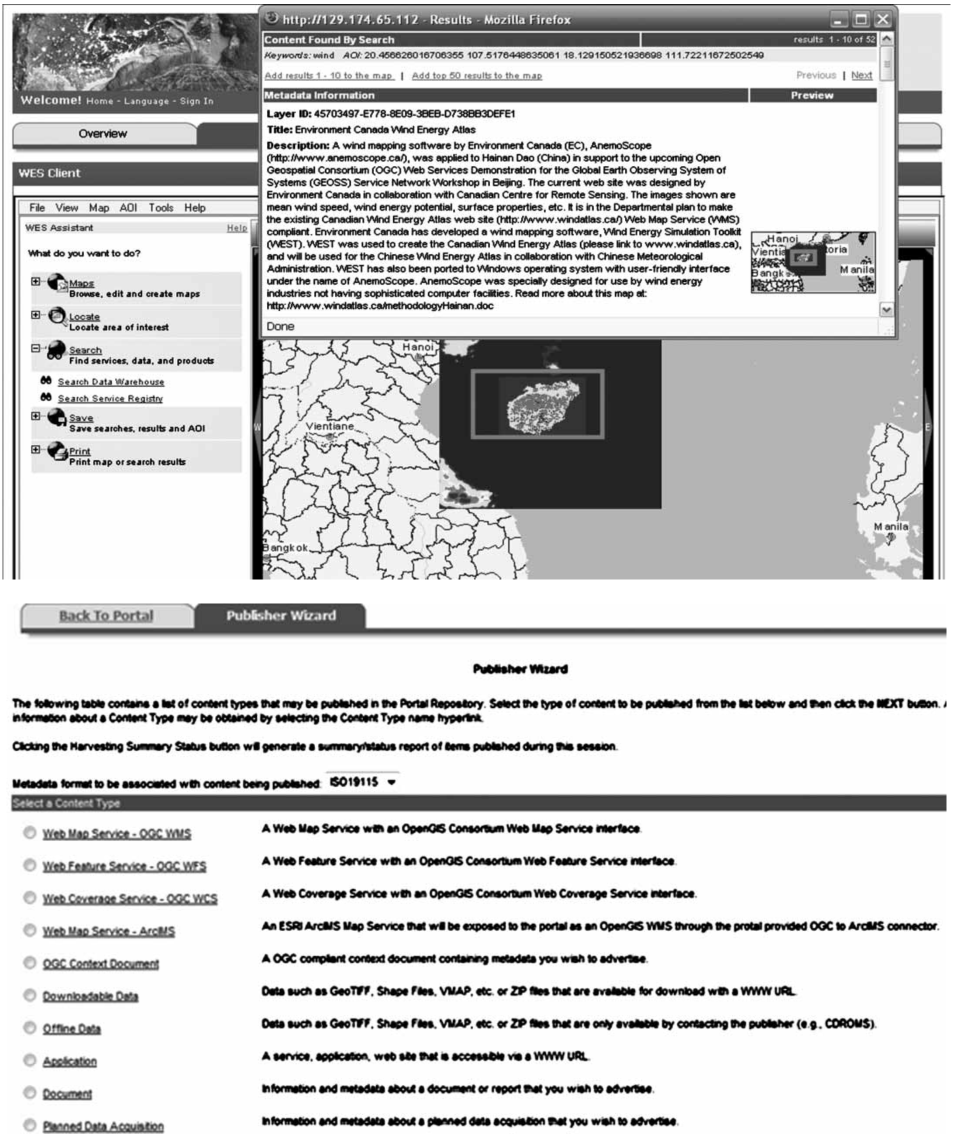

Figure 2. The sharing of data, information, and services to support the GEOSS applications and the DE. (a) The EIE supports the sharing of Earth Observation information, Earth System Models, and Decision Support Tools by providing sensors, geosphysical parameters, models & predications to mapping to the 3 levels. This figure illustrates the Public Health related observations. (b) The EIE supports the sharing of data through the catalog integration. (c) The EIE supports the sharing of service through the ESG's find/view/add functions. (d) The EIE supports publishing/registering OGC compliant Web Services (WS) through ESG's find/view/add functions.

The architecture, algorithm, and interoperability of DGIP will evolve with the development of new open standards, interoperable approaches, and computing devices facilitating our daily life for the DE.

GeoCyberinfrastructure: sharing distributed computing resources to enable geospatial applications

GeoCyberinfrastructure

Both computing power and software purchased and installed on computers are wasted when they are only partially used during their life time (Yang and Tao 2005). These wasted resources can be connected to compose a platform for serving the needs of time-consuming processing required by the DE. A new computing infrastructure providing a tested mechanism for achieving this is Cyberinfrastructure, which refers to the coordinated aggregate of software, hardware and other technologies required to support current and future discoveries in science and engineering (Berman 2007). These resources are normally connected through the internet and relevant middleware, a type of software that facilitates the sharing of resources.

