JOURNAL OF PURE AND APPLIED MICROBIOLOGY, April 2013. Vol. 7(Spl. Edn.), p. 159-166

The Self-organization of Distributed Heterogeneous Spatial Data Sources in Grid-enabled Spatial Data Infrastructure

Li-Xia Guo[1-2], Guo-Qing Li[1] and Yun-xuan Yan[1]

[1]Centre for Earth Observation and Digital Earth, Chinese Academy of Sciences, Beijing, China.
[2]Graduated University of CAS, Beijing, China.

(Received: 03 March 2013; accepted: 14 April 2013)

Spatial data infrastructure (SDI) is a framework of spatial data, metadata, users and tools that are interactively connected in order to use spatial data in an efficient and flexible way. As the development of network information technology and impact of global change, much attention is paid on SDI. However, SDIs have limited capability in dealing with ever-increasing amounts of high resolution data which can be distributed globally and heterogeneous extremely. One of the most important applications in grid is focusing on the collection of data resource. With the availability of large storage and computing facilities offered by Grid infrastructures, new opportunities are emerging. In the context of gird, the paper proposes the concept of "Data Grid Node" to shield the complexity process of heterogeneous spatial data source gridification, and introduces ways of automatic data grid entity building which can be used for the self-organization of distributed heterogeneous spatial data sources. The project "Data sharing infrastructure of global change", funded by "National Research and Development Program of China (973 Program 2009CB723906)", will target these issues.

Key words: Spatial data infrastructure, Grid, Data grid node, Self-organization of spatial data sources.

Global change research is an integrated subject referring to the atmosphere, hydrosphere, biosphere, human circle and outer space, whose research data can be collected from earth observation, field measurements, scientific simulation, base map and so on. As the data sharing infrastructure for global change it is developed for 973 program "Mechanisms and Methods for Space-based Observations of Global Change-sensitive Factors". A large number of scientific data related to global change will be collected and shared in the program. However, as the data is owned by 6 sub-programs separately, the data can be distributed globally and heterogeneous extremely which is not conducive to the exchange and sharing of data. Data sharing infrastructure for global change should be developed to solve the above problems including the collection, organization, archive and publish of scientific information involving global change, making the scientific information within the program shared as convenient as possible.

Spatial data infrastructure (SDI) is a framework of spatial data, metadata, users and tools that are interactively connected in order to use spatial data in an efficient and flexible way[1]. As the development of network information technology and impact of global change, much attention is paid on SDI. However, SDIs have limited capability in dealing with ever-increasing amounts of high resolution data which can be distributed globally and heterogeneous extremely. With the

* To whom all correspondence should be addressed. E-mail: lxguo@ceode.ac.cn; Phone: 15801429520; Address: No.9 Dengzhuang South Road, Haidian District, Beijing 100094, China.

availability of large storage and computing facilities offered by Grid infrastructures, new opportunities are emerging[2].

Grid computing is a term referring to the combination of computer resources from multiple administrative domains to reach a common goal. The grid can be thought of as a distributed system with non-interactive workloads that involve a large number of files. What distinguishes grid computing from conventional high performance computing systems such as cluster computing is that grids tend to be more loosely coupled, heterogeneous, and geographically dispersed. Although a grid can be dedicated to a specialized application, it is more common that a single grid will be used for a variety of different purposes[3,4]. Data resource can be shared as a kind of grid resource and one of the most important applications in grid is focusing on the collection of data resource. As the extention of grid, Spatial Information Grid (SIG) should be also in charge of the collection of the spatial data resource. As we known, the spatial data is heterogeneous because of the mutiple expression on the semantic, spatial-temporal, scale and sensor platform[5]. The heterogeneity brings great trouble not only for the assembling of grid level service from spatial data resources, but also for the provider to publish their data through grid platform.

However, publish of grid resource is a complex process requiring the involvement of professional and technical personnel. When organization out of the grid wants to join it and publish its own data resource, it is both time-consuming and extremely complicated. Consequently it is essential to find ways on how to quickly and automatically build a grid resource node as well as the self-organization of spatial data sources.

Self-organization of spatial data sources

Spatial data source classification model

Generally speaking, heterogeneous spatial data sources always have their own unique data organization and service model. We need to classify them into different types from grid and web service architecture point of view, making each type of data source with similar service logic and hierarchical structure internally, so that we can find suitable level for web service plug-in.

