AGILE: GIScience Series, 3, 49, 2022. https://doi.org/10.5194/agile-giss-3-49-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

New Generation of Geospatial Clearinghouse Networks

Morteza Omidipour, Ara Toomanian, Najmeh Neysani Samany

Department of Remote Sensing and GIS, Faculty of Geography, University of Tehran, Tehran, Iran

Abstract. Clearinghouse is a major component of Spatial Data Infrastructures (SDIs), and plays a key role in the discovery, access, and exchange of geospatial data via a distributed network. Traditionally, the main emphasis of Geospatial Clearinghouse Networks (GCNs) are the availability and accessibility of data. Such approaches, however, have failed to address real-world problem solving and decision-making. The main problem is the lack of a general solution for discovery, access, and knowledge exchange in integrated and online infrastructure. Due to the rapid advancements in technology stacks and innovations in GIS communities, many limitations have been resolved. A new generation of GCNs facilitates and coordinates accessing, and exchange of spatial knowledge between stakeholders. In this paper, we propose the structure of a new generation of GCNs that facilitate the discovery and dissemination of spatial knowledge.

Keywords. geospatial clearinghouse network, spatial data infrastructure, spatial knowledge infrastructure, knowledge discovery, web service

1 Introduction

Sharing geospatial data between other communities need a distributed network named Geospatial Clearinghouse Network (GCN). The method used to set up such networks depends on various technological, legal, organizational, cultural, commercial, and managerial contexts (Toomanian, 2012). A GCN is based on a distributed network, data producers, and users who communicate electronically with each other (Nebert, 2004). The network allows users to obtain information about the availability of data, the way to access them. Generally, the primary goal of geospatial clearinghouses are to provide access to digital spatial data through descriptive information. This descriptive information, known as metadata, is collected in a standard format to facilitate the discovery and accessing of data. Clearinghouse allows individual agencies, consortia, or other communities to band together and promotes their available digital spatial data (U.S. Federal Geographic Data Committee (FGDC), 2022).

Today, geospatial data are growing massively at an exponential rate, and GIscience has been moved from a data-poor to a data-rich age (Miller and Han, 2009). Although the vast availability of geospatial data provides more opportunities for a better understanding of complex phenomena, challenges related to the applicability and usability of the data are remained and need to be addressed (Omidipoor, 2018). In this context, there is a vital need for effective and efficient methods to extract knowledge from the voluminous geospatial data (Longley et al., 2015).

A knowledge-based GCN is a type of clearinghouse that not only carries out direct access to the geospatial data, but also provides utilities for finding, accessing, and sharing knowledge. Knowledge-based GCNs provide a bright vision for geospatial data communities.

To enhance the functionality of GCNs, in this paper, a knowledge-based GCN solution is proposed. This paper first gives a brief overview to the recent evolution and different generations of GCNs. The main components of current GCNs also described in this section (see Sec.2). In Sec.3, components of a Knowledge-based GCN are described. Finally, in Sec.4, properties of the knowledge-based GCN are discussed and conclusions are provided.

