Geospatial semantics, ontology and knowledge graphs for big Earth data

Big Data has attracted a lot of attention from governments, industries and academia, and it has been applied to a large number of fields around the world. Big Earth data refers to big data associated with the Earth sciences that is characterized as being massive, multi-source, heterogeneous, multi-temporal, multi-scalar, highly dimensional, highly complex, nonstationary, and unstructured (Guo, 2017b; Guo et al., 2017a). Most big Earth data is related to a geographical location and is usually referred to as geospatial data (Lee & Kang, 2015).

Geospatial data is not only an important component of and the main way of organizing and visualizing big Earth data, it is also the foundation for integrating multi-source, heterogeneous big Earth data. With the development of Earth observation, deep exploration, computer simulations and other technologies, the capacity to acquire geospatial data has grown very fast, and thus the volume of geospatial data is also increasing exponentially. For example, according to the Union of Concerned Scientists Satellite Database (UCS (Union of Concerned Scientists), 2019), there are 2062 operational satellites currently in orbit around the Earth. These satellites produce huge amounts of remote sensing images with various spatial, temporal and spectral resolutions.

Since the 1950s, in order to promote the full sharing and value-added utility of geospatial data, developed countries as well as some developing countries have launched a number of geospatial data-sharing initiatives and programs that have established many spatial data infrastructures (SDI), data centers, and data-sharing platforms around the world (Bai & Di, 2010; Goodchild, Fu, & Rich, 2007; Harvey & Tulloch, 2006; Hu, Janowicz, Prasad, & Gao, 2015) and great achievements have been made in geospatial data sharing. For instance, the Data Sharing Service System of CAS Earth (http://data.casearth.cn/) launched by the Chinese Academy of Sciences has integrated 5.02PB data resources since 2018, and the National Integrated Earth Observation Data Sharing Platform of China (http://www.chinageoss.org) has collected 2,140,000 scenes acquired by 9 land-observation satellites, 1,440,000 scenes from meteorological satellites, and 10,000 scenes from ocean-observation satellites.

However, for most geospatial data-sharing projects, top-down organizational mechanisms are usually adopted, which results in networks with large amounts of data remaining in the hands of individual scientists and not being effectively shared. In terms of technology, the existing geospatial data-sharing platforms mainly use simple keyword-matching to search metadata, which leads to incomplete and inaccurate search results due to the lack of semantic reasoning. Therefore, in terms of the mechanisms used, geospatial data-sharing is moving towards a combination of top-down and bottom-up mechanisms and, in technical terms, towards precise data searches and
proactive recommendation services. With the support of geospatial semantics and ontology, internet geospatial data can be more effectively mined and VGI (Voluntary Geographic Information) collected; the publication of data can also be promoted. Furthermore, geospatial linked data and knowledge graphs for the implementation of intelligent data searches can be established along with precise data-sharing services.

As well as being an efficient and economic means of developing and applying comprehensive and complex geospatial models, geospatial model-sharing has attracted the attention of researchers. Many model interface specifications and model-sharing platforms have been developed, such as the Open Modelling Interface, the Web Processing Service (WPS) Interface Standard, the Geospatial Model Service Interface, the Universal Data Exchange model for reusing, sharing, and integrating geo-analysis models, and model-sharing platforms based on open standards (Blind & Gregersen, 2005; Feng, Liu, Euliss, Young, & Mushet, 2011; Nativi, Mazzetti, & Geller, 2013; OGC (Open Geospatial Consortium Inc.), 2007; Yue et al., 2015).

One key issue in geospatial model sharing and applications is the preparation of the input data. In particular, as their complexity and simulation accuracy increase, geospatial models need more and more input data. For most geospatial model users, however, preparing large amounts of input data is a time- and labor-consuming task and also expensive (Zhu et al., 2017). In order to save the time and cost involved in data preparation and to enable model users to devote more energy to the analysis of the model calculation results, it is necessary to make full use of the existing openly shared geospatial data online. Therefore, how to automatically match this shared data for use in geospatial models and promote the integrated sharing of “data-models” is an important research direction for the future. The core of the automatic matching of geospatial model data is to realize a consistent description of the model input data and openly shared data, as well as the calculation of the degree of matching, the accurate identification of differences, as well as the intelligent combination of data-processing services and automatic data-processing with the support of geospatial semantics and ontology.

Another key issue in geospatial model sharing is efficient computation. Due to the limits of computing capacity, in traditional geospatial calculations, especially global calculations, either the spatio-temporal resolution or the number of simulated elements has to be reduced and the spatial extent narrowed (Zhu et al., 2016). With the development of high-performance computers and distributed computing, geospatial simulation calculations are well supported. However, within these calculations, semantics and ontology should be used to clearly specify the interface, parameters, the initial and boundary conditions of the model, as well as the structure of the input data so that the parallelization of the algorithm and module, the assignment of computing tasks, and the partitioning of input data can be realized.