Based on the spatial data sources collected for global change, first we need to make sure that whether data access service is provided or not. For those without service additional work is needed, including investigation and study on data format, data organization hierarchy, metadata expression and so on; for those with data access service already, we should continue to judge whether it is OGC service or not, because further analysis on service interface and standard is needed for service not belonging to OGC. According to the classification standards, we can define the data source type and induce its logic and interface characteristic in order to establish a relatively uniform web service framework and model. The classification model is shown as Fig.1.

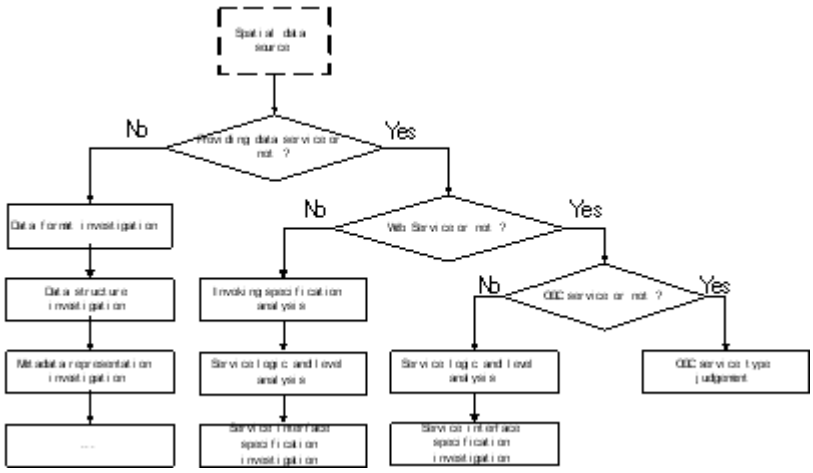

Fig. 1. Classification flow chart of spatial data sources

Data grid entity node

Discovering and evaluating data through their metadata is the first functionality that users can expect from an SDI. Once they know the existence of a specific data set, users want to have the possibility to access it either by direct download or through web services.

Many of the decisions that organizations need to make depend on good, consistent, and readily accessible geospatial data to support decision making processes[6, 7, 8].

Spatial data grid entity node is a kind of service clustering which is devoted to data query and access in SIG. On the one hand, it can provide grid users and other grid entity nodes with numerous data resource, on the other hand it is simple for data owners to construct their own node.

With the concept of grid entity node, it can manage not only internal service resource in SIG, but also the external standard and non-standard web service, registered in UDDI center as grid resource, achieving interoperability among heterogeneous platforms and services. There is standard service description in SIG, service in every node corresponding to a kind of profile[9]. Functions of data grid entity node can be analyzed from both internal aspect and external aspect (Fig.2).

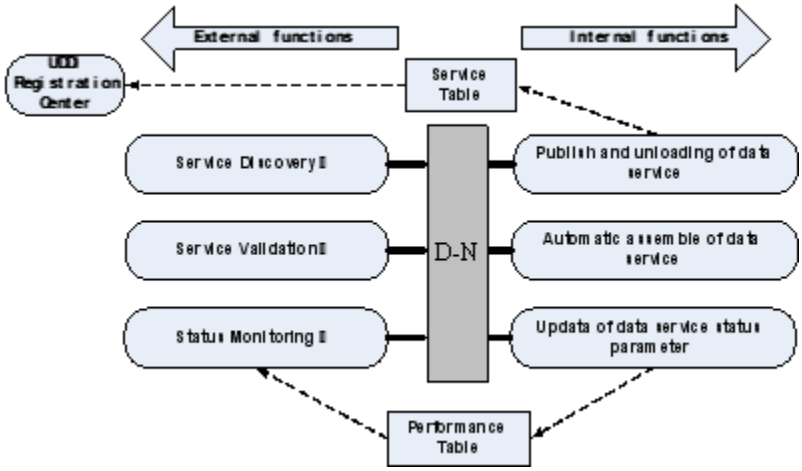

Fig. 2. Internal and external functions of spatial data grid node

Framework of spatial data grid node is show in Fig.3. Data Grid-Node (DGN) is the basic component and function as resource assembler. WPS is used as the middleware container of DGN and WPS extensions are used for effective management and automatic resource registration. Querying and accessing are packaged as internal function. In addition, node engine is used to encapsulate target datasets interoperation interfaces into uniform grid resources. Resource adapter can merge various kinds of interfaces provided by target datasets into a uniform one.

