Chapter 6
Geospatial Information Processing Technologies

Zhenlong Li, Zhipeng Gui, Barbara Hofer, Yan Li, Simon Scheider and Shashi Shekhar

Abstract The increasing availability of geospatial data offers great opportunities for advancing scientific discovery and practices in society. Effective and efficient processing of geospatial data is essential for a wide range of Digital Earth applications such as climate change, natural hazard prediction and mitigation, and public health. However, the massive volume, heterogeneous, and distributed nature of global geospatial data pose challenges in geospatial information processing and computing. This chapter introduces three technologies for geospatial data processing: high-performance computing, online geoprocessing, and distributed geoprocessing, with each technology addressing one aspect of the challenges. The fundamental concepts, principles, and key techniques of the three technologies are elaborated in detail, followed by examples of applications and research directions in the context of Digital Earth. Lastly, a Digital Earth reference framework called discrete global grid system (DGGS) is discussed.

Keywords Geospatial big data · High-performance computing · Online geoprocessing · Distributed geoprocessing · Discrete global grid system

Z. Li (✉)
Geoinformation and Big Data Research Laboratory, Department of Geography, University of South Carolina, Columbia, SC, USA
e-mail: zhenlong@sc.edu

Z. Gui
School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

B. Hofer
Interfaculty Department of Geoinformatics – Z_GIS, University of Salzburg, Salzburg, Austria

Y. Li · S. Shekhar
Department of Computer Science & Engineering, University of Minnesota-Twin Cities, Minneapolis, MN, USA

S. Scheider
Department of Human Geography and Spatial Planning, Universiteit Utrecht, Utrecht, The Netherlands

H. Guo et al. (eds.), *Manual of Digital Earth*,
https://doi.org/10.1007/978-981-32-9915-3_6

6.1 Introduction

With the advancement of sensor and computing technologies, massive volumes of geospatial data are being produced at an increasingly faster speed from a variety of geo-sensors (e.g., in situ and remote sensors) and model simulations (e.g., climate models) with increasing spatial, temporal, and spectral resolutions. For example, satellite sensors are collecting petabytes data daily. Climate model simulations by Intergovernmental Panel on Climate Change scientists produce hundreds of petabytes of climate data (Schnase et al. 2017). In addition to these traditional data sources, geospatial data collected from ubiquitous location-based sensors and billions of human sensors (Goodchild 2007) are becoming more dynamic, heterogeneous, unstructured, and noisy.

These massive volumes of geospatial data offer great opportunities for advancing scientific discovery and practices in society, which could benefit a wide range of applications of Digital Earth such as climate change, natural hazard prediction and mitigation, and public health. In this sense, efficiently and effectively retrieving information and deriving knowledge from the massive geospatial datasets have become critical functions of Digital Earth. The questions that can be (or should be) addressed with Digital Earth include, for example, how to investigate and identify unknown and complex patterns from the large trajectory data of a city to better understand human mobility patterns (e.g., Hu et al. 2019a, b), how to rapidly collect and process heterogeneous and distributed hazard datasets during a hurricane to support decision making (e.g., Martin et al. 2017; Huang et al. 2018), how to synthesize huge datasets to quickly identify the spatial relationships between two climate variables (e.g., Li et al. 2019), and how to find spatial and temporal patterns of human activities during disasters in massive datasets that are notoriously "dirty" and biased population samples (e.g., Twitter data) in a scalable environment (e.g., Li et al. 2018).

Geospatial information computing refers to the computational tasks of making sense of geospatial data. Such tasks mainly include but are not limited to geospatial data storage, management, processing, analysis, and mining. Addressing the above questions poses great challenges for geospatial information computing. First, the volume of the geospatial data at the global scale (e.g., at the petabyte-scale) exceeds the capacity of traditional computing technologies and analytical tools designed for the desktop era. The velocity of data acquisition (e.g., terabytes of satellite images a day and tens of thousands of geotagged tweets a minute) pushes the limits of traditional data storage and computing techniques. Second, geospatial data are inherently heterogeneous. They are collected from different sources (e.g., Earth observations, social media), abstracted with different data models (e.g., raster, vector, array-based), encoded with different data formats (e.g., geodatabase, NetCDF), and have different space and time resolutions. This heterogeneity requires interoperability and standards among the data processing tools or spatial analysis functions. For example, producing timely decision support often requires combining multiple data sources with multiple tools. Moreover, with the involvement of multiple tools and datasets in the problem-solving process, data provenance, analysis transparency, and result reproducibility

become increasingly important. Third, global geospatial data are often physically distributed. They are collected by distributed sensors and stored at data servers all over the world. Moving data from one location such as local server to another such as cloud for processing becomes problematic due to the high volume, high velocity, and necessity of real-time decision making.

A variety of processing and computing technologies have been developed or adapted to tackle these challenges. Figure 6.1 depicts a geospatial information computing framework of Digital Earth, highlighting three types of popular technologies in geospatial information computing: high-performance computing (HPC, Sect. 6.2), online geospatial information processing (or online geoprocessing, Sect. 6.3), and distributed geospatial information processing (or distributed geoprocessing, Sect. 6.4). HPC aims to tackle the large-volume challenge by solving data- and computing-intensive problems in parallel using multiple or many processing units (e.g., GPU, CPU, computers). Online geoprocessing comprises techniques that allow

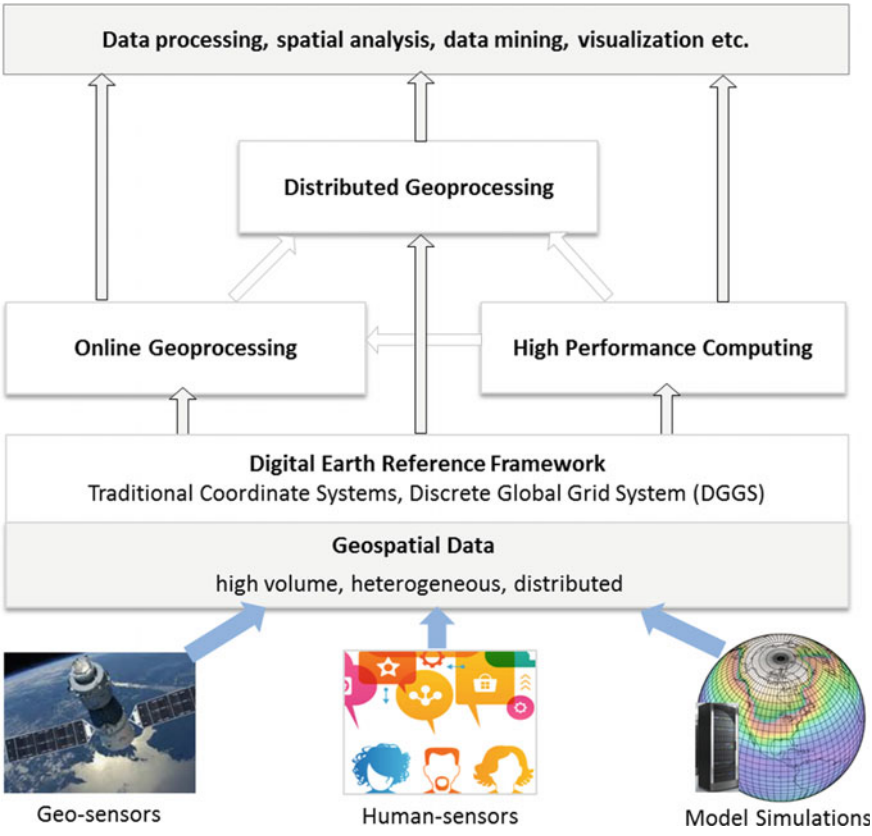


Fig. 6.1 Geospatial information computing framework of Digital Earth composed of high-performance computing, online geoprocessing, and distributed geoprocessing

for performing data processing and spatial analysis tasks on the web using geospatial web services (e.g., OGC web services) or web APIs (e.g., RESTful). Through standardization, these services and APIs are essential for addressing the heterogeneity challenges of geospatial data. Distributed geoprocessing refers to processing geospatial data and information in a distributed computing environment. By chaining a set of distributed data processing services into an executable workflow, the datasets and analysis steps involved in a task are documented, which improves the reproducibility of the analysis.

The following three sections start with a brief introduction and definition of a technology followed by its key principles, techniques, and examples of applications that support Digital Earth. Research challenges and future directions are discussed at the end of each section. A summary of the three technologies and a discussion of the discrete global grid system (DGGS) are provided in the last section. This chapter is not intended to be comprehensive or cover all aspects and technologies of geospatial information computing. The three selected technologies are described to provide the readers with a sense of geoinformation processing and how it is applied to support Digital Earth.

6.2 High-Performance Computing

6.2.1 The Concept of High-Performance Computing: What and Why

HPC aims to solve complex computational problems using supercomputers and parallel processing techniques. Since commodity clusters revolutionized HPC twenty years ago, a price-performance standard has become dominant, which includes inexpensive, high-performance x86 processors, functional accelerators (e.g., Intel Xeon Phi or NVidia Tesla), and open source Linux software and associated toolkits. It has been widely used in various applications such as weather forecasting, nuclear test simulation, and molecular dynamics simulation.

The growing availability of spatial datasets, in the form of GPS vehicle trajectories, social media check-ins, earth observation imagery, and sensor readings pose serious challenges for researchers and tool users in geo-related fields. The currently available computational technology constrains researchers and users in geo-related fields in two ways. First, the size of problems that can be addressed using the currently available methods is limited. Additionally, new problems, patterns, research, and decisions that may be discovered from geospatial big data cannot be found using existing tools. The 3 "V" s of big geospatial data (volume, variety, and velocity) impose new requirements for computational technology for geospatial information processing, for example, large, cheap, and reliable storage for large amounts of data, as well as scalable algorithms to process data in real time. Due to its computational capability, HPC is well suited for geospatial information processing of geospatial

big data. A highly integrated and reliable software infrastructure ecosystem based on HPC will facilitate geo-related applications in two ways. First, it will scale up the data volume and data granularity of data management, mining, and analysis, which has not been possible in the desktop era using the currently available methods. Furthermore, it will inspire and enable new discoveries with novel big-data-oriented methods that are not implementable in the current desktop software.

In the following, we describe HPC platforms frequently used in geospatial information processing, and look at how HPC is applied in spatial database management systems and spatial data mining.

6.2.2 High-Performance Computing Platforms

Since HPC was introduced in the 1960s, parallelism has been introduced into the systems. In parallelization, a computational task is divided into several, often very similar, subtasks that can be processed in parallel and the results are combined upon completion. The direct computational time savings of HPC systems results from the execution of multiple processing elements at the same time to solve a problem. The process of dividing a computational task is called decomposition. Task interaction necessitates communication between processing elements, and thus increasing granularity does not always result in faster computation. There are three major sources of overhead in parallel systems: interprocess interaction, idling, and excess computation. Interprocess interaction is the time spent communicating data between processing elements, which is usually the most significant source. Idling occurs when processing elements stop execution due to load imbalance, synchronization, or the presence of serial components in a program. Excess computation represents the extra time cost of adopting a parallel algorithm based on a poorer but easily parallelizable algorithm rather than the fastest known sequential algorithm that is difficult or impossible to parallelize.

To facilitate the parallelism of HPC systems, the architecture of HPC systems dictates the use of special programming techniques. Commonly used HPC platforms for large-scale processing of spatial data include the Message Passing Interface (MPI), Open Multi-Processing (OpenMP), Unified Parallel C (UPC), general-purpose computing on graphics processing units (GPGPU), Apache Hadoop, and Apache Spark. These platforms can be roughly classified according to the level at which the hardware supports parallelism.

OpenMP, MPI, and UPC support parallelism on central processing units (CPUs). OpenMP is an API that supports multi-platform shared memory parallel programming in C/C++ and Fortran; MPI is the most commonly used standardized and portable message-passing standard, which is designed to function on a wide variety of parallel computing architectures. There are several well-tested and efficient implementations of MPI for users programming in C/C++ and Fortran. They can work cooperatively in a computer cluster such that OpenMP is used for parallel data processing within individual computers while MPI is used for message passing

between computers. UPC extends the C programming language to present a single shared, partitioned address space to the programmer, where each variable may be directly read and written by any processor but is physically possessed by a single processor.