The 4th scientific research paradigm, Data-Intensive Scientific Discovery, emphasizes using big data processing and simulation models to mine and analyze massive scientific data to discover scientific laws and problems hidden behind the data (Hey, Tansley, & Tolle, 2009; Zhu et al., 2016). Earth sciences is a typical data-intensive research field that requires not only geospatial big data but also a highly efficient computing infrastructure to support the operation and application of geospatial models and tools that are used to discover spatiotemporal distribution patterns and differentiation rules (Zhu, Lu, Liu, Qin, & Zhou, 2018). Therefore, for Earth sciences, there is an urgent need to develop an one-
stop scientific research platform (e-Geoscience) that integrates the sharing of geospatial data, models and computing resources (Zhu et al., 2016).

In e-Geoscience, geospatial semantics and ontology are the foundation of the integration and sharing as well as the mining and analysis of geospatial data. They are also the foundation of the intelligent combination and efficient computation of geospatial models, and of the automatic matching of data for use in geospatial models. Moreover, using knowledge graphs, it is possible to achieve the spatio-temporal, semantic linkage of multi-source, heterogeneous Earth Science research resources, such as geospatial data, models, computing resources, and even standards, specimens/samples and literature, that can provide more convenient, efficient and accurate services for georesearchers (Zhu et al., 2016).

In summary, geospatial semantics, ontology and geographic knowledge graphs are some of the most important directions for big Earth data research and development. In this special issue, six papers related to this topic introduce and discuss geospatial data ontology, the extraction of online geographic knowledge, intelligent classification of remote sensing images, global land cover map integration and fusion, the enhancement of VGI application semantics, and spatiotemporal data discovery. We hope this issue will promote research into geospatial semantics and ontology as well as geographic knowledge graphs, and help to promote the development of big Earth data.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by the Natural Science Foundation of China [Nos: 41771430 and 41631177].

References

Bai, Y. Q., & Di, L. P. (2010). Review of geospatial data systems’ support of global change studies. *British Journal of Environment & Climate Change*, 2(4), 421–436.
Blind, M., & Gregersen, J. B. (2005). Towards an Open Modelling Interface (OpenMI) the HarmoniIT project. *Advances in Geosciences*, 4, 69–74.
Feng, M., Liu, S., Euliss, N. H., Young, C., & Mushet, D. M. (2011). Prototyping an online wetland ecosystem services model using open model sharing standards. *Environmental Modelling & Software*, 26(4), 458–468.
Goodchild, M. F., Fu, P. D., & Rich, P. (2007). Sharing geographic information: An assessment of the geospatial one-stop. *Annals of the Association of American Geographers*, 97(2), 250–266.
Guo, H. D., Liu, Z., Jiang, H., Wang, C., Liu, J., & Liang, D. (2017a). Big Earth data: A new challenge and opportunity for Digital Earth’s development. *International Journal of Digital Earth*, 10(1), 1–12.
Guo, H. D. (2017b). Big data drives the development of Earth science. *Big Earth Data*, 1(1–2), 1–3.
Harvey, F., & Tulloch, D. (2006). Local-government data sharing: Evaluating the foundations of spatial data infrastructures. *International Journal of Geographical Information Science*, 20(7), 743–768.
Hey, T., Tansley, S., & Tolle, K. (2009). The fourth paradigm: Data intensive scientific discovery. Microsoft research. ISBN: 978-0-9825442-0-4. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf

Hu, Y., Janowicz, K., Prasad, S., & Gao, S. (2015). Metadata topic harmonization and semantic search for linked-data-driven geoportals: A case study using ArcGIS online. Transactions in GIS, 19(3), 398–416.

Lee, J. G., & Kang, M. (2015). Geospatial Big Data: Challenges and opportunities. Big Data Research, 2(2), 74–81.

Nativi, S., Mazzetti, P., & Geller, G. N. (2013). Environmental model access and interoperability: The GEO modelWeb initiative. Environmental Modelling & Software, 39, 214–228.

OGC (Open Geospatial Consortium Inc.). (2007). OGC 05-007r7. In: OpenGIS® web processing service. Retrieved from http://portal.opengeospatial.org/files/?artifact_id=24151.

UCS (Union of Concerned Scientists). (2019). UCS satellite database. Retrieved from https://www.ucsusa.org/nuclear-weapons/space-weapons/satellite-database.

Yue, S., Wen, Y., Chen, M., Lu, G., Hu, D., & Zhang, F. (2015). A data description model for reusing, sharing and integrating geo-analysis models. Environmental Earth Sciences, 74(10), 7081–7099.

Zhu, A.-X., Lu, G., Liu, J., Qin, C.-Z., & Zhou, C. (2018). Spatial prediction based on third law of geography. Annals of GIS, 24(4), 225–240.

Zhu, Y., Pan, P., Fang, S., Da Xu, L., Song, J., Zhang, J., & Feng, M. (2016). The development and application of e-geoscience in china. Information Systems Frontiers, 18(6), 1217–1231.

Zhu, Y., Zhu, A.-X., Feng, M., Song, J., Zhao, H., Yang, J., … Yao, L. (2017). A similarity-based automatic data recommendation approach for geographic models. International Journal of Geographical Information Science. doi:10.1080/13658816.2017.1300805

Yunqiang Zhu
State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing Normal University, Nanjing, China
zhuyq@igsnrr.ac.cn