GeoCyberinfrastructure refers to the DGIP high-performance computing platform aspect, where Cyberinfrastructure supports the current and future discoveries in geo-science and engineering. For example, the GEOsciences Network (GEON) (http://www.geongrid.org/) is developed using SYNSEIS (synthetic seismogram computation toolkit), GEON LiDAR Workflow (GLW) (an end-to-end system for distributing and processing the DEM generation of LiDAR / ALSM point cloud data), and other middleware to support cross-domain collaborations. The SEE-GEO project (http://edina.ac.uk/projects/seesaw/index.html) is based on the OGSA-DAI (http://www.ogsadai.org.uk/) and it will demonstrate the secure access to DGISs using open interoperability standards running on the National Grid Service. In CISC, we built a GeoCyberinfrastructure using Condor (Litzkow *et al*. 1988), Globus (Foster and Kesselman 1998), and a finer scheduler middleware under in-house development. This GeoCyberinfrastructure is designed to address the distributed geospatial processing requirements. The finer scheduler middleware is the local resource manager, which can handle the job submitted from the internet. The finer scheduler has its own file transfer mechanism, which can be used to replicate data between distributed sites and virtual organisations. It has a flexible script interface which enables users to define the workflow and file transfer process.

GeoCyberinfrastructure requirements

The current developments in the geospatial field require the GeoCyberinfrastructure to provide computing capabilities for leveraging legacy resources and solving time-consuming process. When processing large amounts of geospatial data, the geospatial algorithms, such as adaptive image processing (Zhang *et al.* 1995), data modeling and analysing, and shortest path searching (Worboys and Duckham 2005), require significant computing capability. For example, when routing in real time against a road network with real-time and/or predicted travel time for each road link, we have to simulate a road network every 3–5 min. Figure 3 shows different response times for different simulation link numbers on a computer with 3.06 Hz CPU and 2G

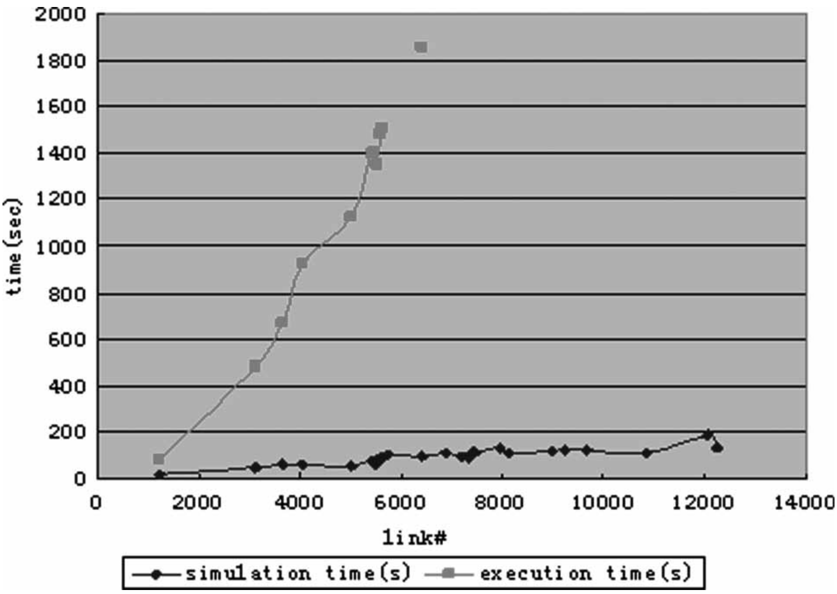

Figure 3. Execution time with different numbers of links (x axis refers to the number of links, y axis refers to time spent in seconds, simulation time is net time consumption of traffic simulation, and execution time includes data preparation).

RAM. The execution time will increase to above 10 min when the link number is over 4000, which is beyond the practical 3–5 min.

When a metropolitan area is equipped with such a routing simulation system, the number of nodes could be up to around 1 million, and the concurrent users could be in the thousands. It will take hours or days to respond to concurrent users if we still used a single computer. In other words, when we apply geospatial algorithms to practical problems, we must have significant computing capabilities.

CISC grid: A GeoCyberinfrastructure example

Observing the computing requirements and scientific and engineering needs of geospatial science, we deployed a CISC grid as a GeoCyberinfrastructure to support geospatial research and development. Figure 4 illustrates the structure of the platform and its components. A central manager is used to schedule and coordinate all resources of the platform. Condor and Globus toolkits are used as middleware to perform the scheduling and coordination tasks. Condor schedules our computing servers and desktops, while Globus manages the interfacing with remote organisations. The CISC grid platform includes 32 CPU cores, 22.238 G bytes RAM, 1.365T hard disk, and 26 G Flops computing power.

The CISC Grid platform is also connected with the SURAgrid, a virtual organisation that shares computing resources from over 20 universities within the southeastern U.S. The SURAgrid provides a uniform portal (https://gridportal.sura.org/) to handle diverse application requirements and to collect the statistics of computing resources. For example, SURAgrid currently has 1221 CPUs and 7148.6GFlops peak computing capacity, using Condor-G (Frey, Tannenbaum,