The GPGPU platform performs computations that are traditionally conducted by CPUs using graphic processing units (GPUs). Architecturally, a CPU is composed of a few cores that can handle complex tasks whereas a GPU is composed of hundreds of cores for simple tasks, so a GPU can dwarf the calculation rate of many CPUs if the computational task can be decomposed to simple subtasks that can be handled by a GPU's core. The GPGPU is programmed using programming models such as CUDA or OpenCL.

Due to the popularity of commodity computer clusters, the MapReduce programming model was introduced to maintain their reliability. Apache Hadoop, which is a collection of open-source software utilities based on the MapReduce programming model, can automatically handle hardware failures that are assumed to be common. Apache Spark was developed in response to limitations in the MapReduce model, which forces a linear dataflow structure to read and write from disk. Instead of a hard drive disk, Apache Spark functions on distributed shared memory. Figure 6.2 illustrates how HPC platforms support both spatial database management systems and spatial data mining.

The abovementioned HPC platforms facilitate the realization of several HPC applications such as cloud computing, newly emerging edge computing (Shi et al. 2016) and fog computing (Bonomi et al. 2012). Cloud computing is the on-demand availability of computational resources such as data storage and computing power without direct active management by the users. It emphasizes the accessibility to HPC over the Internet ("the cloud"). As the cost of computers and sensors continuously decrease and the computational power of small-footprint devices (such as

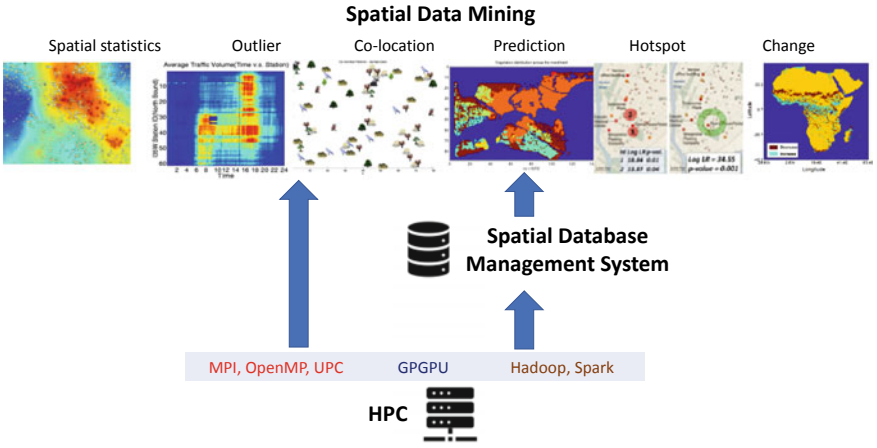


Fig. 6.2 HPC for spatial database management systems and spatial data mining

gateways and sensor hubs) increase, the concepts of edge computing and fog computing include more processing elements such as end devices in the Internet of Things in the computer clusters.

6.2.3 *Spatial Database Management Systems and Spatial Data Mining*

A database management system (DBMS) is a computerized system for defining, creating, querying, updating, and managing a database. It provides persistence across failures, concurrency control, and scalability to search queries of datasets that do not fit inside the main memories of computers. Spatial DBMSs are software modules that can work with an underlying DBMS; they were developed to handle spatial queries that cannot be handled by a traditional DBMS, for example, listing the names of all employees living within one kilometer of a company (Shekhar and Chawla 2003). Spatial DBMSs are an essential component of spatial data storage and management for geospatial information processing.

Spatial data mining is the process of quantifying and discovering interesting, previously unknown, potentially useful pattern families from large spatial datasets such as maps, trajectories, and remote sensing images (Shekhar et al. 2015). Compared with traditional data mining, spatial data mining has three special challenges. First, objects in space exhibit spatial dependence at nearby locations as well as distant locations. The spatial dependence at nearby locations is called the spatial autocorrelation effect. It is also known as Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things." For example, people tend to cluster together with others that share similar characteristics, occupation, and background. Examples of long-range spatial dependence, i.e., spatial tele-coupling, include El Niño and La Niña effects on the climate system. A second challenge is that spatial data is embedded in a continuous space whereas classical datasets are often discrete. Third, spatial heterogeneity and temporal nonstationarity make it difficult to find a global law that is valid across an entire space and for all time. In other words, spatial context matters. Consequently, classical data mining algorithms often perform poorly when applied to spatial data sets and thus more powerful methods such as spatial statistics and spatial data mining are needed.

Spatial statistics (Cressie and Wikle 2015) provides theories (e.g., spatial point process, geostatistics, and lattice statistics), models (e.g., spatial autoregression model), and methods (e.g., Kriging) for spatial data mining. Spatial data mining focuses on five pattern families, namely, outliers, colocations and tele-couplings, location prediction, hotspots, and spatiotemporal change. A spatial outlier is defined as a spatially referenced object whose nonspatial attribute values are inconsistent with those of other objects in its spatial neighborhood (Shekhar et al. 2003). Contrary to global outliers, whose nonspatial attributes are compared with the remainder of the dataset, the attributes of spatial outliers are compared with a local subset of

data around their footprints. For example, a road intersection where the vehicle speed is much higher than in other intersections nearby is a spatial outlier although it may not be a global outlier compared with other intersections in the city.

Spatial colocations represent subsets of spatial event types whose instances are often located in close geographic proximity (Huang et al. 2004). For example, the Nile crocodile and the Egyptian plover are frequently colocated, which indicates their symbiotic relationship. Other common colocation patterns include the colocation of the fast food restaurants McDonald's, Burger King, and KFC; the colocation of shopping malls with movie theaters; and the colocation of bars and drunk driving. Spatial tele-coupling represents interactions across distant locations. For example, the El Niño weather pattern (warming of the Pacific Ocean) affects the weather thousands of miles away in the midwestern and eastern United States.

Location prediction aims to learn a model to infer the location of a spatial phenomenon from maps of other spatial features. Examples include learning land-cover classification maps, predicting yearly crop yield, and predicting habitats for endangered species. Classical data mining techniques yield weak prediction models as they do not capture the spatial autocorrelation and heterogeneity in spatial datasets. Ignoring spatial autocorrelation often results in salt-and-pepper noise, i.e., locations whose predicted land-cover class is very different from the predicted land-cover classes of its neighboring locations. Such problems are significantly reduced by spatial autocorrelation-aware location prediction methods such as spatial autoregression, Markov random field-based Bayesian classifiers, and spatial decision trees (Jiang et al. 2015). Spatial heterogeneity, which prevents single-learner methods (e.g., neural networks and random forests) from accurately learning a global model is considered by spatial ensemble methods as well as Gaussian multiple instance learning methods (Jiang et al. 2017).

Spatial hotspots represent spatial regions where the concentration of objects inside the region is significantly higher than that outside. Hotspot analysis is widely used in public health and public safety to identify hotspots of disease and crime, respectively. False positives and true negatives carry high costs in such settings. Incorrectly labeling a neighborhood a disease or crime hotspot may lead to stigmatization and significant economic loss, and missing true hotspots of disease may lead to preventable mortalities and disease burden.

Spatiotemporal change may be defined in several ways. It may be a change in a statistical parameter, where the data are assumed to follow a distribution and the change is a shift of this distribution. It may be a change in actual value, where the change is defined as the difference between a data value and its spatiotemporal neighborhood. It may also refer to a change in models fitted to data, where the change is defined as a change in the models fitted to the data. Studies have been conducted to find more scalable algorithms for biomass monitoring using Gaussian process learning (Chandola and Vatsavai 2011).

There are many other interesting, useful and nontrivial patterns of interest in spatial data mining. For example, emerging hotspot detection aims to detect disease outbreak well before an outbreak results in a large number of cases. Interested readers are referred to papers on spatial data mining (Shekhar et al. 2011, 2015) and parallel

computing algorithms for GIS (Healey et al. 1997; Shekhar et al. 1996, 1998; Zhao et al. 2016) for additional details.

6.2.4 Applications Supporting Digital Earth

Spatial database management systems using HPC have been studied extensively. Since Hadoop was introduced and its ability to handle big data in computer clusters was demonstrated, researchers and savvy tool users have taken advantage of it in various ways. Some tools and studies use Hadoop as a black box for operations on data, such as GIS tools for Hadoop, a package composed of programming libraries and an add-on toolbox of ArcGIS desktop (ESRI 2018), and Hadoop-GIS, a scalable spatial data warehousing system (Aji et al. 2013). Spatial Hadoop adds native support for spatial data by supporting a set of spatial index structures and developing spatial functions that interact directly with Hadoop base code (Yao et al. 2017). Impala, a distributed SQL query engine for Hadoop, has also been extended for spatial data (Eldawy et al. 2015). Apache Spark's core in-memory data abstraction, called a resilient distributed dataset (RDD), outperforms MapReduce-based approaches. Inefficient handling of interactive operations, the performance bottleneck of Hadoop-based tools, is addressed by GeoSpark, which adds support for spatial data and operations to Spark (Yu et al. 2015). GCMF, an end-to-end software system on GPGPU, illustrates the potential of GPGPU as a platform for geospatial information processing, as it can handle spatial joins over non-indexed polygonal datasets containing more than 600,000 polygons on a single GPU within 8 s (Aghajarian et al. 2016).

HPC is also applied in spatial data mining. Examples of HPC for spatial statistics include parallelizing the computation of statistical measures (e.g., Moran's I and Getis-Ord) using MPI and OpenMP (Wang et al. 2008; Kazar et al. 2004). Parallelization of the interpolation method has also been studied. Parallelized Kriging has been implemented on both MPI and GPGPU (Pesquer et al. 2011; de Ravé et al. 2014). Hadoop and Spark have also been leveraged as platforms to implement Kriging and inverse distance-weighted interpolation algorithms (Xu et al. 2015; Rizki et al. 2017). Parameter estimation for many spatial statistical models (e.g., spatial autoregression and space-time kernel density estimation) relies on matrix operations and may benefit from parallel formulations of linear algebra algorithms. A parallelization of wavelet transform, which can locate frequency outliers, has been implemented on MPI to scale up outlier detection algorithms (Barua and Alhajj 2007). Both GPU-based and OpenMP-based parallel algorithms have been explored for spatial prediction and classification (Gandhi et al. 2006; Rey et al. 2013). Researchers are investigating the use of GPUs as a platform for computing likelihood ratios as well as Ripley's K function (Pang et al. 2013; Tang et al. 2015). GPU-based methods have also been introduced to accelerate the computation of change detection (Prasad et al. 2013, 2015).

6.2.5 Research Challenges and Future Directions

HPC is essential for handling today's growing volumes of spatial data and the ever-increasing size and complexity of geospatial information processing problems. In addition to the existing methods and tools, further study in two focus areas is necessary to take full advantage of HPC for geospatial information processing.

The first focus of study is the parallelization of the currently available methods for HPC. The ubiquitous existence of spatial autocorrelation makes parallelization not applicable for most geo-related algorithms because the dependence between data partitions requires task interaction, which increases the difficulty of parallelizing serial algorithms in spatial database and spatial data mining functions. Additionally, the load balancing between processing elements is complicated when dealing with sparse data structures for which the pattern of interaction among data elements is data-dependent and highly irregular. Spatial networks (e.g., road networks) are an example of these data structures.

The second focus of study is utilization of geospatial big data to discover novel problems, patterns, research, and decisions. For example, most current research in spatial data mining uses Euclidean space, which often assumes isotropic properties and symmetric neighborhoods. However, the distribution of many spatial phenomena is strongly affected by the underlying network space, such as rivers and road networks. Some cutting-edge research has been conducted to generalize spatial analysis and data mining methods to the network space, such as network spatial interpolation (Kriging), network point density estimation, and linear hotspot detection (Okabe and Sugihara 2012). However, more research is needed in the network space. For example, in addition to the shortest paths, simple paths or irregular subgraphs are potential candidates for study in linear hotspot detection problems to discover interesting patterns.

In addition to the network space, the curved surface of the Earth is rarely considered in the currently available spatial database and data mining functions. For example, Chap. 2 discusses extending spatial indexing based on a space-filling curve and coordinate system to the curved surface. However, another family of spatial indexing, R-tree, which is the default spatial indexing supported by major DBMSs such as Oracle, MySQL, and PostGIS, only works in Euclidean space. Additionally, the definition of distance on the curved surface of the Earth is different from that in the Euclidean space, which affects the discovery of spatial patterns such as outliers, hotspots, and colocation.