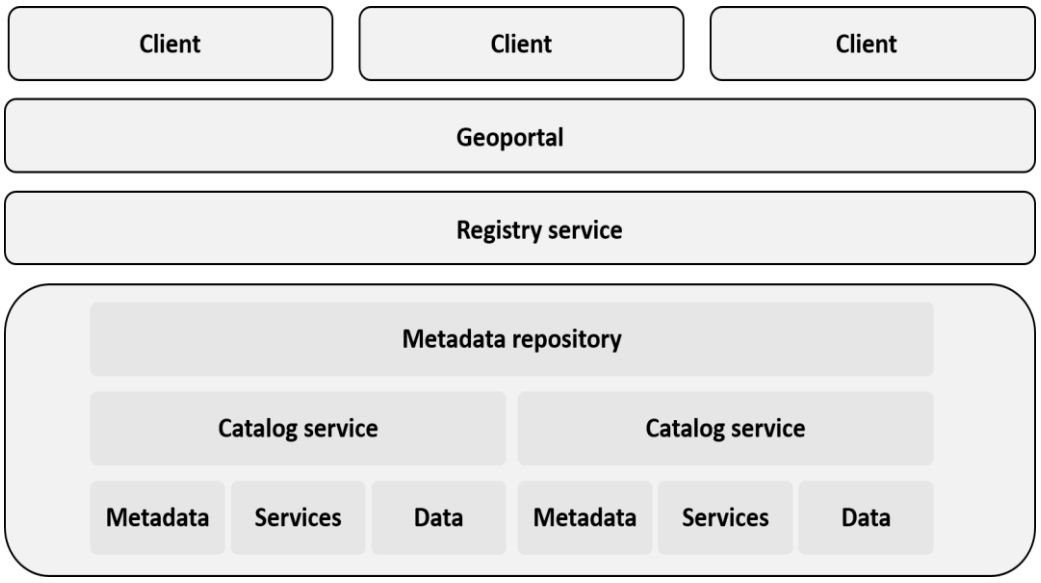

Figure 1. General structure of current generation of GCNs.

2 General components of current GCNs

GCNs consists of a set of components and the relationship between them that has evolved with the advancement of technology. Based on these changes, different generations GCNs have been developed. Over time, various structures and architecture have been used for developing GCNs. Mansourian et al. (2011), draw a distinction between two broad categories of GCNs. The early generation of GCNs provided users with either information about the data, or a link to the data producer website and hints for accessing geospatial data. The development of technology and the advancement of geospatial web services set up a new generation of GCNs that are based on geoportals as a gateway.

Besides the mentioned GCNs, some efforts have been made to enhance the functionality of GCNs services. Mansourian et al. (2012), developed a type of GCNs by integrating expert systems and semantic matching. Such efforts can be incorporated into a different generation of GCNs.

Generally, current GCNs have a number of basic components. In the following, these components will be reviewed (Fig.1).

2.1 Geoportal

Data and client are two general components in GCNs. From an operational point of view, a client (a person, organization or maybe a machine) delegates a series of parameters for a request to find required spatial data. Usually, clients' accessibility to spatial data facilitated through a geoportal. Geoportal can be broadly used in GCNs as a web-based portal for discovering and accessing spatial data and services (Dareshiri et al., 2019). It is the most visible part of Spatial Data Infrastructures (SDIs). Today's geoportals are focusing on interoperability through the implementation of standards for discovery and use of geographic data and services (De Longueville, 2010). One reason why Geoportal is important is that it provides a single and integrated environment for accessing all available spatial resources on the internet.

2.2 Catalog service

In GCNs, a collection of metadata records contains metadata for dataset, as well as services metadata, are managed together by using a catalog service. Catalog services support the ability to publish and search collections of descriptive information (metadata) for data, services, and related resources (Nebert et al., 2016).

2.3 Metadata repository

A metadata repository is a database or data dictionary (Moss and Atre, 2003) that stores descriptive metadata information including ownership, characteristics, rules, and policies. The purpose of the metadata repository is to provide a consistent and reliable means of access to data. From a technical point of view, a metadata repository may be stored in a physical or a virtual database stack.

2.4 Services

The "services" layer refers to a collection of geospatial web services. To manage geospatial data based on the Service Oriented Architecture (SOA) paradigm, a collection of standard and open geospatial web services