Based on function analysis and framework design above, Fig.4 interprets the operational process of spatial data service. Users can access data grid resource through spatial data service in grid environment. Spatial data service mainly

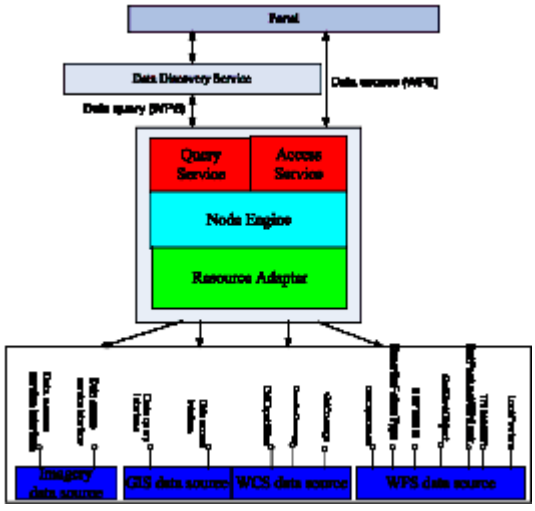

Fig. 3. Framework of Spatial data grid node

translates data query and access request into identified formats, such as SQL sentence or data source service access request and returns the unified response to requester.

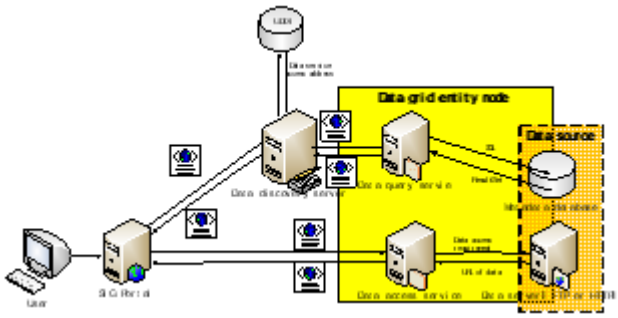

Fig. 4. Operational process of spatial data service

Self-organization model and algorithm of heterogeneous spatial data source

Organization structures of different spatial data sources may vary widely. However, it is supposed to look the same out of grid. Therefore, it is necessary to make use of some middleware to hide the details of these differences between data sources, so as to form relatively uniform spatial data self-organization node outside the grid[10].

Based on the classification model described above, adapters are developed to epitomize what these spatial data source have in common and analyze what category each spatial data source should belong to. Then metadata automatic extraction tool is designed applicable to each category of spatial data source. Next we can integrate these tools as metadata extraction tool set which is capable of category recognition with unified interface. In the extraction process, shared information of different spatial data sources will be expressed consistently and put into common repository according to category information. At the same time the special information will be saved in unique repository. Unified resource expression service will take metadata in common repository as basic information and the ones in unique repository as extended information, providing WPS based self-organization engine with resource expression service by uniform expression method. Besides information extraction, resource adaptor can also add or modify service, as well as providing relatively consistent basic data service to WPS based self-organization engine. Fig.5 shows the internal structure of resource adapter.