Spatial heterogeneity is another topic to be explored. Spatial heterogeneity refers to the uneven distribution of spatial phenomena within an area. Most of the existing methods focus on the discovery rules or patterns valid for the whole dataset. However, the belief that spatial context matters is a major theme in geographic thought (Miller and Goodchild 2015). Different rules or patterns may exist in various places. If a pattern is infrequent relative to the size of the whole dataset, it may be missed if the entire dataset is analyzed. Such localized patterns are easier to find in smaller subsets of the data, around their spatial footprints. Identifying these patterns is challenging

due to the need to enumerate all relevant footprints that may include an exponential number of data partitions (e.g., subgraphs of a road network). Examples of research on this topic include the spatial ensemble classification method (Jiang et al. 2017) and study of local colocation pattern detection (Li and Shekhar 2018).

Both the abovementioned future research directions pose new challenges for the computational capacity of currently available systems and tools. A highly integrated and reliable infrastructure ecosystem of HPC is required for geospatial information processing because most existing approaches focus on parallelization of specific tasks. Such an infrastructure can be utilized to speed up data management, mining, and analysis projects with scale and data granularity that were previously not possible, and enable new discoveries and ways of planning and decision making with novel big-data-oriented tools that are unavailable in the standard software.

6.3 Online Geospatial Information Processing

6.3.1 Web Service-Based Online Geoprocessing

Online geoprocessing refers to the use of spatial analysis functionality (such as buffer, interpolation and filtering operations) on the web to generate the desired output by applying a requested operation or chains of operations on input data. For the client-server interaction to work, clients and servers must be able to exchange requests and responses, for example, in the form of standardized web services. The standardization body in the geoinformatics sector is the Open Geospatial Consortium (OGC) (http://www.opengeospatial.org/standards/owc). OGC standards are aimed to provide syntactically interoperable services to facilitate integration, exchange and reuse of observations, data and geocomputational functions. Current applications of online geocomputation in the context of Digital Earth demonstrate the benefits of this standards-based technology. Some examples of such applications as well as challenges for advancing online geoprocessing for Digital Earth applications are discussed in Sect. 6.3.3. The alternative to standardized web services is application programming interfaces (API) and data formats such as JSON—the JavaScript Object Notation, which are increasingly popular (Scheider and Ballatore 2018). However, the plethora of available APIs limits the reusability of services that is with standardized approaches.

OGC service specifications cover services for raster or vector data, sensor observations, processing services, catalog services, and mapping services. The principle behind these services is that the interfaces are standardized, which means that resources can be requested following a set of defined parameters via the hypertext transfer protocol (HTTP). The requests are processed by a server and a response is sent back to the requesting user or service; the responses are generally encoded in XML (eXtensible Markup Language). Providers of web services can register their services in catalogs such that clients can discover and use these services. This

publish-find-bind principle is fundamental in service-oriented architectures (SOAs) that realize the principle of integrating resources from distributed sources. Service-oriented architectures are commonly used in the context of spatial data infrastructures and (open) data initiatives, for example, GEOSS (http://www.geoportal.org).

According to Yue et al. (2015) such web services have the potential to become *intelligent,* i.e., easing the *automated* discovery and composition of data and processing services to generate the required information at the right time. To realize this vision, a move from the currently supported syntactic interoperability towards *semantic interoperability* is a core requirement (Yue et al. 2015). This section discusses the state-of-the-art of online geoprocessing in the context of Digital Earth as well as current lines of research related to semantics of geocomputational functions and spatial data. The objectives of this section are reflected in its structure: Sect. 6.3.2 introduces the principles of two geoprocessing services—the web processing service (WPS) and the web coverage processing service (WCPS). Section 6.3.3 discusses the state-of-the-art by reviewing successful applications of geoprocessing technology. Some current research trends and future directions to realize intelligent online geoprocessing are discussed in Sect. 6.3.4.

6.3.2 Web (Coverage) Processing Services

The key technologies for web service-based online processing are web processing services (WPSs) and web coverage processing services (WCPSs). As their names suggest, WCPSs provide processing functionality for coverages and is related to the web coverage service (WCS) standard; WPS provide general processing functionality for geospatial data. Both of these services follow the overall design principle of interoperable OGC web services and are briefly introduced below.

WPSs are currently available in version 2.0. A WPS must support the GetCapabilities, DescribeProcess and Execute requests, which are sent to the server using the HTTP GET or POST methods or the simple object access protocol (SOAP) (http://cite.opengeospatial.org/pub/cite/files/edu/processing/basic-index.html). The responses of the GetCapabilities and DescribeProcess requests contain information on parameter values required for an Execute request. These pieces of information cover the input, parameters and output of processes. Input and output data, which are either complex or literal data, are specified with a description and information on mimeType (e.g., "text/xml"), encoding (e.g., "UTF-8") and schema (e.g., "http://schemas.opengis.net/gml/3.2.1/feature.xsd"). It is possible to specify the data types of literal data as well as allowed values. WPS can be executed in synchronous or asynchronous modes; asynchronous execution is preferred for calculations that take longer.

The nature of WPS is generic as the kind of calculation a processing service provides is not specified. The generic nature of WPS is said to be one reason for its slow uptake, as it is difficult for clients to deal with the variety of outputs generated by different WPSs (Jones et al. 2012). The process implementations are hidden from

the users; the information provided on the processes includes their input, output and parameters as well as a title, description, identifier and additional optional metadata. To reuse processes, it is essential to have information on what a process does and the datasets it can be applied to. Thus, process profiles have been revised and modified to describe the meaning of operations and their inputs and outputs in the WPS 2.0 standard (Müller 2015).

The web coverage processing service is an extension of the WCS standard with an explicit focus on the processing of coverages, i.e., multidimensional raster data; it has been available since 2008. The current WCS 2.1 version supports the GetCapabilities, DescribeCoverage and GetCoverage requests. These requests are extended for the ProcessCoverage request in WCPS. Filter mechanisms that restrict the spatial or temporal extent of the processed data are a core requirement for interaction with multidimensional coverages. The WCPS provides a specific syntax, which is somewhat similar to the structured query language SQL, for formulating queries of temporal and spatial subsets of data (Baumann 2010). WCS and WCPS can handle a multitude of different formats of data encodings that are relevant in the context of image data; these include NetCDF, GeoTiff, JPEG, and GRIB2. A tutorial on WCS and its extensions is available on Zenodo (Wagemann 2016).

Although WCPS was specifically designed for coverage data, its reuse across applications is hindered by diverging definitions of data models and the heterogeneity of data formats (Wagemann et al. 2018).

6.3.3 Online Geoprocessing Applications in the Context of Digital Earth

This section presents three recent examples of application of online geoprocessing. These applications were published in a related special issue aimed at promoting online geoprocessing technology for Digital Earth applications (Hofer et al. 2018). The applications demonstrate the use of web processing services and web coverage processing services as extensions of existing infrastructures in a variety of contexts. They derive relevant and timely information from (big) data in efficient and reusable manner, which serves the objectives of Digital Earth.

Wiemann et al. (2018) focus on the assessment of water body quality based on the integration of data available in SDIs to date; the data types considered are feature objects and raster data. Their work introduced a new concept of geoprocessing patterns that suggest the application of processing functionality based on input data selected by the user of the application. The motivation behind this development is to assist users in deriving information from data. Their information system supports determination of river sinuosity as an indicator of the ecological quality of rivers, assessment of real estate values potentially affected by floods, and the discovery of observations made along rivers.

Stasch et al. (2018) present the semiautomatic processing of sensor observations in the context of water dam monitoring. An existing infrastructure makes sensor observations such as water levels and GPS measurements of dam structure available and the objective of their work is to statistically analyze the observations and use them as model inputs. Their motivation to use WPS is related to the possible reuse of services and flexibility regarding the integration of sensor observations from other sources in the final decision making. The coupling of sensor observation services (SOSs) with WPS is not a standard use case. Therefore, Stasch et al. (2018) discuss various approaches of coupling SOSs and WPS and selected a tight coupling approach in which a processing service can directly request observations from an SOS, which reduces overhead in communication. The authors also developed a REST API for WPS to reduce the required parsing of extensive XML files and ease client development; they provided the specification of a REST binding, which is lacking in the current WPS 2.0 standard.

Wagemann et al. (2018) present examples of the application of web coverage processing services in the context of big Earth data. They show how online geoprocessing supports the derivation of value-added products from data collections and how this technology transforms workflows. They state that server-side data processing can overcome issues using different solutions for data access and can minimize the amount of data transported on the web (Wagemann et al. 2018). They described examples of the application of WCPS in the domains of ocean science, Earth observation, climate science and planetary science; all of the examples use the rasdaman server technology. One of the presented applications for marine sciences provides a visual interface where a coverage of interest such as monthly values of chlorophyll concentration that were derived from ocean color satellite data can be specified (http://earthserver.pml.ac.uk/www). The provided coverage data can be compared with in situ measurements via a match-up tool. The match-up is calculated on the server and the users are presented with the results without having to download the chlorophyll data to their machines. The provider of this service must offer the required computing resources and the limitation of requests to a certain data volume is a known issue (Wagemann et al. 2018).

6.3.4 Research Challenges and Future Directions

Online geoprocessing technology has been improved over the last decade and the applications demonstrate its usability in real-world use cases. The potential of standardized web services lies in the flexible integration and reuse of services and computational power from different providers. However, in addition to the costs of service provision to potential clients, the *complexity* and *opacity* of geoprocessing workflows seem to hinder their mass usage. This is indicated by the fact that mapping services and data services are much more widely spread than processing services (Lopez-Pellicer et al. 2012). The reasons for this are manifold and relate to the variety of data models and formats, which limits the applicability of existing processing

services (Wagemann et al. 2018), lacking descriptions of processing services such as those approached with WPS process profiles (Müller 2015) and the required transfer of potentially large data from a data provider to a service provider.

Assuming that geoprocessing services are available for reuse across applications, the most relevant current challenges concern the *opacity* of service, data and tool interfaces, and the corresponding lack of clarity about when a geocomputational service is potentially useful. Applying a geocomputational function is a matter of analytic purpose as well as of the properties of the data sources used. The latter goes well beyond data types and necessarily involves background knowledge about the semantics of spatial data (Hofer et al. 2017; Scheider et al. 2016). Thus, it was recognized early in the field of geocomputation that, in addition to syntactic interoperability (i.e., the matching of formats and data types), *semantic interoperability* must be taken into account (Ouksel and Sheth 1999; Bishr 1998). Since then, many attempts have been made to incorporate semantics into service descriptions, e.g., in the form of Datalog rules and types that restrict the application of geocomputational functions (Fitzner et al. 2011; Klien et al. 2006). The technology evolved as a particular (service-oriented) strand of the semantic web, starting in 2000 (Lara et al. 2004) and resulting in standards such as the semantic markup for web services (OWL-S) (https://www.w3.org/Submission/OWL-S/) and web service modeling language (WSML) (http://www.wsmo.org/).

Researchers of semantic web services have shown that service descriptions and Semantic Web technology can be effectively combined and that abstracting from particular implementations of geocomputational functions remains very difficult (Treiblmayr et al. 2012). Which aspects of such a function are mere technicalities? Which aspects are essential and thus should be represented on the semantic level of the service and data? More generally, what does a reusable representation that is valid across implementation specific details look like (Hofer et al. 2017)? The lack of a good answer to these questions in semantic web service research, e.g., in terms of a *reusable service ontology,* may be the reason why semantic web processing services have become less of a focus in research today. Drawing a line between semantic and syntactic interoperability is not straightforward, and different and incompatible "ontological" views on the world must be acknowledged (Scheider and Kuhn 2015). The need to infuse and reuse such flexible semantics in the age of big data has not lessened and is more urgent than ever (Janowicz et al. 2014; Scheider et al. 2017).