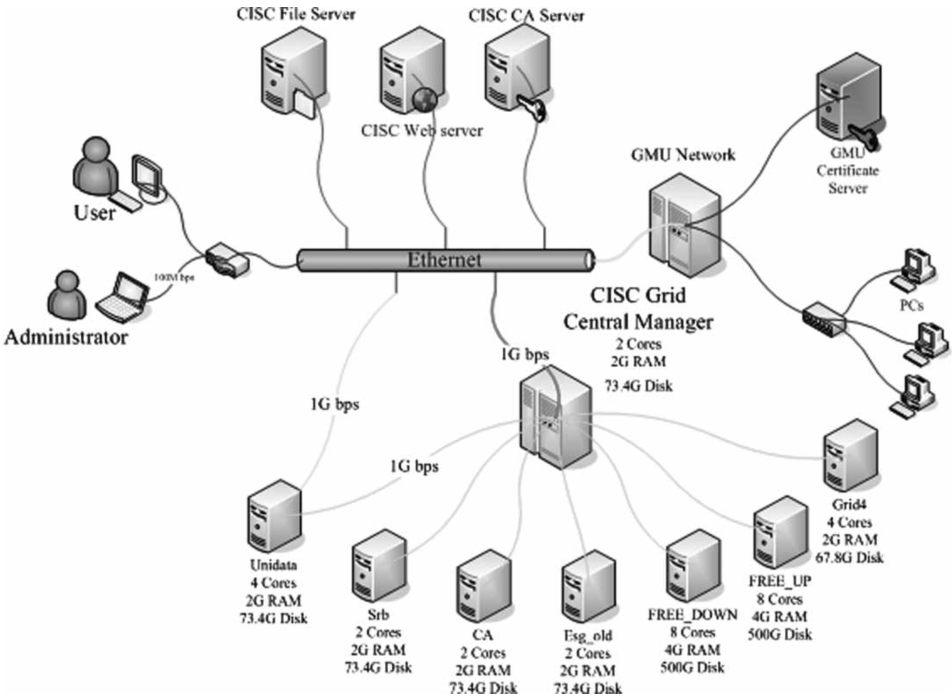

Figure 4. The CISC grid-computing environment.

Livny, Foster and Tuecke 2001), Globus, and Condor to build the SURAgrid platform.

Grid-enabling: leveraging distributed computing to enable applications

There are at least two ways for the DE to leverage the capabilities of the GeoCyberinfrastructure: a) GeoCyberinfrastructure services can be used to integrate and share distributed Earth observations, simulations, and other geospatial data, or b) the GeoCyberinfrastructure can be utilised to leverage distributed computing resources to enhance efficiency of geospatial algorithms and Earth science models. Many geospatial algorithms and Earth science models, such as Weather Research and Forecasting (WRF, http://www.wrf-model.org), are very time-consuming (hours or days) when processing large amounts of data to obtain prediction or simulation results if operating on a single computer. GeoCyberinfrastructure can be leveraged to reduce the time from days/hours to a practical level of minutes/seconds, therefore, better enabling scientific and engineering research. To leverage the GeoCyberinfrastructure to support a specific application, we need to first select a middleware to construct a grid platform, and then grid enable the application, developing an interface with the grid. Grid enabling has two approaches a) to decouple the algorithms or models into different parallel steps, or b) to decouple the study area into different parallel executable sections.

For example, a real time routing simulation will require predictions to generate travel time for future time steps. We chose MITSIMLab (a popular routing simulation platform, Yang 1997) as the simulator for the operations of integrated

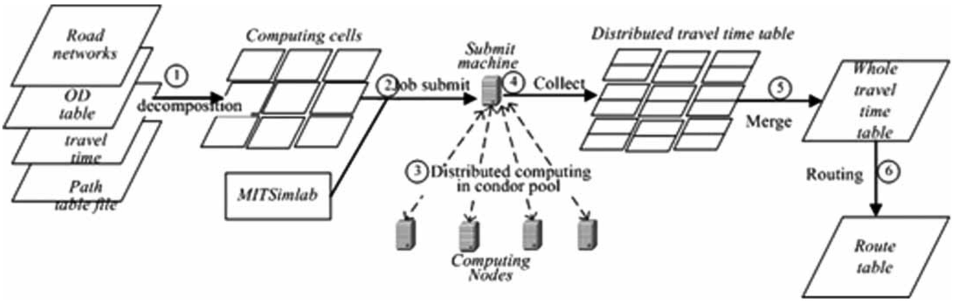

Figure 5. Work flow of grid enabled real time traffic simulation.

traffic networks. The simulation is a very time-consuming task as illustrated in Figure 3. To solve this problem, we use the second grid-enabled method and take the six steps listed in Figure 5. Condor (http://www.cs.wisc.edu/condor/) was used as the grid middleware to schedule idle servers connected by the local area network.

The workflow for grid-enabling real-time traffic simulation includes the following six steps (Cao 2007):

1. Data decomposition: all Washington, D.C. traffic data is decomposed into sections
2. Each section is submitted as a separate job to the computing platform
3. Match computing nodes and distributed computing
4. Return results separately
5. Merge the results into a complete travel time table
6. Ready for routing using complete travel time table

To give an example of the performance of the grid enabled real-time traffic simulation, part of the Washington D.C. region with over 50,000 links (Figure 6) is used as the study area. The study area is divided into 25 sections with an overlap of 1000 m for each section.

A grid with 2, 4, 8, 22 CPU cores are used to test the results as illustrated in Figure 7. As a result, the grid-enabled traffic simulation can enhance the computing efficiency by harnessing distributed computing resources.

Another example is the WRF, which is a next-generation mesoscale numerical weather prediction system designed to serve both operational forecasting and atmospheric research needs. The model is widely adopted to simulate different global and environmental issues, such as dust storms and hurricanes, for emergency response and preparedness (FCM 2007). The current software architecture of WRF allows for computational parallelism and system extensibility based on RSL, OpenMP, and MPICH. Figure 8 shows that with more CPUs involved, the simulation time can be significantly reduced within a certain level (up to 8 cores in this case).