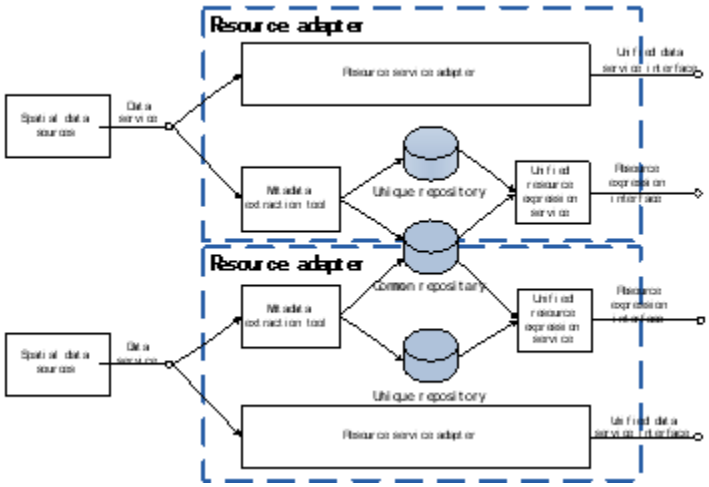

Fig. 5. Internal structure of resource adapter

High performance, scalable, highly available service interface

There are two kinds of service interfaces for grid users. One is traditional query-access mode with query interface and access interface which is the basic service mode within self-organization node, meanwhile providing interface to traditional remote sensing information platform. For query interface, standard query expression and unified metadata information are supposed to take into consideration. Mandatory metadata information includes data formats, acquisition time, geographic range, resolution and so on. In addition, metadata information of raster data also includes data collection information such as spacecraft, sensors, product level and so on. For access interface, parameter expression of on-demand data service is supposed to take into consideration. The other is OGC based service mode, including WCS mode and WFS mode, providing coverage service and feature service respectively. It can provide interoperability to users outside the node, even the grid. Based on research of OGC-WCS and OGC-WFS services, data service interface and resource expression interface are transformed into interfaces conforming OGC standards[11].

Project use-case and applications

According to theory proposed above, we developed several applications, mainly 863 project named "Grid-based Distributed Heterogeneous Spatial Data on-Demand Service Technology". Based on service-oriented grid structure, distributed spatial data sources can be collaborative and controlled universally in the project, which have heterogeneous standards, system, work pattern and data content, so that they can make no difference in term of grid service. It can bring huge data advantage as the result of heterogeneity and distribution characteristics for on-demand customer service system, rather than the huge complexity.

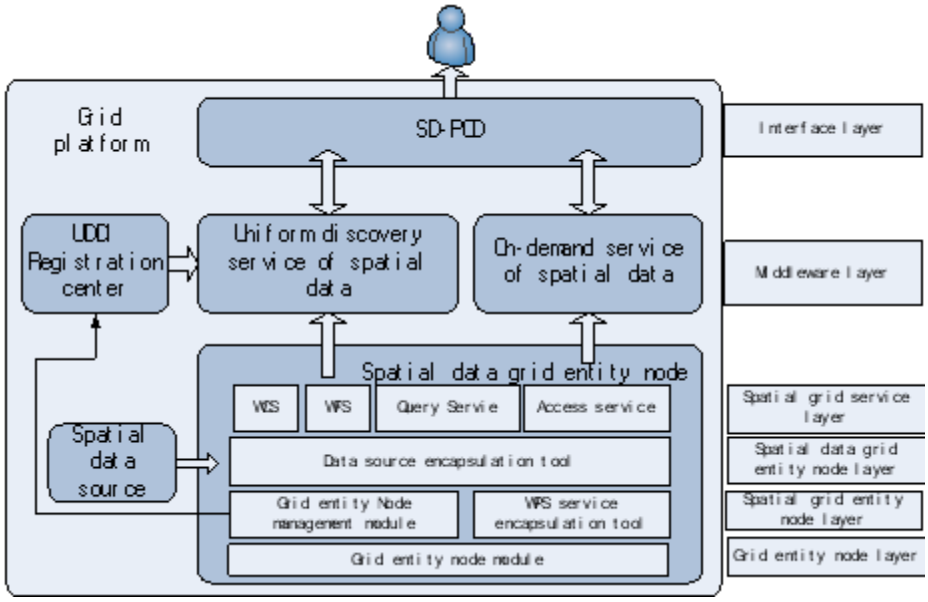

Fig. 6. System Architecture

System architecture

Fig.6 shows system architecture. In spatial data grid entity node, heterogeneous spatial data source is encapsulated into homogeneous one, which can be accessed by top spatial data service. In addition, spatial data gird entity node management module and WPS service packaging tool share the service management responsibility of spatial data grid entity node. Management module is responsible for registering services and resources in the node into UDDI, therefore spatial data uniform discovery service can find the required spatial data grid entity node through UDDI and invoke corresponding query service, when user query data on SD-POD. At the same time, spatial data uniform discovery service is capable of

consistent metadata expression of different data, which is convenient for users. After accessing data, spatial data on-demand service can process data according to users' need.