We currently lack *reusable representations* of the different views that make geoprocessing operations and data sources useful for a specific purpose. We also lack *neat theories* that tell us which concepts and aspects should be retained to describe data and geocomputational functions from the practical viewpoint of data analysis. Ontology design patterns have been proposed as a means to create such representations (Gangemi and Presutti 2009) and have recently gained popularity. Furthermore, it is an open question how geocomputational functions relate to the *purposes of analysis*. Finally, we need *computational methods* that allow for us to *infuse* the needed background knowledge into service and data descriptions to enable *publishing* and *exploiting* it.

Current research on semantic issues in geoprocessing tackles these challenges to support spatial analyses. We summarize three main lines of research that have evolved in recent years that may be promising for progress on a *semantically interoperable* Digital Earth:

6.3.4.1 Service Metadata, Computational Core Concepts, Linked Data and Automated Typing

In the current web processing service standards, to reuse a service it is necessary to describe the capabilities of the service and the service parameters (including data sources) in terms of metadata. However, the current metadata standards do not specify how to do this. It remains unclear which language and concepts should be used in these descriptions; it is also unclear how these concepts can be shared across communities of practice and how they can be automatically added without manual intervention. Regarding the first problem, several recent investigations attempted to identify a necessary and sufficient set of "core" concepts of spatial information (Kuhn 2012; Kuhn and Ballatore 2015; Scheider et al. 2016), which remain to be tested in diverse practical analytical settings. Regarding the second problem, linked open data (LOD) provides a way to remove the distinction between metadata and data, enabling us to publish, share and query data and its descriptions at the same time (Kuhn et al. 2014). Similarly, Brauner (2015) investigated the possibilities of describing and reusing geooperators with linked data tags on the web, and Hofer et al. (2017) discussed how such geooperator descriptions can be used for workflow development. Regarding the third problem, it has long been recognized that semantic labeling is a central automation task for the semantic web, as users tend to avoid the extra manual work involved. For this purpose, it has been suggested that the provenance information contained in workflows can be used to add semantic labels to the nodes in such a workflow (Alper et al. 2014). For the geospatial domain, it was demonstrated that the information contained in GIS workflows can be used to enrich geodata as well as GIS tools with important semantic types by traversing such a workflow, and share this information as linked data (Scheider and Ballatore 2018). Furthermore, certain semantic concepts such as the distinction between extensive and intensive attributes, which is central for geocomputation and cartography, can be automatically added as labels using machine learning classifiers (Scheider and Huisjes 2019).

6.3.4.2 From Service Chaining to Automated Workflow Composition

Automated service chaining has been a scientific goal and research topic since the start of the development of semantic web services (Rao and Su 2005). Ontologies are used to describe the restrictions on input and output types, which can be exploited by service chaining algorithms to suggest syntactically valid workflows. This idea has also been adopted for the geospatial domain (Yue et al. 2007), where the ontological concepts were mainly based on geodata types. However, in the wider area

of workflow composition (Gil 2007; Naujokat et al. 2012), finding efficient composition algorithms is not the issue, finding the relevant semantic constraints that render the problem tractable is. Once such constraints are found, it is much easier to devise an algorithm that makes service composition computable for practical purpose, and it becomes possible to filter out syntactically valid but *nonmeaningful* workflows that are currently clogging the workflow composition flows (Lamprecht 2013). Thus, similar to the metadata challenge discussed above, scientific progress largely depends on whether we will be able to devise a set of reusable valid semantic concepts for both geocomputation and geodata. In the future, it would be valuable to measure the effectiveness of spatial semantic concepts in reducing computational time and increasing accuracy in automated GIS workflow composition.

6.3.4.3 From Geocomputation to (Indirect) Question Answering

Since the application of geocomputational tools and the chaining of services require lots of background knowledge and GIS skills, their usage is currently restricted to GIS experts. However, those with little or no technical expertise in this area would benefit most, as well as those with a *relevant spatial question* about Digital Earth. How can Digital Earth technology help such users answer their questions? Question-based spatial computation was proposed as a research topic by Vahedi et al. (2016) and Gao and Goodchild (2013). The question-answering (QA) computational technique has been investigated during the last two decades from the information retrieval perspective (Lin 2002). Standard QA technology parses a natural language question and matches it with answers available in a database or the web. Recently, linked data-based data cubes were proposed as a way to realize question answering on a web scale (Höffner et al. 2016). However, question answering for geocomputation and analysis requires handling questions that *do not yet have an answer* but could be answered using appropriate tools and data. The latter problem was therefore termed *indirect question answering* by Scheider et al. (2017). A semantically informed retrieval portal that can answer such questions should be able expand a data query in a way that encompasses data sets that do not directly answer a given query but can be made to do so via appropriate analysis steps. For this purpose, geocomputational tools and datasets need to be described by the questions they answer, so that they can match the questions posed by a user. A recent first step in developing such a system for a set of common GIS tools was made based on SPARQL query matching (Scheider et al. 2019), following the idea of query matching for service descriptions proposed by Fitzner et al. (2011). However, similar to the previous two computational challenges, the kinds of questions and the matching language and technology are dependent on our theories of spatial (interrogative) concepts used to formulate these questions. In the future, we should investigate what kinds of spatial questions are relevant and how they can be formally captured in terms of core concepts. For related work, refer to the ERC-funded project QuAnGIS: Question-based analysis of geographic information with semantic queries (https://questionbasedanalysis.com).

6.4 Distributed Geospatial Information Processing

6.4.1 The Concept of Distributed Geospatial Information Processing: What and Why

Distributed geospatial information processing (DGIP) (Yang et al. 2008; Friis-Christensen et al. 2009) refers to geospatial information processing (geoprocessing for short) in a distributed computing environment (DCE). With the development of the Internet and world wide web, architecture modes of software have changed dramatically. Decentralization and cross-domain collaboration under a loosely coupled and dynamically changed DCE has become an emerging trend. Adoption of service-oriented architecture (SOA) and cloud computing is a promising and prevalent solution for modern enterprises to enhance and rebuild their cyberinfrastructure. Using these technologies, it is more agile and much easier to build cooperation networks and adjust cross-enterprise business workflows dynamically.

Following this trend, geographical information systems (GISystems) are also experiencing an evolution from traditional stand-alone toolkits to web service-based ecosystems (Gong et al. 2012), e.g., the geospatial service web (GSW). The GSW is a conceptual framework for a loosely coupled geospatial collaboration network through which the end users can share and exchange geospatial resources and conduct geoprocessing online by using distributed geographical information services (GIServices). In the GSW, everything is encapsulated as a service (XaaS), as shown in Fig. 6.3, including computing resources (CPU, memory, storage and network, etc.), geospatial data, models, algorithms and knowledge. The wide adoption of the enabling technologies such as web services, SOA and cloud computing make such a distributed geospatial collaboration network possible but there are also challenges. One of the major challenges is how to guarantee the reliability of geoprocessing in a mutable DCE (Gong et al. 2012; Wu et al. 2015). Traditionally, geoprocessing is conducted on a single machine with a stand-alone GISystem toolkit installed. Since the functional components of a GISystem are tightly coupled, it is relatively easy to capture and handle geoprocessing exceptions and ensure the whole geoprocessing process, e.g., a workflow synthesized with coordinate transformation, buffering and overlay operations. In comparison, in a DCE, it is complicated to define, coordinate and guarantee such a process due to the complexities in data transmission, workflow control and exception handling.

Therefore, DGIP has become a research hotspot as well as an application trend (Yang et al. 2008; Wu et al. 2014; Jiang et al. 2017). In this section, we introduce the basic concept and key techniques of DGIP, and demonstrate its applications in Digital Earth. Finally, we discuss the technical challenges and future directions.

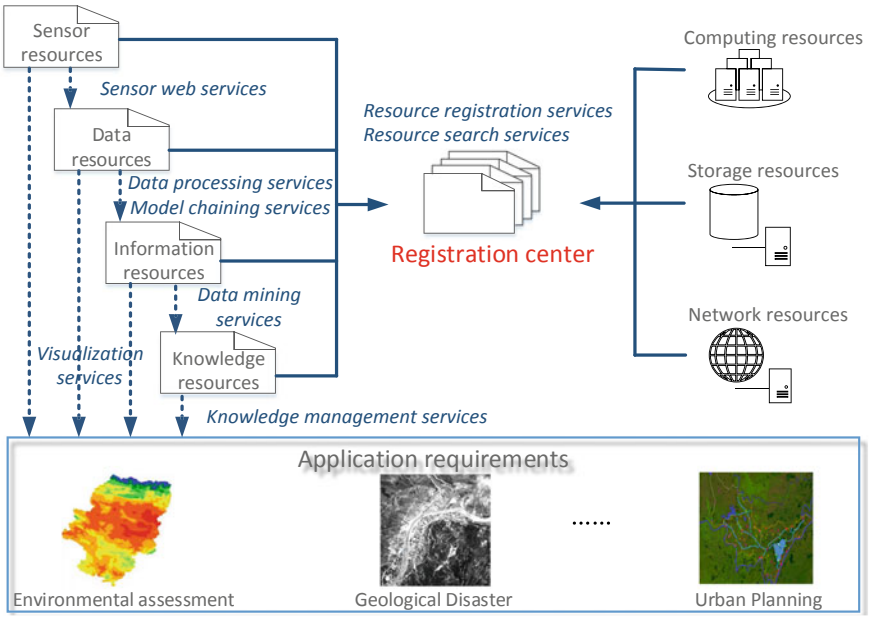


Fig. 6.3 The Conceptual framework of the geospatial service web (Gong et al. 2012)

6.4.2 *Fundamental Concepts and Techniques*

6.4.2.1 Collaboration Mode (Orchestration vs. Choreography)

For a distributed workflow to operate appropriately, the coordination and controlling mechanism is critical. In an SOA context, there are two basic collaboration modes (Peltz 2003), choreography and orchestration, based on the control flow patterns and how messages are exchanged, as illustrated in Fig. 6.4.

Web service orchestration (WSO) employs a centralized approach for service composition. A workflow is represented by a centralized coordinator that coordinates the interaction among different services. The coordinator or so-called composite service is responsible for invoking service partners, manipulating and dispatch messages. The relationships between the participating services are maintained by the coordinator. Since WSO adopts a hierarchical requester and responder model, it is process-centralized and the cooperation among participating services is weakened. The participating services do not need to know about each other in collaboration. In WSO, the status maintenance and error handling are relatively easier since it can be monitored and controlled by the coordinator. When an exception occurs, the coordinator can trigger exception handling or a compensation mechanism before the workflow progresses into the next step.

In comparison, web service choreography (WSC) adopts a decentralized peer-to-peer model. There is no a centralized compose service acting as the coordinator to

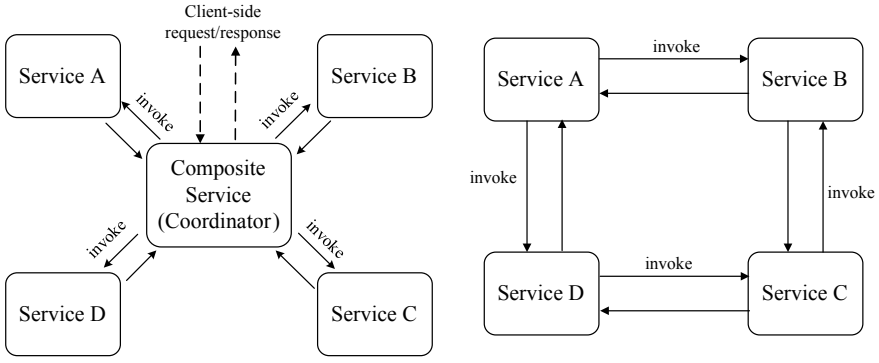


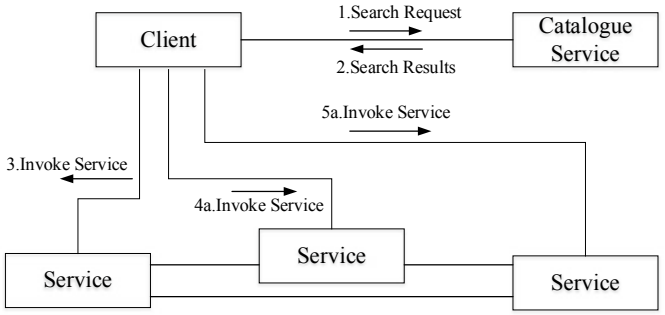
Fig. 6.4 Architectures of web service orchestration and web service choreography

control participating services, which makes it much more loosely coupled. The whole workflow is defined by exchanging messages, rules of interaction and agreements between services. Each participating service knows when and how to interact with each other, as well as whom to interact with, i.e., it is self-described and highly autonomous. One side effect is that it is difficult to detect errors in a timely manner and conduct exception handling from the workflow perspective. However, it can avoid the performance bottleneck problem for the coordinator in message exchange and data transmission.