To leverage the power and understand the limit of the GeoCyberinfrastructure fully, we need to develop geospatial middleware that integrate and process geospatial data (Yang *et al.* 2004). These middleware should have the capability to integrate Earth Observations, Earth system models, decision support tools, and other geospatial resources.

Figure 6. The study area is located in southwestern Washington DC and is divided into a 5X5 section matrix.

Semantic search: sharing knowledge for discovering geospatial resources.

Supporting the DE requires sharing geospatial resources in an intelligent manner by leveraging the knowledge gained to help us locate relevant information in a timely manner. Although popular search engines, such as Google and Yahoo, provide very convenient information retrieval methods for discovering information, they still need to be improved (Guha *et al.* 2001), especially for discovering Earth science information, where a large number of professional terms and term relationships are accumulated and can be used to help with the information discovery. Take *Air Quality* as an example; if we try to search for *air components that affect air quality*, traditional keyword-based search engines can only return documents or contents containing those keywords, while the real results we need are *Ozone*, *Ammonia* and *Nitrogen Oxides*. Therefore, utilising the terms and their relationships, i.e. ontology (Boriana 2007), to accurately discover expected content from distributed geospatial information resources becomes a DGIP challenge. This challenge is also called Semantic Search (Alesso and Smith 2005), a popular application of Semantic Web (Berners-Lee 2001), aiming to conduct searches through the meanings of words instead of the spellings of words. However, two main problems need to be addressed in semantic searches: 1) how to explicitly represent domain knowledge, and 2) how to conduct semantic inference based on the knowledge?

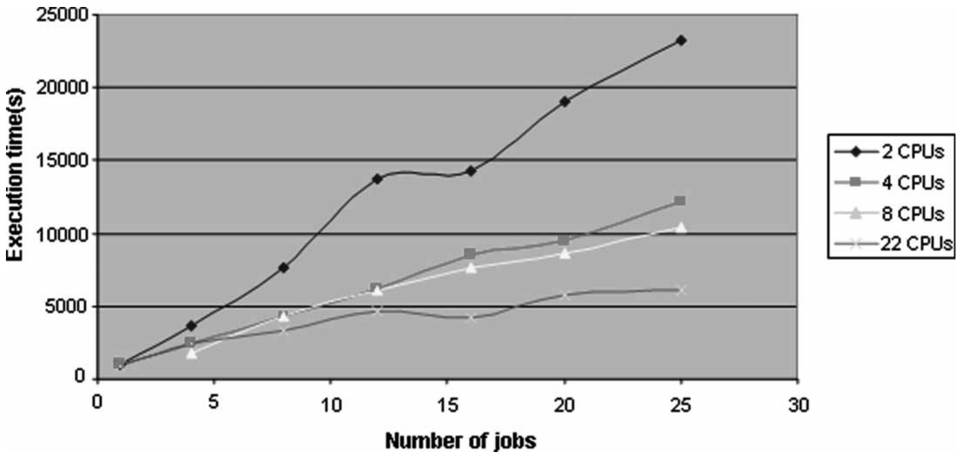

Figure 7. The grid platform effectively improves the performance of traffic simulations.

A knowledge representation system specifies how to represent concepts (Heflin 2001) to make the data machine understandable for searching. It relies on Ontology to represent concepts and interrelationships of the concepts. Ontology refers to vocabulary based on taxonomy and domain knowledge, and the semantic relationships such as subsumption, combination, and partition on the vocabulary. Ontology can be represented as a hierarchical structure in an object-oriented or thesaurus manner (such as NASA 2007). Each concept in the structure is a Class with a parent and properties inherited from its parent(s). The machine understandable ontology can be formed through three steps (Figure 9): a) XML and XML Schema (http://www.w3.org) can help formalise the concept, b) Resource Description Framework (RDF) and RDF Schema (http://www.w3.org/TR/rdf-schema/) can model the objects, and c) the OWL (Web Ontology Language, http://www.w3.org/TR/owl-features/) is used to capture more vocabulary and constraints for supporting the semantic search.

Semantic Web for Earth Environmental Terminology (SWEET, Raskin 2005) is one of the most populated Ontologies within Earth science. We adopted SWEET to

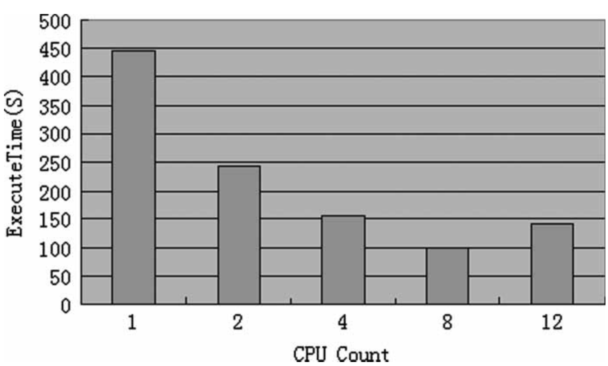

Figure 8. Execution time of parallel WRF application.

Figure 9. Ontology Language Evolutions.

describe the *Air* related ontology (different circles in Figure 10) and interrelationships (arrows in Figure 10) of the terms are built to support semantic inference.

