Big Data Era: Challenges and Opportunities for Geology (Postprint)

Authors: Zhai Mingguo, Yang Shufeng, Chen Ninghua, Hanlin Chen

Date: 2023-03-19T00:00:00+00:00

Abstract

Big data is profoundly influencing human life and transforming the paradigms through which humanity understands and investigates the world. As a quintessential data-intensive discipline, geology is confronting unprecedented challenges and opportunities. To address this challenge, geologists must not only refine conventional research methodologies, but more fundamentally, transform traditional modes of thinking to embrace the advent of the big data era. The integration of geology and big data not only substantially expands the cognitive horizons of geology and enhances the capacity for acquiring novel geological knowledge, but also injects innovative vitality into social production and public services underpinned by geology—including energy and mineral resource exploration, rational utilization of environmental resources, and disaster prevention and mitigation. Based on an analysis of the current state of geological big data research in China, this article articulates the frontier scientific questions in China's geological big data research, proposes strategic development objectives for geological big data, and explores the principal challenges and potential solutions confronting its advancement. Big data will revolutionize the thinking patterns of geologists, and the data-driven paradigm of scientific discovery will usher in an entirely new landscape for geological development. This article calls upon the geological community to devote greater attention and support to big data.

Full Text

Big Data Epoch: Challenges and Opportunities for Geology

Big data is profoundly influencing human life and transforming the way we understand and study the world. As a typical data-intensive discipline, geology is facing unprecedented challenges and opportunities. To meet these challenges, geologists must not only improve traditional research methods but, more importantly, change conventional mindsets to embrace the era of big data. The inte-

gration of geology and big data greatly expands the cognitive space of geology, enhances the ability to acquire new geological knowledge, and simultaneously provides innovative vitality for social production and public services supported by geology, including energy and mineral resource surveys, rational utilization of environmental resources, and disaster prevention and mitigation. Based on an analysis of the current status of geological big data research in China, this paper elaborates on the frontier scientific issues in geological big data research, proposes strategic development goals for geological big data, and explores the main problems and solutions in the development of geological big data. Big data will change the way geologists think, and the data-driven scientific discovery model will bring a completely new outlook to the development of geology. This paper calls on the Chinese geological community to pay more attention to and support big data initiatives.

Keywords: geology, big data, data-intensive, data mining

Geology is a natural science that studies how the Earth evolves, focusing primarily on the solid lithosphere to investigate the material composition, internal structure, external characteristics, and interactions and evolutionary history of Earth's various spheres. The emergence of geology stems from human society's demand for mineral resources such as petroleum, coal, and metallic and non-metallic minerals. As social productivity develops, human activities increasingly impact the Earth, and the constraints of geological environments on humanity become more pronounced. How to rationally and effectively utilize Earth's resources and maintain the environment for human survival has become a common concern worldwide. Consequently, geological research has expanded to include the rational use of resources, sustainable development of resources and environment, and harmony between humans and Earth.

Geology is essentially an information discipline and a typical data-intensive science that discovers inherent regularities in observational data through observation of natural phenomena. Since the beginning of the 21st century, Earth information detection technology has advanced rapidly, continuously improving data acquisition capabilities and resulting in exponential growth in accumulated Earth observation data. It is estimated that by 2020, global data volume will reach 40 ZB[1]. Geological big data is characterized by multi-source, multi-dimensional, heterogeneous, spatiotemporal, directional, correlated, random, fuzzy, spatiotemporally non-uniform, and nonlinear features. While geological big data shares similarities with general big data, significant differences also exist, bringing unprecedented opportunities and challenges to geology.

Geological big data exhibits the traditional "4V" characteristics of big data: large volume, great variety, high velocity, and low value density. It also possesses the "three highs" features of scientific big data: high dimensionality, high computational complexity, and high uncertainty. Due to the vast spatiotemporal ranges of geological object evolution and the numerous influencing

factors of geological processes, these high-dimensionality, high-complexity, and high-uncertainty features are particularly pronounced. The characteristics of geological big data are mainly manifested in the following four aspects.