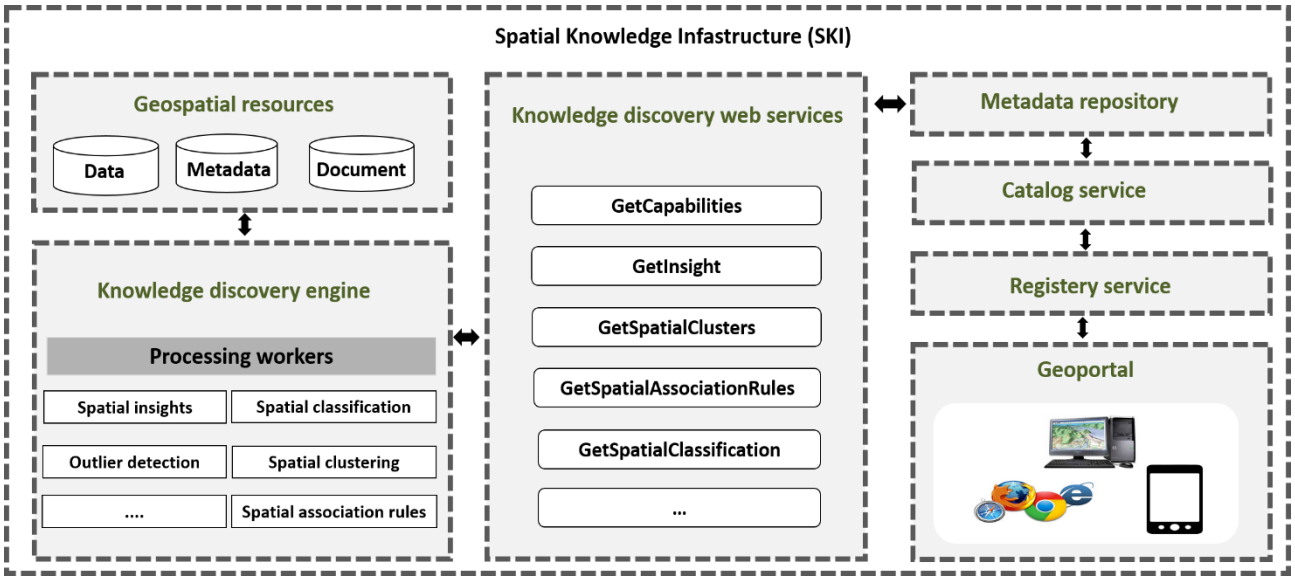

Figure 2. General components of a knowledge-based GCN.

are used in GCNs. These standards include Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), and Web Processing Service (WPS) are utilized for various purposes such as visualization, downloading, or processing of geospatial datasets.

2.5 Registry service

Because of distributed nature of GCNs, the geospatial resources such as datasets, servers, and geospatial services should be maintained by participants. Participants can include governmental agencies, non-governmental organizations (NGOs), universities, and companies in a community. In this context, a registry service (sometimes named a directory service) allows an individual to be known to the community with a central authority. In a nutshell, registry services are where that catalog services are registered to be discoverable by a geoportal.

A brief overview of the mentioned components indicates that current GCNs are well suited for distributed data-driven by accessing and sharing spatial data, but it is not yet adapted for knowledge discovery and sharing in an interoperable network. Knowledge-based GCNs needed an interoperable and scalable computing platform. Running spatial data mining algorithms usually is a time-consuming procedure that requires interoperable, and distributed big data processing frameworks. This indicates that without such infrastructure it is not possible to facilitate knowledge extraction process in current GCNs. Fortunately, with the advent of new big data frameworks, the limitations of traditional processing systems have been improved. Cloud-based computing services, modern storage solutions, and big data software utilities are technologies that enhance the functionalities of current GCNs.

3 Towards knowledge-based GCNs

Currently, the main purpose of GCNs is to share geospatial data. In this regard, a kind of GCN can be imagined that facilities sharing hidden knowledge in distributed geospatial data. To enhance GCN functionalities, a solution is proposed (see Fig.2). The solution is enclosed by a Spatial Knowledge Infrastructure (SKI) and integrates the capabilities of Knowledge Discovery Web Service (KDWS), and knowledge discovery engine into the current generation of GCNs.

As Fig.2 shows, in a knowledge-based GCN, a user requests and sets a series of parameters via a geoportal. Subsequently, the geoportal searches metadata repositories (which are already recorded in the registry services) through catalog services. By delegating the request, a spatial data mining task is assigned to the knowledge extraction engine. At this stage, multiple workers begin to process, and after ending the process, extracted knowledge can be shared by the user.