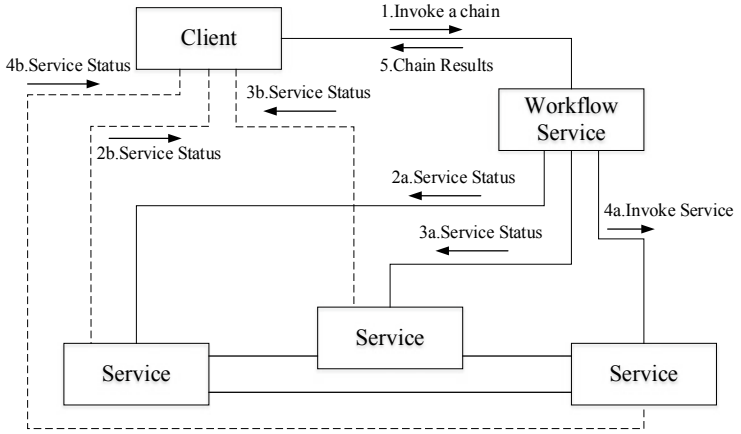
In summary, the WSC describes the interactions between multiple services from a global view whereas WSO defines control from one party's perspective, and the control logic of the interactions between services is explicitly modeled in the composite service (Peltz 2003). Therefore, WSO is generally an intraorganization workflow modeling solution whereas WSC is more suitable for interorganizational or cross-domain workflow modeling when it is difficult to set up a centralized coordinator across the boundary of management.

Learning from service composition in the IT domain, the geospatial domain proposed the concept of a geospatial service chain, which is defined as a model for combining services in a dependent series to achieve larger tasks for supporting DGIP. According to the definition of international standard ISO 19119 (2002), there are three types of architecture patterns to implement a service chain, as illustrated in Fig. 6.5, by giving different controlling authorities to clients (Alameh 2003; ISO 19119 2002), i.e., user-defined chaining, workflow-managed chaining and aggregated chaining.

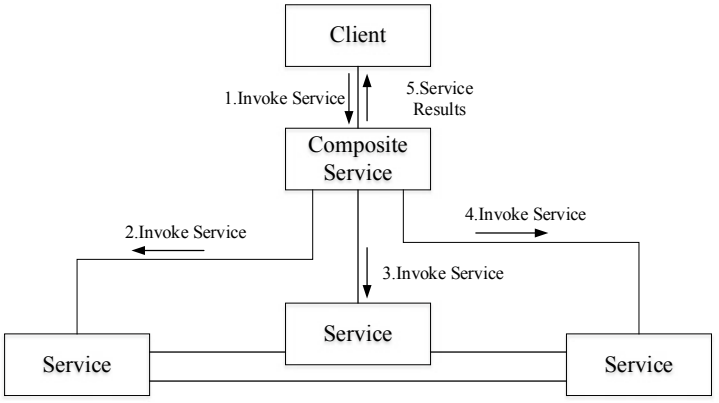
- In user-defined (transparent) chaining, the client defines and controls the entire workflow. In this case, the client discovers and evaluates the fitness of available services by querying a catalog service, which gives most freedom to the client to make the control decision and ask for workflow modeling knowledge.
- In workflow-managed (translucent) chaining, the workflow management service controls the service chain and the client is aware of the participating services. In



(a) Transparent Chaining



(b) Translucent Chaining



(c) Opaque Chaining

Fig. 6.5 Three architecture patterns for geospatial service chains defined in ISO 19119

this mode, the client can check the execution status of individual participating services, and the workload on workflow control is reduced.
- In aggregated service (opaque) chaining, the client invokes a compose service without awareness of the individual participating services. The compose service manages all the details of chain execution.

Although ISO 19119 gives the architecture patterns of service chaining, there is no de facto domain-specific standard on modeling language. The modeling languages of web service composition introduced in the next section use service chain modeling as a reference.

6.4.2.2 Workflow Modeling Language

A formalized model description language is desired to allow for a service chain be understood and shared among heterogeneous systems. Computer-aided business process management (BPM) has been widely used in modern enterprises for decades. Due to the variety of the backend IT enabling technologies and application scenarios, there are hundreds of workflow languages developed by different communities. These languages have different capabilities for flow rule expression (Aalst et al. 2003). In general, the languages can be classified into industrial specifications and academic models (Beek et al. 2007).

Industrial workflow specifications are model languages that target a certain technique implementation, and are widely supported by companies and standardization organizations. Web services business process execution language (WS-BPEL) and web service choreography description language (WSCDL) are two workflow standards specialized for web service composition. There are many open-source and commercial toolkits for reliable workflow modeling and execution management based on these specifications. However, these specifications are usually mixed with lower-level techniques such as XML encoding, XQuery, SOAP, WSDL and WS-addressing. These technical details increase the learning curve for users that lack or have little background knowledge of programming and web service standards. Workflow Management Coalition (WfMC) created an XML-based process definition language (XPDL) to store and exchange workflow models defined by different modeling language that is independent of concrete implementation techniques. XPDL is considered one of the best solutions to formally describe workflow diagrams defined using business process modeling notation (BPMN).

Academic workflow models express abstract process structures and rules that are not bound by a concrete runtime environment, lower-level implementation details and protocols (Beek et al. 2007; Gui et al. 2008), e.g., automata and process algebras. Directed graph and mathematical notations are widely used for workflow description, e.g., Petri nets (Hamadi and Benatallah 2003). Academic workflow models can express abstract process knowledge and have strict mathematical logics for process validation. However, these models are less used in industrial environments, and software to support workflow modeling and runtime management is lacking.

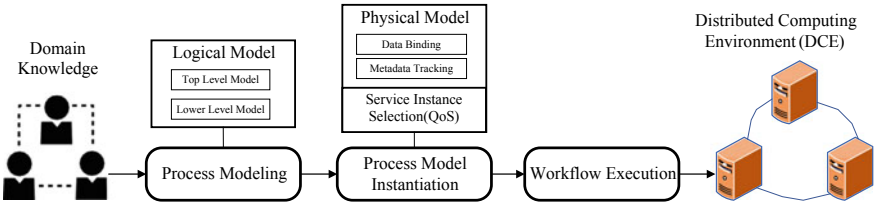


Fig. 6.6 A multistage geospatial service chaining process

In terms of geospatial service chaining, there is no well-accepted model language and domain-specific modeling methods should be developed. The European Space Agency (ESA) adopted WS-BEPL and established a service partner network to support global collaboration on earth observation. WS-BPEL is a de facto and widely used standard of the Organization for the Advancement of Structured Information Standards (OASIS) derived from the combination of IBM's Web Services Flow Language (WSFL) and Microsoft's notation language for business process design (XLANG). However, the lower-level technique details in WS-BPEL may be beyond the expertise of domain experts without web service knowledge. WS-BPEL adopts static binding to specify service partners and communication rules in advance and makes it difficult to adopt a dynamic and mutable environment. Therefore, a multistage geospatial service chaining method is highly desired to separate abstract geoprocessing workflow knowledge and lower-level implementation details, and make service partner binding dynamic, as illustrated in Fig. 6.6.

As shown in Fig. 6.6, Di et al. (2006) divided geospatial service chaining process into three steps, geoprocessing modeling, geoprocessing model instantiation and workflow execution. In the processing modeling stage, geoscience domain experts use a logical model language to depict abstract geoprocessing workflows based on process knowledge. In the process model instantiation stage, the logical model is mapped into a physical model that binds with implementation details. The data sources of the input data and service instances of participating services are specified during instantiation. Then, the physical model can be deployed into a workflow engine for workflow execution and runtime management.

6.4.3 Application Supporting Digital Earth

6.4.3.1 Development of Geospatial Workflow Modeling Languages and Tools

Based on the concept of multistage geospatial service chaining, various model languages have been proposed and modeling platforms have been developed. Chen et al. (2009) defined a geospatial abstract information model (AIM) language to describe logical models, which can be considered a new virtual geospatial product for process

knowledge sharing and reuse. A logical model has a directed graph expression as well as an XML presentation that can be instantiated and translated into an executable WS-BPEL model for reliable execution management. By adopting such technologies, Di (2004) developed GeoBrain, a geospatial knowledge building system based on web service technologies, to automate data discovery and facilitate geoprocessing workflow modeling. Gui et al. (2008) also proposed an abstract geospatial service chain model language, DDBASCM, by combining data-dependency directed graph and block structures. In DDBASCM, the data flow is represented using a directed graph structure, and the control flow and aggregated service are depicted as block structures by learning from the concept of a transition-bordered set in Petri Net. Based on DDBASCM and WS-BPEL, a geospatial web service chain visual modeling and execution platform called GeoChaining was developed, which integrates catalog-based geospatial resource searching, service chain visual modeling, execution status monitoring and data visualization (Wu et al. 2011, 2014). Sun et al. (2012) developed a task-oriented geoprocessing system called GeoPWTManager to design, execute, monitor and visualize the workflow. In GeoPWTManager, the entire modeling and execution process is ensured by the collaboration of three components, i.e., a task designer, a task executor and a task monitor. Based on GeoPWTManager, GeoJModelBuilder, an open source geoprocessing workflow tool, was developed by leveraging open standard, sensor web, geoprocessing service and OpenMI-compliant models (Jiang et al. 2017).

Although implementation technologies are continuously evolving and new tools will be developed, the demand for development of a domain-specific workflow modeling language that can explicitly describe the terminologies and geoprocessing knowledge in the geospatial domain is still high. Cloud-based services that integrate online resource discovery, visualization, automatic/semiautomatic modeling, model sharing and reuse will be a trend to facilitate DGIP workflow modeling and execution management.

6.4.3.2 Digital Earth Applications

Geospatial service chaining provides an agile and loosely coupled approach to arrange the cooperation of dispersed GIServices to achieve DGIP. Based on the aforementioned technologies and platforms, DGIP-supported earth science applications have been developed. For example, the ESA created a net primary productivity (NPP) workflow in its online collaboration platform, Service Support Environment (SSE), for repeatable estimates of the net flux of carbon over user-specified AOI areas using SPOT vegetation S10 data. There are also more than 30 DGIP workflow-based applications provided by 23 service partners from 10 countries, including for oil spill detection, fire risk, Kyoto protocol verification, desert locusts, land use, snow cover, tidal currents, and multiple catalog access. In GeoBrain, many DGIP workflow models have been developed based on the proposed logical modeling language. A landslide susceptibility model (Chen et al. 2009) that integrates terrain slope and aspect analysis services as well as landslide susceptibility analysis services has been

used to analyze landslide susceptibility in California, USA. GeoChaining (Wu et al. 2014) also provides workflow models by integrating third-party developed GIServices such as OpenRS (Guo et al. 2010) and GIServices developed by encapsulating the open-source GIS tool GRASS (https://grass.osgeo.org). A flood analysis model was developed to analyze flooding in the Boyang Lake area using remote sensing data before a flood, during flooding and after flooding. By developing web-based human-computer interaction interfaces using Rich Internet Application (RIA) technologies, workflow models involving human participation have also been developed in GeoSquare for educational and research purposes (Wu et al. 2015; Yang et al. 2016), including remote sensing image geometrical rectification and classification. Through integration with NASA world wind, GeoJModelBuilder (Jiang et al. 2017) also provides many hydrological models such as for water turbidity, watershed runoff, and drainage extraction.

In addition to applications comprised of geospatial service chaining, there are other forms of DGIP. For example, volunteer computing (VC) is a type of distributed computing that incorporates volunteered computing resources from individual persons and organizations. The VC usually adopts middleware architecture containing a client program that is installed and running on volunteer computers. VC has been successfully applied to many scientific research projects such as SETI@home (https://setiathome.berkeley.edu) and Folding@home(https://foldingathome.org). In the earth science domain, NASA launched a VC project named Climate@home to create a virtual supercomputer to model global climate research. This project utilizes worldwide computing resources to establish accuracy models for climate change prediction (Li et al. 2013).