The semantic inference is supported by Description Logics (DL), which contains a set of instructions in a knowledge representation language that can be used to represent the knowledge of an application domain in a structured and formal manner to be executed by a computer (Baader *et al.,* 2005). There are several inference engines that support DL reasoning, such as Racer (http://www.racer-systems.com/products/racerpro/index.phtml), FaCT (http://www.cs.man.ac.uk/FaCT) and Pellet (http://pellet.owldl.com). The reasoning is based on domain ontology. For example, to answer the question of *what air components affect air quality*, the components (*Ozone*, *Ammonia*, *Nitrogen Oxides*) that affect air quality are connected with the concept 'air' by properties *QualityAffectedby* (Figure 10). Once the query is submitted, not only will the answer to a certain question be given, other *Air* related processes or phenomena will also be provided as *refine search* choices (Figure 12 (b)). Therefore, a user's query can be extended and the meaning of the query can be well understood and executed by computers (Figure 12).

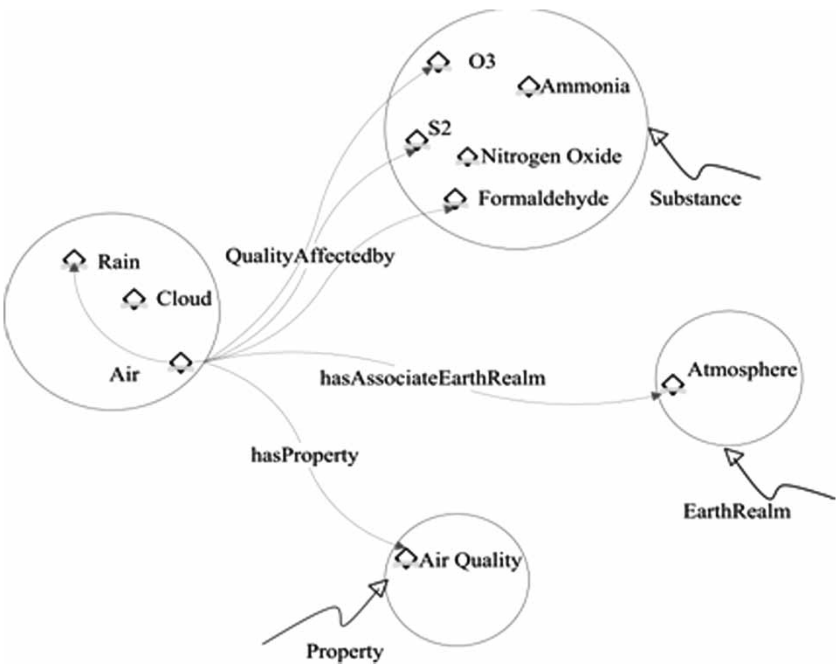

Figure 10. Air ontology (Li 2007) example: Air related terms are divided into four facets (Phenomena, Property, Substance, EarthRealm) shown in circles. The terms in different facets are connected by properties, such as QualityAffected by, has Associated Earth Ealm, and has Property, depicted by arrows.

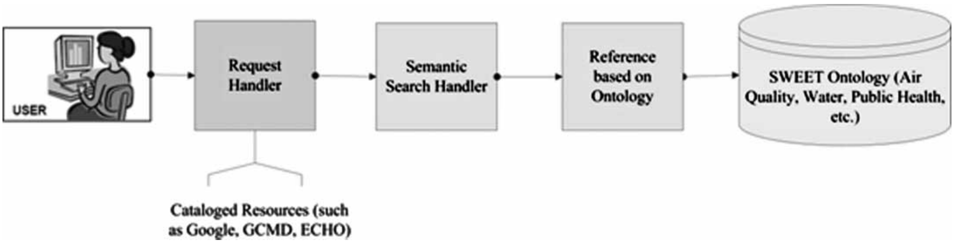

Figure 11. Workflow of a semantic search engine (modified from Ramachandran 2005).

Semantic search can be conducted based on the ontology and the reference engine as illustrated in Figure 11: 1) the user's requests in the format of either keyword or natural language description is issued to *Request Handler*. 2) The *Request Handler* will search *Cataloged Resources* and invoke *Semantic Search Handler* if semantic search is enabled (Figure 12 (a)). 3) The *Semantic Search Handler* will call *Ontology Reference* to reason based on ontology, which is stored in a database supported by SWEET with national applications populated. 4) The reasoning results (Figure 12(b)) will be returned to the *Request Handler* to refine the search results (Figure 12(c)).

Because of the complexity of the Earth system and the human understanding of it, capturing all the knowledge and utilising the knowledge in intelligent reasoning for information discovery and decision making will 1) be grand challenges, 2) require