In addition to the common components described in Sec.2, the most important components of the solution are described in the following.

3.1 Spatial knowledge Infrastructure (SKI)

The increasing availability of geospatial data offers great opportunities for discovering valuable knowledge (Le et al., 2020, Alkathiri et al., 2019). However, some restrictions and difficulties such as conventional data

storage, computing technologies, heterogeneity, and interoperability concerns of spatial data are led to a delay in the development of such an architecture ecosystem. A more comprehensive description of these challenges can be found in (Li et al., 2020).

Extracting knowledge from distributed data requires a set of facilities, and a unique infrastructure (Talia and Trunfio, 2013). Without such infrastructure, it is not feasible to enable knowledge management procedures including discovery and sharing.

By drawing on the concept of SKI, Omidipoor et al. (2018) has defined SKI as follows: "The SKI is a widespread interoperable framework through which spatial knowledge is created, organized, shared, managed and used in many domains. It creates a mechanism to make the necessary processes of geographic knowledge with the highest efficiency and usability. SKI's primary goal is to combine SDI, spatial web services (SWS), and spatial data mining (SDM) concepts to facilitate geographic knowledge discovery as a collection of services".

3.2 Knowledge discovery engine

The extraction of appropriate knowledge from big geospatial data calls for innovative platforms tackling the data storage, processing, and analysis dimensions (Soille et al., 2018). Within data-intensive environment, parallel processing or cluster computing strategies can be used to solve latency problems.

A Knowledge discovery engine is a core processing component of knowledge-based GCNs. The component supports high-performance geospatial data mining techniques across clusters of computers named processing workers. In addition to general data mining methods, spatial clustering, spatial classification, and spatial association rule mining are the most popular data mining algorithms which are supported by this component. From a technical point of view, this component can be implemented in the grid or cloud environment. In this regard, innovative open-source parallel computing frameworks such as Apache Hadoop, and Apache Spark can be used.

3.3 Knowledge discovery web services

In the proposed solution, seamless and interoperable interaction between geoportal and knowledge discovery engines is provided by a set of Knowledge Discovery Web Services (KDWSs). The KDWS is formed the central focus of a study by Omidipoor et al (2020). The utilization of KDWS provides usable and interoperable knowledge that can be used in various applications (Omidipoor et al., 2020).

The service inherits the GetCapabilities operation from the Open Geospatial Consortium Web Services (OWS) interface and adds four operations named GetInsight, GetSpatialClusters, GetSpatialClassification, and GetSpatialAssociationRules.

To understand how to set request parameters, an Extensible Markup Language (XML) document is used to describe valid KDWS requests. Operations descriptions, supported data mining algorithms, access levels, header information, and available dataset are the most important information available in the XML document. Organizations and participants can implement different data mining algorithms, but what is important is that descriptive information is documented on the GetCapabilities.

The request and response life cycle of introduced structure can be understood in a scenario. Three organizations, A, B and C, have registered and shared their KDWS metadata. By using a geoportal, user wants to check available knowledge related to COVID-19. Through metadata repository explored by catalog services a list of items (formed from a combination of data and spatial algorithms) is delivered to the user. For example, "COVID-19 cases and deaths by counties" data are available in organization A. Since knowledge discovery engine of the organization are support spatial clustering algorithms, list of available algorithms for the COVID-19 dataset delivered to the user. As soon as the required item selected by the user, processing workers start to do the assigned tasks. Finally, the result is presented to the user. It should be noted that the processes is done via the infrastructure of organization A. To represent a user-friendly result, a set of visualization and mapping techniques may be used in geoportals.