Multi-Source Heterogeneity

Geological data acquisition platforms and methods are diverse, and data obtained through different means have different data organization and management formats. Examples include field outcrop descriptions, drill core descriptions, various geological report documents, numerous field mapping sketches and photos, remote sensing imagery, and real-time point data from geological disaster monitoring. Some data are stored and managed in paper form, while others undergo structured conversion and are aggregated into GIS databases. Different data organization methods create different data structures, and descriptions of the same geological entity also form semantic gaps due to differences in spatial reference frames and spatiotemporal scales. Different data acquisition methods and multi-angle descriptions result in severe heterogeneity and multimodality of geological big data.

Spatiotemporal Correlation

Geological bodies, structures, resources, environments, and disasters typically occupy vast spatial ranges and undergo long-term evolutionary processes. The "time dimension" of geological big data has complex characteristics of long-term duration and stage development, which other Earth science data do not possess. The high spatiotemporal characteristics of geological big data are reflected in two aspects: First, geological objects themselves have specific geological ages, and geological research has obvious regional characteristics. Research objects in specific periods and regions often have distinct differential features. Second, geological data describe the attributes of objects at specific locations at certain time points, and these inherent attributes are generated when the data are acquired. The temporal scale of geological data can span from seconds to hundreds of thousands of years, and variations in coordinate systems, projection parameters, detection accuracy, and basic granularity for describing spatial locations further increase the complexity of geological data. Therefore, geological data without temporal and spatial context are meaningless, and geological big data must be unified under consistent spatiotemporal benchmarks when conducting fusion analysis.

Complexity and Ambiguity

The Earth is a complex giant system, and the participation of geological data reduces the complexity of this system to some extent, making modeling and solving possible. However, due to interactions among various Earth sphere factors and the high complexity of geological processes themselves, many geological 规律 explanations and conclusions remain controversial. Additionally, the difficulty in quantitatively describing objects with geological data determines the

difficulty of geological data analysis, modeling, and computation. One technical orientation of big data is to "emphasize correlation rather than causality." We cannot understand the mechanisms of geological phenomena through data alone, and collecting global sample data is currently impossible. Therefore, the results of geological big data analysis are mostly fuzzy and uncertain[6].

Global Nature of Geological Bodies and National Interests

The distribution of geological bodies and units does not respect national boundaries, and the distribution of geological resources does not follow national or population demands. This creates difficulties in constructing global databases due to "national interest" interventions.

Research Progress in Geological Big Data

Overall, geological big data research in China is still in its infancy. On one hand, many question the applicability of big data to Earth sciences as an observational discipline. On the other hand, most researchers have not yet realized the importance of accumulating and sharing data, which hinders the development of geological big data to some extent. Additionally, the "correlation" pursuit in big data research presents a significant contradiction with the "causality" knowledge discovery in scientific research, challenging scientists to transform their thinking patterns[8]. Based on recent research achievements, the current status of domestic geological big data research can be summarized in three areas.

Storage Management of Geological Big Data

Geology has accumulated vast amounts of geological data, and with the rapid development of Earth information detection technology, new geological data are continuously generated. Geological big data includes not only qualitative and quantitative data but also text descriptions, geological maps, and even video and audio files left by geological workers. The long-term directory file storage method greatly reduces the efficiency of data query, retrieval, statistics, and update operations, resulting in low data service capabilities[3]. Therefore, constructing a geological information system that can effectively achieve integrated storage and management of structured, semi-structured, and unstructured data, static and dynamic data, and geological data and models is crucial for stable and efficient storage and retrieval of massive geological data[4].