Various applications have been developed, and the potential application scenarios are unlimited. As more GIServices for geoprocessing and big data analysis are developed using cloud computing and VC technologies, more interdisciplinary applications in earth science and social science will be developed.

6.4.4 Research Challenges and Future Directions

6.4.4.1 Communication Mechanism and Code Migration

Optimized network communication is critical for efficient and reliable DGIP because it relies on network communication for data transmission and service collaboration. The simple object access protocol (SOAP) is a widely used messaging protocol for exchanging information and conducting remote procedure calls (RPCs) in DCE using multiple lower-level transportation protocols such as HTTP, SMTP and TCP. SOAP is extensible to support functions such as security, message routing and reliability by compositing with web service specifications. SOAP supports multiple message exchange patterns (MEPs) such as one-way messages, request/respond mode and asynchronous messages. However, SOAP is not efficient in encoding due to its XML-based hierarchical envelope structure, for example, when transmitting vector data

represented in GML or a raster image formatted using base64 binary encoding. As a result, SOAP message transmission optimization technologies have been developed. Binary data code can be sent as multipart MIME documents in SOAP attachments, and XML-binary Optimized Packages (XOP) provide a reliable approach to refer external data in the SOAP messaging, as proposed in SOAP standard version 1.2.

The development of HTTP Representational State Transfer (RESTful) (Fielding 2000) brings new challenges for DGIP. The RESTful architecture style has been widely adopted in web application development (Pautasso et al. 2008). OGC GIService standards use RESTful APIs as the major interoperating approach. Considering this trend, service composition technologies and tools should support RESTful services. Compared with SOAP, RESTful is lightweight and stateless, but the security, routing and reliable message transmission are weakened. Therefore, making DGIP reliable and secure has become critically important. Robust flow control, exception handling and compensation mechanisms must be developed for both the workflow engine and the participating services.

Communication issues have also inspired new ideas and research directions. Geoprocessing usually involves a large data volume and intensive geo-computation. The intensive data transmission increases the workload of the network infrastructure, as well as those of the participating services and workflow coordinator, and makes time efficiency a troublesome issue. To improve the user experience for DGIP, an asynchronous execution status-tracking method has been developed (Wu et al. 2014). Version 2.0 of the OGC web processing service (WPS) standard officially supports asynchronous execution of geoprocessing by the conjunction of *GetStatus* and *GetResult* operations. The *GetStatus* operation provides status information of a processing job for query, and *GetResult* allows for the client to query the result of a processing job. Through an asynchronous mechanism, a geoprocessing workflow engine can actively and instantly push the latest execution status of dispersed services to clients. Data transmission may also introduce data security risks, especially for classified or sensitive data. As the volume of software programs may be much smaller than the data volume, researchers proposed the idea of code migration. However, it is not easy to migrate code in heterogeneous systems due to the complex dependency of software packages. VC provides an alternative solution by installing a specified client to set up a unified runtime environment, e.g., BOINC (https://boinc.berkeley.edu). This problem is eliminated in a clustered computing environment because the computing nodes are equipped with the same operating system and distributed computing framework and thus the code can be migrated smoothly. For example, the high-performance frameworks introduced in Sect. 6.2, e.g., Apache Hadoop and Spark, migrate codes to computing nodes according to the locality of the dataset in the distributed file system to avoid IO overhead and optimize computing performance.

6.4.4.2 Quality-Aware Service Chain Instantiation

As global providers deliver more GIServices with similar functions but diverse quality, it has become challenging to select appropriate service instances from similar

service candidates. To enable quality-aware service chain instantiation, quality evaluation methods and mathematical planning methods must be developed (Hu et al. 2019b). Quality evaluation assesses the fitness of individual participating services or aggregated services according to user quality requirements, and mathematical planning assists the service instance selection for each individual participating service by considering the overall quality of the service chain.

Multiple quality dimensions such as time efficiency and reliability must be leveraged to evaluate the quality of a participating service. Operations research methods such as multiple attribute decision making (MADM) and the analytic hierarchy process (AHP) provide solutions for quality dimension integration (Zeng et al. 2003). However, the control-flow and data-flow structures must be considered to determine the aggregated quality of a service chain (Jaeger et al. 2004).

In terms of service chaining, quality metrics have different aggregation behaviors under different flow structures (Aalst et al. 2003). For example, the total response time of a service chain with a sequential control-flow structure is the sum of the response times of all the participating services, and the total reliability is calculated by multiplying the availability of all the participating services. Quality computation can be more complicated in service chains with nested flow structures. If only the quality status of participating services is considered and the workflow structure is ignored, then the overall optimization of a service chain cannot be guaranteed (Jaeger et al. 2004; Gui et al. 2009; Hu et al. 2019a, b), especially when multiple quality metrics must be balanced.

To support quality-aware geospatial service chain instantiation, sophisticated GIS-ervice selection methods must be developed. Mathematical programming approaches such as Linear Programming (LP) can be used in service chains (Zeng et al. 2003; Gui et al. 2009) with a limited number of participating services. When the scale of the service chain increases, these methods become less efficient due the computing complexity. Furthermore, LP can only provide one optimized solution in the planning stage, which may not be optimal when one of the quality metrics slightly changes, since service runtime and network environments are typically mutable. Evolutionary methods (Canfora et al. 2005) such as genetic algorithms and swarm intelligent algorithms provide strong search capabilities and robustness in dynamic situations (Jula et al. 2014) and can be applied for geospatial service chain optimization. Considering the nature of complex flow structures and high dimensions of the quality metrics of a geospatial service chain, more research on quality evaluation and GIService selection must be conducted.

6.4.4.3 Semantic-Aided Automatic Service Chaining

With the development of artificial intelligence (AI) and semantic web technologies, automatic service chaining has been a research hotspot for many years and is still evolving. The goal of automatic service chaining is to make the computer capable of discovering web service resources and automatically building the service chain

according to the requirements and constraints of the end user. In contrast to quality-aware service chain instantiation, there is no logical model available in advance for automatic service chaining. Thus, the computer must build the logical chain and instantiate it upon domain knowledge and the timeliness of the service resources, i.e., whether the service instance or data provider is available or not. To achieve this goal, a formal description of knowledge is required. The development of semantic web, ontology web language (OWL) and domain ontologies facilitates GIService semantic markups. For example, ontology-based description languages and rule languages are used for semantic discovery and geospatial service chaining (Lutz and Klien 2006; Yue et al. 2009), including semantic markup for web services (OWL-S), web service modeling ontology (WSMO), description logics (DL) and first-order logic (FOL). GeoBrain provides a web-based semiautomatic chaining environment by allowing for end-users to participate in human-computer interaction during the backwards reasoning (Di 2004; Han et al. 2011). The degree of suitability for candidate workflows is calculated by using the semantic similarity to support semiautomatic chaining (Hobona et al. 2007).

Semantic-aided chaining approaches have been developed and verified in laboratory environments; however, more research must be conducted to make them feasible in real-world applications. Currently, semantic markups for describing content, functions or prerequisites lack in most online-accessible geospatial resources. In addition, spatial data infrastructures (SDIs) and geoportals such as GEOSS clearinghouse, Data.gov, and INSPIRE do not provide semantic-aware discovery functions. The challenges include determining how to provide a semantic-enabled metadata Registry Information Model (RIM) for GIService semantic description, retrieval and validation (Qi et al. 2016; Zhang et al. 2017). W3C semantic standards such as the resource description framework (RDF) and OWL-S provide promising solutions for describing domain knowledge, enabling intelligent and efficient service discovery. However, these semantic languages must be linked with existing metadata standards in global SDIs (Gui et al. 2013). From the chain modeling perspective, AI reasoning technologies require further development to enable automatic and intelligent chaining. The rapid development of knowledge graph and mining technologies may provide a potential solution, which has been widely adopted in domain knowledge modeling and reasoning (Lin et al. 2015). Furthermore, to conduct DGIP-supported geoscience data analysis using heterogeneous Earth observation and socioeconomic data, we need to establish and advocate for standardization of the Discrete Global Grid System (DGGS) (Mahdavi-Amiri et al. 2015). It is critically important to promote heterogeneous earth science data fusion and interoperability, and the related standards and data models should be integrated into global SDIs (Purss et al. 2017).

6.5 Discussion and Conclusion

Geospatial information processing and computing technologies are essential for Digital Earth, as they enable various Digital Earth applications by turning geospatial

data into information and knowledge. By identifying the challenges of geospatial data manipulation in the big data era, including massive *volume*, *heterogeneous*, and *distributed*, this chapter introduced three population technologies for geospatial data processing: high-performance computing, online geoprocessing, and distributed geoprocessing. Each of the three technologies focuses on addressing a specific challenge, though there are some overlaps. High-performance computing primarily deals with the *volume* challenge by solving data- and computing-intensive problems in parallel. Online geoprocessing tackles the *heterogeneous* challenge through standardized and interoperable geospatial web services and web APIs. Distributed geoprocessing addresses the *distributed* challenge by processing geospatial data and information in a distributed computing environment. The fundamental concepts, principles, and key techniques of the three technologies were elaborated in detail. Application examples in the context of Digital Earth were also provided to demonstrate how each technology has been used to support geospatial information processing. Although the three technologies are relatively mature and have a broad range of applications, research challenges have been identified and future research directions are envisioned for each technology to better support Digital Earth.

For high-performance computing (Sect. 6.2), one research challenge and direction is to continue the efforts to parallelize existing serial algorithms in spatial database and spatial data mining functions considering the dependence and interactions between data and problem partitions. Another direction is to develop new parallel algorithms to mine geospatial big data in the network space instead of in Euclidean space, as many spatial processes and interactions often occur in the network space. The third direction is to explore new and efficient computing methods to identify patterns from massive volumes of geospatial data considering the spatial heterogeneity. For online geoprocessing (Sect. 6.3), the main challenge is the lack of opacity in the services, data, and tool interfaces. This hinders the interoperability among the diverse services and creates a challenge when a problem needs to be solved by processing multi-sourced data using different services and tools. One promising solution is to incorporate semantics into web services to increase the interoperability among heterogeneous resources. In semantic web service research, three research directions are envisioned to achieve a semantically interoperable and intelligent Digital Earth: linked data and automated typing, automated workflow composition, and question answering. For distributed geoprocessing (Sect. 6.4), one challenge arises from reliability and security concerns. More efforts are needed to ensure a reliable and secure distributed computing environment considering aspects of the flow control, exception handling, compensation mechanism, and quality-aware service chains. The large volumes of geospatial data also lead to challenges in moving distributed data to the processing tools/services. Although moving code to data (code migration) is a promising solution, further research is needed to migrate code among the heterogeneous systems due to the complex dependency of software packages. In addition, more efforts are needed to move semantic-aided automatic service chaining techniques from the laboratory environment to real-world applications.

Lastly, the Digital Earth reference framework (Fig. 6.1) aims to integrate heterogeneous data sources with a harmonious high-level data model of the Earth so that data

can be handled seamlessly with different tools, protocols, technologies. Currently, most of the tools use the framework of traditional coordinate systems such as the geographic coordinate system based on the continuous latitude and longitude or the projected coordinate system that projects the curved Earth surface to a flat surface. Although the traditional coordinate systems have been successful, another reference framework called the discrete global grid system (DGGS, see Chap. 2 *Digital Earth Platforms* for more details) is considered better for data associated with the curved heterogeneous surface of the Earth (Sabeur et al. 2019). We believe that the DGGS will play an increasingly important role in geospatial information processing in the big data era because (1) the DGGS provides a single and relatively simple framework for the seamless integration of heterogeneous distributed global geospatial data from different sources and domains; (2) the DGGS works with high-performance computing to handle big data extremely well because data managed with the DGGS is already decomposed into discrete domains and can be processed in parallel; and (3) by providing a single framework, the DGGS benefits interoperability among different tools and geoprocessing technologies and is a promising solution to build a semantically interoperable Digital Earth. However, most available analysis tools are designed to work with the traditional reference framework. Thus, more efforts are needed to design and develop storage mechanisms, spatiotemporal indexes, computing algorithms, and big data computing platforms that are compatible with the DGGS framework.