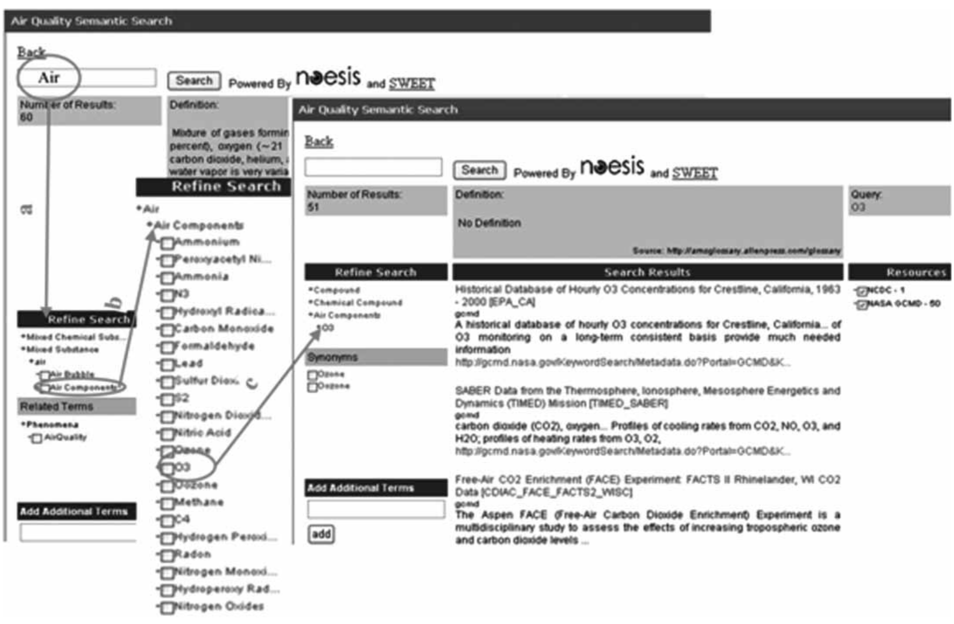

Figure 12. Semantic inference for the refine search example. (a) Searching Air in the semantic search engine will bring up a definition of Air, geospatial resources found within catalogs and ontology of Air for refining search. (b) The Air and Air-Components ontology can be provided and used to refine search. (c) The search for ozone will bring up ozone related and Air- and Air-Components related geospatial resources as Search Results.

the collaboration among experts from different science, technology, and management domains, and 3) be a long-term advancing process aligned with the progress of our understanding of the planet Earth.

Conclusion and Discussions

As a term coined to capture the recent developments in utilising computing and networking infrastructures to support the processing of distributed geospatial resources, DGIP is itself a brand new field and will evolve by itself with relevant information technology advancements and the implementation of the DE. This paper reports our research on utilising different aspects of the DGIP 1) to implement the vision of the DE of sharing data, information, services, knowledge, and computing resources, 2) to provide examples/applications of importance to the societal benefit areas (GEO 2005), and 3) to illustrate how the DGIP and relevant research fields can contribute to, enable, and advance the DE.

Besides the applications discussed, DGIP can also help facilitate addressing other global issues, like environmental issues. For example, as an incredibly complex ecosystem that includes important habitats and food webs, the Chesapeake Bay (the largest estuary in the US, http://www.chesapeakebay.net/) has been polluted for the past several decades. An effort is trying to restore the bay by mandating the reduction of pollution from each source. A Chesapeake watershed model is developed to simulate the effectiveness of optimal reduction values and, therefore, assist in the decisions on nutrient and sediment reduction to protect the environment. However, the input data of the model include 308 land segments and 25 land use types (which means 7700 independent land simulations), 930 rivers, and 45000 land use/river connections (Shenk 2006). All of these components have their own input and reside at different locations. Therefore, it is very computing intensive and requires one week to do one simulation, which is beyond the ideal result of minutes or hours. Therefore, the grand challenge posed to DGIP for both 1) integrating the heterogeneous resources in a timely fashion and 2) grid enabling the simulation, will not only contribute to the DE with the challenge but will also benefit the users of the Bay.

In this regard, all five aspects of DGIP need to be addressed to respond to the requirements for supporting the simulation and building the DE: 1) The architecture and algorithms need to be advanced when new applications and devices are integrated (Yang and Tao 2005); 2) A thorough interoperability study needs to investigate the implementation of a fully interoperable GEOSS and DE environment; 3) More middleware and scheduling algorithms need to be developed to share the widely available computing resources for the computing and heterogeneous challenges; 4) The knowledge gained in the past needs to be sorted and integrated for training the next generation of new users of a specific domain (Zhou *et al*. 2002), such as Air Quality; and 5) best practices need to be developed and put into operation to benefit the users and to implement the vision of the DE.

Besides these technical challenges and issues for building the DE, we also face the challenges of policy, management, workforce development, and politics, such as the security and privacy issues popularised by Google Earth when people began to find that too detailed information is available to the global community through the internet (Mclaughlin 2005).

Acknowledgements

The reported research is supported by a Chinese 973 project (2006CB701306), a NASA Geosciences Interoperability project (NNX07AD99G), and FGDC 2005 CAP award (05HQAG0115). The anonymous reviewers' and the editor's comments are greatly appreciated.

Notes on contributors

Dr Chaowei Yang is the director of the Joint Center for Intelligent Spatial Computing at George Mason University. He is the PI for over 10 funded projects focusing on the five streams of Distributed Geospatial Information Processing. He has published over 40 peer reviewed articles in journals and conference proceedings.

Ms Wenwen Li is a graduate student and a researcher in the Joint Center for Intelligent Spatial Computing, College of Science at George Mason University. Her research interests include geospatial information retrieval and processing, and issues relating to ontology and semantic searches. She has published relevant research studies in conference proceedings and academic journals.

Dr Jibo Xie is a postdoctoral researcher in the Joint Center for Intelligent Spatial Computing (CISC), George Mason University. His research interests include parallelising and grid enabling computing-intensive geospatial applications, such as digital terrain analysis, real-time traffic simulation, and weather forecasting models.