Currently, some scholars have proposed using cloud platforms, Hadoop, and NoSQL technologies, drawing on real-time GIS spatiotemporal data models[9], to achieve dynamic management of geological spatiotemporal big data models. Hadoop is currently the standard platform for big data storage and processing, supporting large-scale data parallel processing through MapReduce. NoSQL databases use distributed node sets to dynamically handle loads. Distributed file system technology can be used to store geological big data and improve

data fault tolerance and reliability[10]. For example, the China Geological Survey Cloud Platform developed by the Key Laboratory of Geological Information Technology of the Ministry of Land and Resources and the Development Research Center of the China Geological Survey establishes a non-structured geological data storage organization model under this framework. By changing the storage, reading, search, and application patterns of non-structured data, it lays the foundation for providing accurate and rapid services for intelligent geological surveys[11].

Mining and Analysis of Geological Big Data

Three important technical orientations in the big data era are: use all data rather than sampling; prioritize efficiency over absolute precision; and emphasize correlation over causality[7]. This forces us to rethink data analysis from three dimensions: data types, data operations and maintenance, and the challenges brought by big data. Zhou Yongzhang et al.[12] believe that the core application technologies of big data and mathematical geoscience should include high-dimensional data dimensionality reduction, image data processing, infinite data stream mining, machine learning, association rule algorithms, and recommendation system algorithms.

Data mining refers to the process of searching for hidden information in large amounts of data through algorithms[13]. Compared with data retrieval and information extraction, data mining requires theoretical and technical support based on big data and knowledge-based intelligent reasoning[14]. Geological big data mining is the process of finding implicit features and patterns from data warehouses and applying them to geological 规律 research, metallogenic prediction, resource evaluation, environmental protection, and geological disaster prevention. This process involves using related methods and technical means such as artificial intelligence, machine learning, pattern recognition, inductive reasoning, statistics, databases, high-performance computing, and data visualization to automatically or semi-automatically acquire new understandable knowledge from multi-theme, multi-modal geological data, thereby providing decision-making support for geological thematic research and applications.

Currently, the task of digital geology is to vigorously promote the updating of data mining and data analysis methods in geological science. How to effectively mine and extract information from massive but low-value-density big data is a key problem to be solved in current geological big data research. The key technologies of geological big data analysis mainly involve comprehensive analysis of multi-source heterogeneous geological data, including correlation analysis of structured data, information extraction from semi-structured data, and verification analysis using non-structured data as validation for the above processing results. Furthermore, the rise of technologies such as the Internet of Things, virtual reality, and cloud computing makes it possible to develop Internet-based geological data resource sharing platforms and provides conditions for complex geoscientific computing. Integrating cloud computing and artificial intelligence

into geological big data mining and analysis has become a new development trend. For example, some scholars have borrowed big data thinking to use Bayesian networks to explore the genetic mechanisms of ore deposits, thereby constructing big data-intelligent metallogenic and prospecting models[15], promoting a revolution from "digital geology" to "intelligent geology."

Application Services of Geological Big Data

Geological big data has not only changed the thinking paradigm of geologists in studying scientific problems but also brought technological innovation to the geology industry based on data analysis. The improvement of data digitization levels in various geological fields has effectively broken down information silos, enabling quantitative analysis to advance further. The application services of geological big data are mainly reflected in five aspects.

(1) Basic Geological Survey. The "13th Five-Year Plan for Scientific and Technological Innovation Development of Land and Resources" points out the need to promote the development of digital geological survey systems toward intelligence, gradually achieving rapid geological data collection, real-time aggregation, efficient analysis and processing, and modeling. It also advocates promoting innovation in intelligent geological survey and service models supported by big data technology, deepening applications in geological mapping, mineral geological surveys, oil and gas geological surveys, and comprehensive coastal zone geological surveys. How to apply distributed data cloud storage, cloud management, and cloud service systems to various basic geological survey databases to achieve efficient and rapid storage of massive, fragmented, unstructured, and diverse data is a hot topic in basic geological survey research in the big data era[16]. Additionally, China is currently carrying out digital geological surveys, and the "Geological Cloud 1.0" developed by the China Geological Survey was officially released and launched in 2017. This system provides multi-professional data sharing services for various geological survey professionals, including basic geology, mineral geology, hydrogeology and engineering geology, and marine geology. It also provides multiple types of geological information product services for the public. The upgraded and improved intelligent geological survey system has been demonstrated and applied in basic geology and mineral geological surveys.