References

Aalst WMPVD, Hofstede AHMT, Kiepuszewski B et al (2003) Workflow patterns. Distrib Parallel Databases 14(1):5–51

Aghajarian D, Puri S, Prasad S (2016) GCMF: an efficient end-to-end spatial join system over large polygonal datasets on GPGPU platform. In: Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems, Burlingame, CA, 31 October–3 November 2016. ACM, New York, p 18

Aji A, Wang F, Vo H et al (2013) Hadoop-GIS: a high performance spatial data warehousing system over MapReduce. Proc VLDB Endow 6(11):1009–1020

Alameh N (2003) Chaining geographic information web services. IEEE Internet Comput 7(5):22–29

Alper P, Belhajjame K, Goble CA et al (2014) LabelFlow: exploiting workflow provenance to surface scientific data provenance. In: International provenance and annotation workshop, Cologne, Germany, 9–13 June 2015. Springer, Heidelberg, pp 84–96

Barua S, Alhajj R (2007) Parallel wavelet transform for spatio-temporal outlier detection in large meteorological data. In: International conference on intelligent data engineering and automated learning, Birminghan, UK, 16–19 December 2007. Springer, Heidelber, pp 684–694

Baumann P (2010) The OGC web coverage processing service (WCPS) standard. GeoInformatica 14(4):447–479

Beek MT, Bucchiarone A, Gnesi S (2007) Web service composition approaches: from industrial standards to formal methods. In: Second international conference on internet and web applications and services (ICIW'07), Morne, Mauritius, 13–19 May 2007

Bishr Y (1998) Overcoming the semantic and other barriers to GIS interoperability. Int J Geogr Inf Sci 12(4):299–314

Bonomi F, Milito R, Zhu J et al (2012) Fog computing and its role in the internet of things. In: Proceedings of the first edition of the MCC workshop on mobile cloud computing, Helsinki, Finland, 17 August 2012. ACM, New York, pp 13–16

Brauner J (2015) Formalizations for geooperators-geoprocessing in spatial data infrastructures. http://tud.qucosa.de/api/qucosa%3A28979/attachment/ATT-1. Accessed 11 Jul 2019

Canfora G, Penta MD, Esposito R et al (2005) An approach for QoS-aware service composition based on genetic algorithms. In: Proceedings of the 7th annual conference on genetic and evolutionary computation, Washington DC, USA, 25–29 June 2005. ACM, New York, pp 1069–1075

Chandola V, Vatsavai RR (2011) A scalable gaussian process analysis algorithm for biomass monitoring. Stat Anal Data Min ASA Data Sci J 4(4):430–445

Chen A, Di L, Wei Y et al (2009) Use of grid computing for modeling virtual geospatial products. Int J Geogr Inf Sci 23(5):581–604

Cressie N, Wikle C (2015) Statistics for spatio-temporal data. John Wiley and Sons, Hoboken, New Jersey

de Ravé EG, Jiménez-Hornero FJ, Ariza-Villaverde AB et al (2014) Using general-purpose computing on graphics processing units (GPGPU) to accelerate the ordinary kriging algorithm. Comput Geosci 64:1–6

Di, L (2004) GeoBrain-a web services based geospatial knowledge building system. In: Proceedings of NASA earth science technology conference, Palo Alto, CA, 22–24 June 2004

Di, L., Zhao P., Yang W., and Yue P., 2006. Ontology-driven Automatic Geospatial-Processing Modeling based on Web-service Chaining, Proceedings of the Sixth Annual NASA Earth Science Technology Conference. June 27-29, 2006. College Park, MD, USA 7p

Eldawy A, Elganainy M, Bakeer A et al (2015) Sphinx: distributed execution of interactive sql queries on big spatial data. In: Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems, Seattle, Washington, 3–6 November 2015. ACM, New York, p 78

ESRI (2018) GIS tools for hadoop by Esri. http://esri.github.io/gis-tools-for-hadoop. Accessed 11 Jul 2019

Fielding R (2000) Architectural styles and the design of network-based software architectures. Doctoral Dissertation, University of California

Fitzner D, Hoffmann J, Klien E (2011) Functional description of geoprocessing services as conjunctive datalog queries. GeoInformatica 15(1):191–221

Friis-Christensen A, Lucchi R, Lutz M et al (2009) Service chaining architectures for applications implementing distributed geographic information processing. Int J Geogr Inf Sci 23(5):561–580

Gandhi V, Celik M, Shekhar S (2006) Parallelizing multiscale and multigranular spatial data mining algorithms. In: Partitioned global address space programming models conference, Washington DC, 3–4 October 2006

Gangemi A, Presutti V (2009) Ontology design patterns. In: Staab S, Studer R (eds) Handbook on ontologies. Springer, Heidelberg, pp 221–243

Gao S, Goodchild MF (2013) Asking spatial questions to identify GIS functionality. In: 2013 fourth international conference on computing for geospatial research and application, San Jose, CA, 22–24 July 2013. IEEE, New Jersey, pp 106–110

Gil Y (2007) Workflow composition: semantic representations for flexible automation. In: Taylor IJ, Deelman E, Gannon DB et al (eds) Workflows for e-science: scientific workflows for grids. Springer, Heidelberg, pp 244–257

Gong J, Wu H, Zhang T et al (2012) Geospatial service web: towards integrated cyberinfrastructure for GIScience. Geo-Spat Inf Sci 15(2):73–84

Goodchild M (2007) Citizens as voluntary sensors: spatial data infrastructure in the world of web 2.0. Int J Spat Data Infrastruct Res 2:24–32

Gui Z, Wu H, Chen Y et al (2009) The research on QoS assessment and optimization for geospatial service chain. In: 2009 17th international conference on geoinformatics, Fairfax, VA, 12–14 August 2009

Gui Z, Wu H, Wang Z (2008) A data dependency relationship directed graph and block structures based abstract geospatial information service chain model. In: Proceedings of the 2008 fourth international conference on networked computing and advanced information management, Gyeongju, South Korea, 2–4 September 2008

Gui Z, Yang C, Xia J et al (2013) A performance, semantic and service quality-enhanced distributed search engine for improving geospatial resource discovery. Int J Geogr Inf Sci 27(6):1109–1132

Guo W, Gong J, Jiang W et al (2010) OpenRS-cloud: a remote sensing image processing platform based on cloud computing environment. Sci China Technol Sci 53(1):221–230

Hamadi R, Benatallah B (2003) A Petri net-based model for web service composition. In: Schewe K, Zhou X (eds) Proceedings of the 14th Australasian database conference on database Technologies, Adelaide, Australia, pp 191–200

Han W, Di L, Zhao P et al (2011) GeoBrain online analysis system: an SOA-based geospatial web portal. In: Zhao P, Di L (eds) Geospatial web services: advances in information interoperability. IGI Global, Pennsylvania, pp 455–474

Healey R, Dowers S, Gittings B et al (1997) Parallel processing algorithms for GIS. CRC Press, Florida

Hobona G, Fairbairn D, James P (2007) Semantically-assisted geospatial workflow design. In: Proceedings of the 15th annual ACM international symposium on advances in geographic information systems, Seattle, Washington, 7–9 November 2007. ACM, New York, pp 194–201

Hofer B, Granell C, Bernard L (2018) Innovation in geoprocessing for a digital earth. Int J Digit Earth 11(1):3–6

Hofer B, Mäs S, Brauner J et al (2017) Towards a knowledge base to support geoprocessing workflow development. Int J Geogr Inf Sci 31(4):694–716

Höffner K, Lehmann J, Usbeck R (2016) CubeQA—question answering on RDF data cubes. In: Groth P, Simperl E, Gray A et al (eds) The semantic web – ISWC 2016, Kobe, Japan, 17–21 October 2016. Lecture Notes in Computer Science. Springer, Heidelberg, pp 325–340

Hu F, Li Z, Yang C et al (2019a) A graph-based approach to detecting tourist movement patterns using social media data. Cartogra Geogr Inf Sci 46(4):368–382

Hu K, Gi Z, Cheng X et al (2019b) The concept and technologies of quality of geographic information service: improving user experience of GIServices in a distributed computing environment. ISPRS Int J Geo-Inf 8(3):118

Huang X, Wang C, Li Z (2018) Reconstructing flood inundation probability by enhancing near real-time imagery with real-time gauges and tweets. IEEE Trans Geosci Remote Sens 56(8):4691–4701

Huang Y, Shekhar S, Xiong H (2004) Discovering colocation patterns from spatial data sets: a general approach. IEEE Trans Knowl Data Eng 16(12):1472–1485

ISO 19119 (2002) International standard ISO 19119: geographic information – services. ISO, Geneva, Switzerland

Jaeger MC, Rojec-Goldmann G, Muhl G (2004) QoS aggregation for web service composition using workflow patterns. In: Proceedings of eighth IEEE international enterprise distributed object computing conference (EDOC 2004). Monterey, CA, 24 September 2004. IEEE, New Jersey, pp 149–159

Janowicz K, Van Harmelen F, Hendler J et al (2014) Why the data train needs semantic rails. AI Mag 36(1):5–14

Jiang Z, Li Y, Shekhar S et al (2017) Spatial ensemble learning for heterogeneous geographic data with class ambiguity: a summary of results. In: Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems, Redondo Beach, CA, USA, 7–10 November 2017. ACM, New York, pp 23–32

Jiang Z, Shekhar S, Zhou X et al (2015) Focal-test-based spatial decision tree learning. IEEE Trans Knowl Data Eng 27(6):1547–1559

Jones R, Cornford D, Bastin L (2012) UncertWeb processing service: making models easier to access on the web. Trans GIS 16(6):921–939

Jula A, Sundararajan E, Othman Z (2014) Cloud computing service composition: a systematic literature review. Expert Syst Appl 41(8):3809–3824
Kazar B, Shekhar S, Lilja D et al (2004) A parallel formulation of the spatial auto-regression model for mining large geo-spatial datasets. In: SIAM international conference on data mining workshop on high performance and distributed mining (HPDM2004), Florida, 22–24 April 2004
Klien E, Lutz M, Kuhn W (2006) Ontology-based discovery of geographic information services—an application in disaster management. Comput Environ Urban Syst 30(1):102–123
Kuhn W (2012) Core concepts of spatial information for transdisciplinary research. Int J Geogr Inf Sci 26(12):2267–2276
Kuhn W, Ballatore A (2015) Designing a language for spatial computing. In: Bacao F, Santos M, Painho M (eds) AGILE 2015, Springer, Heidelberg, pp 309–326
Kuhn W, Kauppinen T, Janowicz K (2014) Linked data - a paradigm shift for geographic information science. In: International conference on geographic information science, Springer, Heidelburg, 3–6 June 2014
Lamprecht AL (2013) User-level workflow design: a bioinformatics perspective. Springer, Heidelburg
Lara R, Roman D, Polleres A et al (2004) A conceptual comparison of WSMO and OWL-S. In: Zhang L, Jeckle M (eds) Web services, Springer, Heidelberg, pp 254–269
Li Y, Shekhar S (2018) Local co-location pattern detection: a summary of results. In: LIPIcs-Leibniz international proceedings in informatics, Melbourne, Australia, 28–31 August 2018
Li Z, Huang Q, Jiang Y et al (2019) SOVAS: a scalable online visual analytic system for big climate data analysis. Int J Geogr Inf Sci 1–22. https://doi.org/10.1080/13658816.2019.1605073
Li Z, Wang C, Emrich CT et al (2018) A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods. Cartogra Geogr Inf Sci 45(2):97–110
Li J., Li, Z., Sun M., Liu K. (2013). Cloud-enabling Climate@Home. In Yang C., Huang Q., Li Z., Xu C., Liu K. (Eds.), Spatial cloud computing: a practical approach (pp. 143–160). CRC Press/Taylor & Francis
Lin, J. J. (2002, May). The Web as a Resource for Question Answering: Perspectives and Challenges. In LREC. Available at: https://cs.uwaterloo.ca/~jimmylin/publications/Lin_LREC2002.pdf
Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015, February). Learning entity and relation embeddings for knowledge graph completion. In Twenty-ninth AAAI conference on artificial intelligence
Lopez-Pellicer et al. (2012):Lopez-Pellicer, F. J., Rentería-Agualimpia, W., Béjar, R., Muro-Medrano, P. R., & Zarazaga-Soria, F. J. (2012). Availability of the OGC geoprocessing standard: March 2011 reality check. Computers & Geosciences, 47, 13–19
Lutz, M., & Klien, E. (2006). Ontology-based retrieval of geographic information. International Journal of Geographical Information Science, 20(3), 233–260
Mahdavi-Amiri A, Alderson T, Samavati F (2015) A survey of digital earth. Comput Graph 53:95–117
Martin Y, Li Z, Cutter SL (2017) Leveraging Twitter to gauge evacuation compliance: spatiotemporal analysis of Hurricane Matthew. PLoS ONE 12(7):e0181701
Miller HJ, Goodchild MF (2015) Data-driven geography. GeoJournal 80(4):449–461
Müller M (2015) Hierarchical profiling of geoprocessing services. Comput Geosci 82:68–77
Naujokat S, Lamprecht AL, Steffen B (2012) Loose programming with PROPHETS. In: International conference on fundamental approaches to software engineering, Tallinn, Estonia, 24 March–1 April 2012. Springer, Heidelberg, pp 94–98
Okabe A, Sugihara K (2012) Spatial analysis along networks: statistical and computational methods. John Wiley & Sons, New Jersey
Ouksel AM, Sheth A (1999) Semantic interoperability in global information systems. ACM Sigmod Rec 28(1):5–12
Pang LX, Chawla S, Scholz B et al (2013) A scalable approach for LRT computation in GPGPU environments. In: Asia-pacific web conference, Sydney, Australia, 4–6 April 2013, Springer, Berlin, Heidelberg, pp 595–608