Achievements show

As data are collected by 6 sub-programs in the program, the data formats and organization methods can be different extremely. The diversity of spatial data, as well as the distribution of data agencies, brought a great obstacle for data sharing and collaboration. However, from a practical point of view, heterogeneity and distribution is the inevitable trend in the field of global change research, and has its positive meaning. Dynamic data allows the expression more comprehensive and objective to the real world. The distribution of data sources will enable the establishment and maintenance more professional and precision. The heterogeneity can maximize the protection of the integrity of the various types, and facilitate the provision of specific applications.

Fig. 7. Portal of data sharing infrastructure for global change

The figure above is data sharing infrastructure for global change. It provides portal for data query and access, involving various kinds of spatial data related to global change, mainly base map data, earth observation data, in-situ data, scientific model and simulation data. These data are collected from 6 sub programs covering 3 test areas. It provides an easy way for both scientists and general users to query and access the data they need. Additionally, data owners can publish their data conveniently by the portal as quickly as possible. Consequently, it plays an important role in the program and bridge the gap between researchers and users, as well as professional grid technicians and scientists. Fig.8 shows data query and access client. Fig.8 (a) is portal for data search, users can select data source, geographic bonding box, acquisition time, spacecraft, sensor, resolution and so on. Fig.8 (b), (c) and (d) are search results, respectively ESA-ASAR, DEM and MODIS data. Users can download data they need.

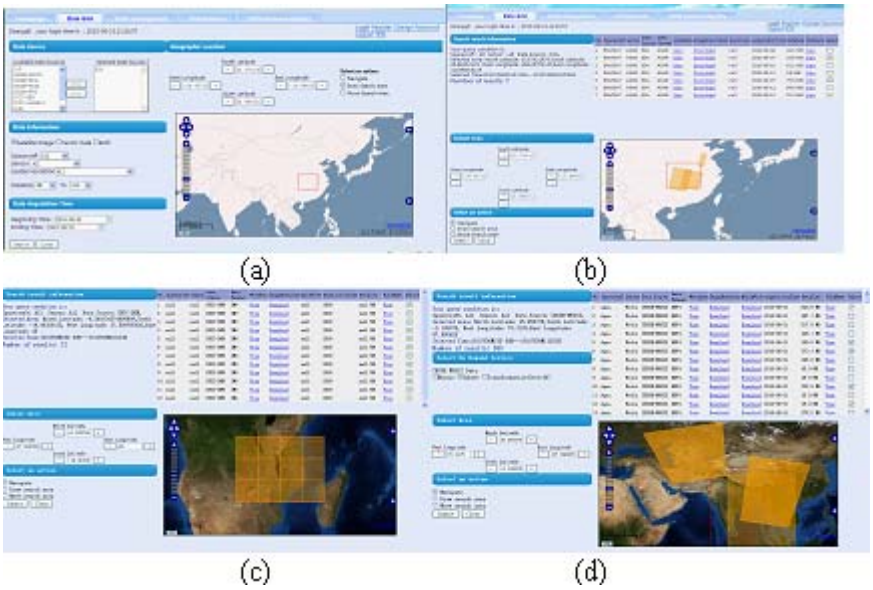

Fig. 8 . Data query and access client

CONCLUSIONS

In the paper "The Digital Earth: Understanding our Planet in the 21st Century", written by the former American Vice President Al Gore you can find the following: "A new wave of technological innovation is allowing us to capture, store, process and display an unprecedented amount of information about our planet and a wide variety of environmental and cultural phenomena. Much of this information will be "geo-referenced"– that is, it will refer to some specific place on the Earth's surface. The hard part of taking advantage of this flood of geospatial information will be making sense of it, turning raw data into understandable information[12]. This requires analysis capabilities that current Spatial Data Infrastructures cannot fully provide. Moreover, the increasing spatial and temporal resolution of geospatial data causes a tremendous challenge for their computation, together with the heterogeneous standards, system, work pattern and data content, with which traditional SDIs cannot cope. To address these challenges, the environmental science community is looking with interest to Grid computing infrastructures because these can satisfy the increasing need for processing power and storage capacity, can improve accessibility to distributed storage and computing resources, and can provide a reliable and secure infrastructure. In other words, Grids have the potential to underpin SDIs services and resources[13].