(2) Land Resource Management. Land and resources departments have accumulated massive amounts of land data through years of information technology practice, leading to the proposal of full-scale land resource data integration and big data construction technology. In 2016, the Ministry of Land and Resources proposed to continuously improve the "one map" data resource system for land and resources and build a unified land and resources data sharing and opening platform. Big data collection and analysis technology has become an important technical means for building decision support systems and think tank information work platforms, gradually forming a new "Internet Plus" think tank operation system under information technology conditions. This is crucial

for enhancing decision support capabilities in macro-control, management monitoring, situation analysis, policy evaluation, and public opinion analysis of land and resources.

(3) Geological Disaster Monitoring. Supported by the Internet of Things and big data technology, fully mining the potential information value from massive geological disaster data, combined with multi-orbit, multi-scale, and multi-temporal remote sensing environmental monitoring technology, can establish intelligent geological disaster, groundwater, mine geological environment, land subsidence, soil and water environment, and geological heritage investigation and monitoring data collection systems and early warning and forecasting systems. This strengthens the analysis and prediction of disaster occurrence trends, enhances real-time monitoring and early warning, and uses the power of data to prevent and control geological disasters.

(4) Mineral Resource Exploration. Mineral resources are an important material foundation for national economic development, and mineral resource prediction is guiding work in resource discovery and exploration. Previously, professionals relied on existing knowledge and experience under certain theories and methods, using qualitative or quantitative methods for prospecting. However, with continuous progress in mineral resource prediction theory and the organic integration of geoscience information with virtual reality technology, 3S technology, database technology, 3D modeling, and visualization technology, significant progress has been made in understanding new metallogenic 规律. This method mines massive data related to geological science, conducts multi-dimensional and multi-feature descriptions and modeling of various deposit types, thereby replacing prediction models composed of few parameters. It achieves mineral resource prediction and evaluation that combines geological theory with practical problem-solving, mathematical application with mathematical model research, and information technology application[17]. Additionally, the emergence of big data-driven metallogenic prediction theory has further spawned numerous 3D visualization software systems and mineral resource prediction systems based on spatial databases, laying the foundation for intelligent prospecting[18].

(5) 3D Visualization. Data visualization is the best method and means to describe, express, and understand the relationships and models of various semi-structured or even unstructured problems[19]. Based on geological spatial big data, combined with 3D visualization and virtual reality technology, 3D dynamic visualization modeling of geological bodies and structures can constitute a "Glass Earth," helping researchers analyze, predict, evaluate, and make decisions. Taking the development of digital mine technology as an example, 3D visualization technology can more vividly display mine geological and geomorphological information and clearly reflect the occurrence state of ore bodies[20], thereby comprehensively and dynamically guiding researchers in ore body location and metallogenic prediction.

Challenges and Prospects

The big data era has brought both opportunities and challenges to the development of geology. On one hand, geological big data opens new vistas for comprehensively perceiving and understanding the Earth and provides new means and approaches for knowledge discovery and scientific and technological innovation in geological science. On the other hand, due to the "three highs" characteristics of scientific big data, geological big data presents difficulties for mining and utilization. Additionally, immature data exchange and sharing mechanisms have become obstacles to geological big data research and development. How to establish efficient big data service platforms and promote collaborative research among various disciplines with big data sources are important issues to be addressed in the future. China's geological big data research is still in its infancy, but its important strategic significance and development prospects should be affirmed. To accelerate the construction process of China's geological big data, three recommendations are proposed:

(1) Promote the establishment of a "Geology + Big Data" talent training system. Universities must respond to the challenges of the big data era by establishing geological big data talent training programs. We call on the Ministry of Education and the Ministry of Science and Technology to increase support for geological big data projects, cultivating talent through projects and nurturing interdisciplinary professionals who have solid geological foundations while being familiar with algorithm development, data modeling, and data architecture. These professionals should be competent in geological big data system development, mining and analysis, and application development.