Pautasso C, Zimmermann O, Leymann F (2008) Restful web services vs. "big"' web services: making the right architectural decision. In: Proceedings of the 17th international conference on world wide web pages, Beijing, China, 21–25 April 2008
Peltz C (2003) Web services orchestration and choreography. Computer 36(10):46–52
Pesquer L, Cortés A, Pons X (2011) Parallel ordinary kriging interpolation incorporating automatic variogram fitting. Comput Geosci 37(4):464–473
Prasad SK, McDermott M, Puri S et al (2015) A vision for GPU-accelerated parallel computation on geo-spatial datasets. SIGSPATIAL Spec 6(3):19–26
Prasad SK, Shekhar S, McDermott M et al (2013) GPGPU-accelerated interesting interval discovery and other computations on GeoSpatial datasets: a summary of results. In: Proceedings of the 2nd ACM SIGSPATIAL international workshop on analytics for big geospatial data, Orlando, FL, 4 November 2013. ACM, New York, pp 65–72
Purss M, Gibb R, Samavati F et al (2017) Discrete global grid systems abstract specification–Topic 21. In: Purss M (ed) Open Geospatial Consortium
Qi K, Gui Z, Li Z et al (2016) An extension mechanism to verify, constrain and enhance geoprocessing workflows invocation. Trans GIS 20(2):240–258
Rao J, Su X (2005) A survey of automated web service composition methods. In: International workshop on semantic web services and web process composition, Springer, Heidelberg, 6 July 2004
Rey SJ, Anselin L, Pahle R et al (2013) Parallel optimal choropleth map classification in PySAL. Int J Geogr Inf Sci 27(5):1023–1039
Rizki P, Eum J, Lee H et al (2017) Spark-based in-memory DEM creation from 3D LiDAR point clouds. Remote Sens Lett 8(4):360–369
Sabeur Z, Gibb R, Purss M (2019) Discrete global grid systems SWG. http://www.opengeospatial.org/projects/groups/dggsswg. Accessed 13 Mar 2019
Scheider S, Ballatore A (2018) Semantic typing of linked geoprocessing workflows. Int J Digit Earth 11(1):113–138
Scheider S, Ballatore A, Lemmens R (2019) Finding and sharing GIS methods based on the questions they answer. Int J Digit Earth 12(5):594–613
Scheider S, Gräler B, Pebesma E et al (2016) Modeling spatiotemporal information generation. Int J Geogr Inf Sci 30(10):1980–2008
Scheider S, Huisjes MD (2019) Distinguishing extensive and intensive properties for meaningful geocomputation and mapping. Int J Geogr Inf Sci 33(1):28–54
Scheider S, Kuhn W (2015) How to talk to each other via computers: semantic interoperability as conceptual imitation. In: Zenker F, Gärdenfors P (eds) Applications of conceptual spaces. Springer, Heidelberg, pp 97–122
Scheider S, Ostermann FO, Adams B (2017) Why good data analysts need to be critical synthesists. Determining the role of semantics in data analysis. Future Gener Comput Syst 72:11–22
Schnase JL, Duffy DQ, Tamkin GS et al (2017) MERRA analytic services: meeting the big data challenges of climate science through cloud-enabled climate analytics-as-a-service. Comput Environ Urban Syst 61:198–211
Shekhar S, Chawla S (2003) Spatial databases: a tour. Prentice Hall, Saddle River, NJ
Shekhar S, Evans MR, Kang JM et al (2011) Identifying patterns in spatial information: a survey of methods. Wiley Interdiscip Rev Data Min Knowl Discov 1(3):193–214
Shekhar S, Jiang Z, Ali YR et al (2015) Spatiotemporal data mining: a computational perspective. ISPRS Int J Geo-Inf 4(4):2306–2338
Shekhar S, Lu C-T, Zhang P (2003) A unified approach to detecting spatial outliers. GeoInformatica 7(2):139–166
Shekhar S, Ravada S, Chubb D et al (1998) Declustering and load-balancing methods for parallelizing geographic information systems. IEEE Trans Knowl Data Eng 10(4):632–655
Shekhar S, Ravada S, Kumar V et al (1996) Parallelizing a GIS on a shared address space architecture. Computer 29(12):42–48

Shi W, Cao J, Zhang Q et al (2016) Edge computing: vision and challenges. IEEE Internet Things J 3(5):637–646
Stasch C, Pross B, Gräler B et al (2018) Coupling sensor observation services and web processing services for online geoprocessing in water dam monitoring. Int J Digit Earth 11(1):64–78
Sun Z, Yue P, Di L (2012) GeoPWTManager: a task-oriented web geoprocessing system. Comput Geosci 47:34–45
Tang W, Feng W, Jia M (2015) Massively parallel spatial point pattern analysis: Ripley's K function accelerated using graphics processing units. Int J Geogr Inf Sci 29(3):412–439
Treiblmayr M, Scheider S, Krüger A et al (2012) Integrating GI with non-GI services—showcasing interoperability in a heterogeneous service-oriented architecture. GeoInformatica 16(1):207–220
Vahedi B, Kuhn W, Ballatore A (2016) Question-based spatial computing—a case study. In: Sarjakoski T, Santos M, Sarjakoski L (eds) Geospatial data in a changing world. Lecture notes in geoinformation and cartography, Springer, Heidelberg, pp 37–50
Wagemann J (2016) OGC web coverage service tutorial. Zenodo https://doi.org/10.5281/zenodo.205442
Wagemann J, Clements O, Figuera RM et al (2018) Geospatial web services pave new ways for server-based on-demand access and processing of big earth data. Int J Digit Earth 11(1):7–25
Wang S, Cowles MK, Armstrong MP (2008) Grid computing of spatial statistics: using the TeraGrid for G(d) analysis. Concurr Comput Pract Exp 20(14):1697–1720
Wiemann S, Karrasch P, Bernard L (2018) Ad-hoc combination and analysis of heterogeneous and distributed spatial data for environmental monitoring – design and prototype of a web-based solution. Int J Digit Earth 11(1):79–94
Wu H, Li Z, Zhang 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. Comput Geosci 37(4):485–494
Wu H, You L, Gui Z et al (2014) FAST: a fully asynchronous and status-tracking pattern for geoprocessing services orchestration. Comput Geosci 70:213–228
Wu H, You L, Gui Z et al (2015) GeoSquare: collaborative geoprocessing models' building, execution and sharing on Azure Cloud. Ann GIS 21(4):287–300
Xu Z, Guan J, Zhou J (2015) A distributed inverse distance weighted interpolation algorithm based on the cloud computing platform of Hadoop and its implementation. In: 2015 12th international conference on fuzzy systems and knowledge discovery (FSKD), Zhangjiajie, China, 15–17 August 2015. IEEE, New Jersey, pp 2412–2416
Yang C, Li W, Xie J et al (2008) Distributed geospatial information processing: sharing distributed geospatial resources to support digital earth. Int J Digit Earth 1(3):259–278
Yang Z, Cao J, Hu K et al (2016) Developing a cloud-based online geospatial information sharing and geoprocessing platform to facilitate collaborative education and research. In: The international archives of photogrammetry, remote sensing and spatial information sciences, XLI-B6, 3–7. XXIII ISPRS Congress, Prague, Czech Republic, 12–19 July 2016
Yao X, Mokbel MF, Alarabi L et al (2017) Spatial coding-based approach for partitioning big spatial data in Hadoop. Comput Geosci 106:60–67
Yu J, Wu J, Sarwat M (2015) GeoSpark: a cluster computing framework for processing large-scale spatial data. In: Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems, Seattle, Washington, 3–6 November 2015. ACM, New York, p 70
Yue P, Baumann P, Bugbee K et al (2015) Towards intelligent GIServices. Earth Sci Inform 8(3):463–481
Yue P, Di L, Yang W et al (2007) Semantics-based automatic composition of geospatial web service chains. Comput Geosci 33(5):649–665
Yue P, Di L, Yang W et al (2009) Semantic web services-based process planning for earth science applications. Int J Geogr Inf Sci 23(9):1139–1163

Zeng L, Benatallah B, Dumas M et al (2003) Quality driven web services composition. In: Proceedings of the 12th international conference on world wide web, Budapest, Hungary, 20–24 May 2003. ACM, New York, pp 411–421

Zhang M, Yue P, Wu Z et al (2017) Model provenance tracking and inference for integrated environmental modelling. Environ Model Softw 96:95–105

Zhao L, Chen L, Ranjan R et al (2016) Geographical information system parallelization for atial big data processing: a review. Clust Comput 19(1):139–152

Zhenlong Li is an Assistant Professor with the Department of Geography at the University of South Carolina, where he leads the Geoinformation and Big Data Research Laboratory. His primary research focuses on geospatial big data analytics, high performance computing, and CyberGIS with applications to disaster management, climate analysis, and human mobility. He serves as the Chair of the Association of American Geographers CyberInfrastructure Specialty Group.

Zhipeng Gui is Associate Professor of Geographic Information Science at School of Remote Sensing and Information Engineering, Wuhan University. His research interest is geospatial service chaining, high-performance spatiotemporal data mining and geovisual analytics. He serves as the Co-chair of International Society for Photogrammetry and Remote Sensing (ISPRS) Working Group V/4—Web-based Resource Sharing for Education and Research.

Barbara Hofer is GIScientist and works as Associate Professor at the Interfaculty Department of Geoinformatics—Z_GIS at the University of Salzburg, Austria. She is co-leader of the research group "geographic information infrastructure", which relates to her interest in the field of spatial data infrastructures, Digital Earth, online geoprocessing and reproducible research. Currently, Barbara also serves as councillor for AGILE—the Association of Geographic Information Laboratories in Europe.

Yan Li is a Ph.D. student in Computer Science at the University of Minnesota. His research interest includes spatial computing, data mining, and machine learning. He got his bachelor's degree in Remote Sensing at Wuhan University, China, and master's degree in Geography at the University of Tennessee.

Simon Scheider is an Assistant Professor in Geographic Information Science at the Department of Human Geography and Spatial Planning, University Utrecht. His research focuses on understanding the concepts underlying spatio-temporal data, including reference systems, fields, places, objects, events, trajectories and their relationship to human activities. He uses linked data and Semantic Web technology to support GIS analysts and to automate analytic workflows.

Shashi Shekhar is a Professor in Computer Science at the University of Minnesota. He is a spatial database and data mining researcher and a GeoInformatica co-Editor-In-Chief, co-authored a textbook Spatial Database, and co-edited an Encyclopedia of GIS. Honors include IEEE-CS Technical Achievement Award, AAAS Fellow, and IEEE Fellow.