Aiming at the problem of heterogeneity of huge amount of spatial data, the concept of "Data Grid Entity Node" is proposed to package

various kinds of spatial data into services, providing users with universal method of data query and access. Besides, grid-enable WPS is represented by the encapsulation of Grid processing services with a standard WPS request by encoding directly Job Submission Description Language (JSDL) to describe job and resource requirements (disk space, CPU and other parameters) directly into the Execute request of WPS interface[14, 15]. Hence, users can invoke computing resources in the grid to accomplish their on-demand services, selecting ROI, not only time saving and efficient, but also practical.

Nowadays, more and more attention is paid on global change as various environmental problems are emerging. Along with development of scientific simulation and algorithm, there are more severe challenges in the field of spatial data infrastructure, involving storage, management, query, access, process, visualization and publish of massive spatial data, which may be distributed and heterogeneous. However, the diversity of spatial data, as well as the distribution of data agencies, brought a great obstacle for data sharing and collaboration. NASA conducted a study shows that, in the field of Earth Observation, scientists need to spend 50% to 80% of the time on data search and integration, even the final calculation and analysis is not difficult. The self-organization of spatial data sources can simplify the process of gridfication and encapsulate the data sources uniformly, making the query and access of data source both easy and convenient for users without professional knowledge on data management or network, which is beneficial for not only scientific researchers on global change, but also for the general public.

ACKNOWLEDGEMENTS

This work is supported by "Digital Earth Science Platform project" —Informatics study of Digital Earth Systems studies (Y01011101A), together with National Research and Development Program of China (973 Program 2009CB723906).

REFERENCES

1. Foley, R., Integrated Spatial Data Infrastructures. In Kitchin, R. and Thrift, N. (eds.) International Encyclopaedia of Human Geography, Volume V. Oxford, Elsevier, 2009; 507-511.
2. Di,L.P., McDonald,K.,Next Generation Data and Information System for Earth Sciences Research. The first International Symposium on Digital Earth, Beijing, China, 1999; 92-101.
3. http://en.wikipedia.org/wiki/Grid_computing
4. Foster, I., Kesselman, C., Nick M. J. And Tuecke, S. Grid Services for Distributed System Integration. *Computer*, 2002; **35**(6): 37-46.
5. Zeng Yi, Liu Dingsheng, Li Guoqing. Using Data Grid technology to build MODIS data management and distribution system based on Spatial Information Grid. The 7th International Grid and Cooperative Computing, 2008.
6. Rajabifard, A. and Williamson,I.P., Spatial data infrastructures: concept, SDI hierarchy and future directions, in: Geomatics'80, Tehran, Iran 2001.
7. Kiehle, C., Greve, K., Heier, C., Standardized geoprocessing—taking spatial data infrastructures one step further, in: 9th AGILE Conference on Geographic Information Science, Visegrad, 2006; 273-282.
8. Stollberg, B., Zipf, A., OGC web processing service interface for web service orchestration: aggregating geo-processing services in a bomb threat scenariom, Web andWireless Geographical Information Systems. in: Proceedings 7th International Symposium,W2GIS 2007; **4857**: 239-251.
9. Zhenchun Huang, Guoqing Li, Yi Zeng. Building data infrastructure for geo-computation by spatial information grid. GCC 2007: 630-635
10. Zeng Yi, Li Guoqing. Encapsulation method for the database access in grid application based on OGSA. *Journal of Huazhong University of Science and Technology*, 2005; **33**: 32-37.
11. OGC Consortium. http://www.opengeospatial.org/standards
12. Gore Al., The Digital Earth: Understanding our Planet in the 21st Century. Given at the California Science Center, Los Angeles, California, 1998.
13. Masser, Ian., GIS Worlds: Creating Spatial Data Infrastructures (First ed.). Redlands CA: ESRI Press, 2005.
14. Woolf, A. and Shaon, A., An approach to encapsulation of grid processing within an OGC web processing service, GIS Science 2009; **3**: 82–88.
15. Jes RYTTERSGAARD, Denmark, SPATIAL DATA INFRASTRUCTURE Developing Trends and Challenges. CODI-2, Addis Ababa, 2001.