4 Conclusion and future work

Nowadays, the central emphasis of GCNs is to share geospatial data. The availability of geospatial data beside technological innovations in geospatial services provide an opportunity to enhance the functionalities of current GCNs. In this regard, a type of GCN can be developed that facilitate sharing of hidden knowledge in distributed geospatial data. After reviewing the current GCN components, the general structure and required components of knowledge-based GCNs are proposed. To enhance the functionality of GCNs, the solution integrates the capabilities of KDWS, and a knowledge discovery engine enclosed by a SKI. Finding, accessing, and sharing geospatial knowledge are the most important features of the proposed structure. To investigate a full picture of knowledge-based GCNs, additional studies will be needed. Further work to implementing a geoportal that

supporting the proposed structure are therefore recommended.

References

Alkathiri, M., Jhummarwala, A., & Potdar, M. B.:Multi-dimensional geospatial data mining in a distributed environment using MapReduce, Journal of Big Data, 6(1), 1-34, https://doi.org/10.1186/s40537-019-0245-9, 2019.

Dareshiri, S., Farnaghi, M., & Sahelgozin, M.: A recommender geoportal for geospatial resource discovery and recommendation. Journal of Spatial Science, 64(1), 49-71, https://doi.org/10.1080/14498596.2017.1397559, 2019.

De Longueville B.: Community-based geoportals: The next generation? Concepts and methods for the geospatial Web 2.0. Computers, Environment and Urban Systems, 34(4), 299–308, https://doi.org/10.1016/j.compenvurbsys.2010.04.004, 2010.

Kargupta, H., Next generation of data mining. CRC Press, 2009.

Li, Z.; Gui, Z.; Hofer, B.; Li, Y.; Scheider, S.; Shekhar, S.:Geospatial information processing technologies. In Manual of Digital Earth; Springer: Singapore; pp. 191–227, https://doi.org/10.1007/978-981-32-9915-3_6, 2020.

Longley, P. A. et al., Geographic information science and systems. 4th edn. Wiley. Available at: https://www.amazon.com/Geographic-Information-Science-Systems-Longley/dp/1118676955.2015.

Mansourian, A., Omidi, E., Toomanian, A., & Harrie, L.: Expert system to enhance the functionality of clearinghouse services. Computers, Environment and Urban Systems, 35(2), 159-172, 2011.

Miller, H. J. and Han, J., Geographic data mining and knowledge discovery. CRC Press, 2009.

Moss, L. T., & Atre, S., Business intelligence roadmap: the complete project lifecycle for decision-support applications. Addison-Wesley Professional, 2003.

Nebert, D., Developing Spatial Data Infrastructures: The SDI Cookbook v. 2.0. Global Spatial Data Infrastructure, 2, 39-56, 2004.

Nebert, D., Voges, U., & Bigagli, L., OGC® Catalogue Services 3.0-General Model, Version 3.0,2016.

Omidipoor, M., Toomanian, A., Neysani Samany, N., & Mansourian, A.: Knowledge discovery web service for spatial data infrastructures. ISPRS International Journal of Geo-Information, 10(1), 12, https://doi.org/10.3390/ijgi10010012, 2020.

Omidipoor, Morteza, Ara Toomanian, and Najmeh Neisany Samani. "Towards Spatial Knowledge Infrastructure (SKI): Technological Understanding." In Proceedings of the 21st AGILE International Conference on Geographic Information Science, Lund, Sweden, pp. 12-15. 2018.

Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V., & Vasilev, V.: A versatile data-intensive computing platform for information retrieval from big geospatial data. Future Generation Computer Systems, 81, 30-40, https://doi.org/10.1016/j.future.2017.11.007, 2018.

Talia, D. and Trunfio, P. ,Service-oriented distributed knowledge discovery. CRC Press, 2013.

Toomanian, A., Methods to Improve and Evaluate Spatial Data Infrastructures. Lund University, Sweden. Available at: http://www.lunduniversity.lu.se/lup/publication/40d993fc-98ca-4f12-becd-a0e8d25c9048, 2012.

U.S. Federal Geographic Data Committee (FGDC), available at: https://www.fgdc.gov/dataandservices/clearinghouse_qanda, last access: 10 April 2022, 2022.