Mr Bin Zhou is a visiting student in the Joint Center for Intelligent Spatial Computing, College of Science at George Mason University, and a PhD student of the department of electronic engineering, Tsinghua University, China. His main research interests are grid computing, scheduling techniques and parallel algorithms for high performance computing and high speed signal processing.

References

Alesso, H.P. and Smith, C.F., 2005. *Developing semantic web services.* Natick, MA: A K Peters.

Atkinson, R. and Fitzke, J., 2002. Web Gazetteer Service, OGC Implementation Specification [online]. Available from: https://portal.opengeospatial.org/files/?artifact_id =7175 (Accessed 9 August, 2007).

Baader, F., *et al*. 2005. Integrating description logics and action formalisms for reasoning about web services. Technical report LTCS-05-02 [online]. Available from: http://lat.inf.tu-dresden.de/research/reports.html (Accessed 28 February 2008).

Berners-Lee, T., Hendler, J. and Lassila, O., 2001. The semantic web. *Scientific American*, 2001, May, 29–37.

Berman, F., 2007. SBE/CISE Workshop on Cyberinfrastructure for the Social Sciences [online]. Available from: http://vis.sdsc.edu/sbe/SBE-CISE_Workshop_Intro.pdf (Accessed 9August 2007).

Birk, R.M., *et al*., 2006. NASA's Applied sciences program: transforming research results into operational success. *Earth Imaging Journal*, 3 (3), 18–23.

Boriana, D., 2007. Thesaurus and domain ontology of geoinformatics, *Transactions in GIS* 11 (4), 637–651.

Cao, Y., 2007. *Transportation routing with real-time events supported by grid computing*. Thesis (PhD). George Mason University. 104 p.

Chen, J. and Wu, L., 2003. *Geospatial fundamental framework of digital China*. Beijing: China Science Press.

Cheng, J.C., *et al*., 2000. *Introduction to Digital Earth*. Beijing: China Science Press.

Coleman D.J., 1999. Geographical information systems in a networked environemnt. In Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds). *Geographic information systems*. New York: Wiley.

De La Beaujardiere, J., 2004. Web map service, Ver.1.3, OGC implementation specification [online]. Available from: http://portal.opengeospatial.org/files/?artifact_id = 14416, OGC (Accessed 9 August 2007).

Evans, J., 2003. Web coverage service, Ver.1.0, OGC implementation specification [online]. Available from: https://portal.opengeospatial.org/files/?artifact_id = 18153, (Accessed 9 August 2007).

Evans, J. and Bambacus, M., 2005. NASA's Earth-Sun system gateway: an open standards-based portal to geospatial data and services. *Proceedings of IEEE IGARSS,* 25–29 July 2005, Seoul, Korea, pp. 4228–4231.

FCM, 2007, *Interagency Strategic Research Plan for Tropical Cyclones: The Way Ahead* [online]. Available from: http://www.ofcm.gov/p36-isrtc/ (Accessed 13August 2007).

Foote K. and Kirvan T., 1998. WebGIS, *NCGIA Core Curriculum in GIScience* [online]. Available from http://www.ncgia.ucsb.edu/giscc/units/u133/u133_f.html (Accessed 9 August 2007).

Foster, I. and Kesselman, C., 1998. The Globus project: a status report. *Proceedings of the Seventh Heterogeneous Computing Workshop.* 30–31 March. Washington, DC: IEEE Computer Society, 4.

Frey, J., *et al*., Condor-G: a computation management agent for multi-institutional grids 2001. In *Proceeding of the 10th IEEE International Symposium on High Performance Distributed Computing*.

GEO, 2005. *The Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan* [online]. Available from: http://earthobservations.org/docs/10-Year%20Implementation%20Plan.pdf (Accessed 8 August 2007).

Gore, A., 1998. *The Digital Earth: Understanding our planet in the 21st Century, Given at the California Science Center*. Los Angeles, California, 31 January 1998.

Goodchild, M. F., Fu, P. and Rich, P., 2007. Sharing geographic information: an assessment of the Geospatial One-Stop. *Annals of the Association of American Geographers,* 97 (2), 249–265.

Goodchild, M. F., *et al*. (eds), 1999. *Interoperating geographic information systems*. Boston: Kluwer Academic.

Guha, R., Mccool, R. and Miller, E., 2001. Semantic search. *Proceedings of the 12th international conference on World Wide Web*, Budapest, Hungary, 20–24 May 2003.

Guo, H.D., 1999. *Earth observation and Digital Earth*. Beijing: Chinese Science Press.

Guo, H.D. and Yang, C. J., 1999. *Proceedings of the International Symposium on Digital Earth (ISDE'99)*, Beijing (China): Chinese Science Press.

Heflin, J., 2001. *Towards the semantic web: knowledge representation in a dynamic, distributed environment*. Thesis. (PhD), University of Maryland, College Park.

Huffman, G., *et al*., 2007. The TRMM multisatellite precipitation analysis (TMPA): quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, 8 (1), 38–55.

IT Professional, 2005, News Briefs: Hurricane Katrina reveals strengthens of emerging technologies. *IT Professional*, 7 (5), 5.

Joseph, S. and Shmuel, M., 2007. *Information technology-transmission, processing, and storage.* New York: Springer.

Li, D., 2001. Function of digital province and digital city in territorial planning and urban construction. *Engineering Journal of Wuhan University*, 34 (6), 47–59.

Li, Q. and Lin, S., 2001. Research on digital city framework architecture. *Proceedings of ICII 2001* Beijing, China, Vol. 1, pp. 30–36.

Li, Q., Yang, C. and Yi, S., 1999. Architecture of Digital Earth. *Journal of Remote Sensing*, 3 (4), 281–287.

Li, W., 2007. *Interoperability based Spatial Information Retrieval and Ontology based Semantic Search*. Thesis (PhD). Institute of Remote Sensing Applications, Chinese Academy of Sciences, 2007.

Lieberman, J., Pehle, T. And Dean, M., 2005. Semantic evolution of geospatial web services [online]. Available from: http://www.w3.org/2005/04/FSWS/Submissions/48/GSWS_Position_Paper.html (Accessed 13 August 2007).

Litzkow, M.J., Livny, M. and Mutka, M.W., 1988. Condor-a hunter of idle workstations. *Proceeding of 8th International Conference on Distributed Computing Systems*. San Jose, USA: IEEE-CS Press, 104–111.

Maguire, D. J., and Longley, P.A., 2005. The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment and Urban Systems,* 29 (1), 3–14.

Mclaughlin, L., 2005. From AWK to Google: Peter Weinberger Talks Search. *IEEE Security and Privacy*, 3 (5), 11–13.

Mörner, 2004, Estimating future sea level changes from past records. *Global Planet. Change,* 40 (1–2), 49–54.

NASA, 2007. NASA Thesaurus Project [online]. Available from: http://www.sti.nasa.gov/98Thesaurus/vol1.pdf (Accessed 20 September 2007).

Nebert, D., (ed.) 2004. *The SDI Cookbook*, version 2.0. Association of Global Spatial Data Infrastructure.

Nebert, D., and Whiteside, A., 2005. *Catalog Services, Version 2, OGC Implementation Specification* [online]. Available from http://portal.opengeospatial.org/files/?artifact_id = 20555, OGC (Accessed 9 August 2007).

Patrice, F., 2007. Carey-Libbrecht L. (translation) 2007. *The Internet imaginaire*. Cambridge, MA: MIT Press.

Peng Z.R., and Tsou M.H., 2003. *Internet GIS: Distributed Geographic Information Services for the Internet and Wireless Networks*. Hoboken, New Jersey: John Wiley.

Ramachandran, R., et al., 2005. *NOESIS: Ontology-based Semantic Search Tool for Atmospheric Science* [online]. Available from: *http://ams.confex.com/ams/pdfpapers/102272.pdf*. (Accessed 13 August 2007).

Raskin, R., 2005. *Semantic Web for Earth and Environmental Terminology (SWEET)* [online]. Available from: http://archive.esipfed.org/tech_center/presentations/SWEET.ppt (Accessed 13 August 2007).

Rae-Dupree, J., 2006. Dash navigation – real-time GPS driving. *Business Journal*, 2006, [online]. Available from: http://sanjose.bizjournals.com/sanjose/stories/2006/11/20/focus15.html (Accessed 8 August 2007).

Shenk, G., 2006. Development of the next generation of Chesapeake Bay Program Watershed Model. EPA/Chesapeake Bay Program presentation.

Vretanos, P.A., 2002. Web feature service, Version 1.0 [online]. Available from: URL: http://portal.opengeospatial.org/files/?artifact_id =8339, OGC (Accessed 9 August 2007).

Worboys, M. and Duckham, M. 2005. *GIS A Computing Perspective*. 2nd ed. Boca Raton, FL: CRC Press.

Xu, C.Z., Jin, H., and Srimani, P.K., 2003. Special issue on scalable web services and architecture. *Journal of Parallel and Distributed Computing*, 63 (10), 913–915.

Yang C., 2000. Theory, techniques and implementation methods of WebGIS. Thesis. (PhD). Peking University. (In Chinese). Unpublished.

Yang, C., 2007. Earth information exchange- sharing earth information through interoperable and cyber infrastructure. *The 5th International Symposium on Digital Earth*, San Francisco, CA, 4–9 June 2007.

Yang, C., *et al.*, 2004. GridGIS: A next generation GIS. *CITSA 2004*, 21–25 July, Orlando, FL, pp. 22–27.

Yang, C., and Raskin, R., 2007, Sessions of Distributed Geospatial Information Processing (DGIP). In *AAG 2007 annual meeting* [online]. Available from: http://www.cisc.gmu.edu/meetings/aag07.htm (Accessed 8 August 2007).

Yang C. and Tao V., 2005. Distributed geospatial information service. In S. Rana and J. Sharma (eds). *Frontiers of geographic information technology.* Chap. 5. London: Springer, pp. 113–130.

Yang, C., *et al.*, 2007. The emerging concepts and applications of the spatial web portal. *PE&RS*, 73 (6), 691–698.

Yang, Q., 1997. *A simulation laboratory for evaluation of dynamic traffic management system*. Thesis. (PhD). Massachusetts Institute of Technology.

Zhang,T., Lu W.X. and Peng J.X., 1995. Adaptive image matching via spatial varying gray-level correction. *IEEE Transactions on Communications*, Vol. 43, pp. 1970–1981.

Zhou, L., Booker, Q.E., and Zhang, D., 2002. ROD-toward rapid ontology development for underdeveloped domains. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, pp. 957–965.