(2) Accelerate the establishment of geological big data sharing and exchange platforms. The free flow and collaborative sharing of data are key to realizing the value of data resources. Currently, most geological data resource construction is driven by major scientific research projects with certain implementation cycles, and their data service platforms suffer from single functionality, low retrieval efficiency, and inconsistent database construction standards, resulting in poor data circulation and usability. A specialized national-level institution should coordinate the construction of a geological big data center participated by universities, research institutes, and geological production units. Under the premise of protecting national interests, the construction of standardized and unified geological big data sharing and exchange platforms should be accelerated to promote geological big data research and application.

(3) Mindset transformation for geologists and geological workers. Scientific big data has become an important approach to scientific research, and the data-intensive scientific paradigm has gradually been accepted. Geologists and geological workers should seize this historical opportunity, embrace big data, change traditional empirical thinking patterns, adopt new attitudes toward data, and use new thinking modes to obtain new knowledge and create new value from data.

References

1. Gantz J, Reinsel D. The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Framingham: IDC Analyze the Future, 2012.
2. Zhao Pengda. The Era of Big Data Calls for Data Science in All Scientific Fields. China Science and Technology Awards, 2014, (9): 29-30.
3. Chen Jianping, Li Jing, Xie Shuai, et al. Research Status of Geological Big Data in China. Journal of Geology, 2017, 41(3): 353-366.
4. Wu Chonglong, Liu Gang, Zhang Xialin, et al. Discussion on Several Issues Concerning Big Data and Its Utilization in Geoscience. Chinese Science Bulletin, 2016, 61(16): 1797-1807.
5. Wang Denghong, Liu Xinxing, Liu Lijun. Characteristics of Geological Big Data and Its Application in the Study of Metallogenic Regularity and Metallogenic Series. Mineral Deposits, 2015, 34(6): 1143-1154.
6. Tan Yongjie, Wen Min, Zhu Yueqin, et al. Study on the Big Data Characteristics of Geological Data. China Mining Magazine, 2017, 26(9): 67-71.
7. Zhou Yongzhang, Wang Liben, Chen Jun, et al. Big Data and Mathematical Geoscience. Acta Petrologica Sinica, 2018, 34(2): 255-263.
8. Zhai Mingguo. Preface to the 2017 Big Data Special Issue of "Geological Science". Chinese Journal of Geology, 2017, 52(3): 637-648.
9. Gong Jianya, Li Xiaolong, Wu Huayi. Real-time GIS Spatiotemporal Data Model. Acta Geodaetica et Cartographica Sinica, 2014, (3): 226-232.
10. Chen Jianping, Li Jing, Cui Ning, et al. Construction and Application of Geological Cloud Under the Background of Big Data. Geological Bulletin of China, 2015, 34(7): 1260-1265.
11. Wu Chonglong, Liu Gang. Status, Problems, Trends and Countermeasures of "Glass Earth" Construction. Geological Bulletin of China, 2015, 34(7): 1280-1287.
12. Zhou Yongzhang, Li Zhen, Chen Jun, et al. Big Data and Mathematical Geoscience: Background and Progress. Bulletin of Mineralogy, Petrology and Geochemistry, 2017, 36(2): 327-331.
13. Han J. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2005.
14. Li Deren, Zhang Liangpei, Xia Guisong. Automatic Analysis and Data Mining of Remote Sensing Big Data. Acta Geodaetica et Cartographica Sinica, 2014, 43(12): 1211-1216.
15. Chen Jianping, Li Jing, Xie Shuai, et al. Big Data-Based Metallogenic Prediction: Theories, Methods, Structures and Key Technologies. Geological Bulletin of China, 2015, 34(7): 1288-1299.
16. Wang Denghong, Liu Lijun, Li Jinyi, et al. Application of Big Data Technology in Quantitative Prediction of Mineral Resources. Geological Bulletin of China, 2015, 34(7): 1333-1343.
17. Wu Chonglong, He Zhenwen, Weng Zhengping, et al. Attributes, Classification and Key Technologies of 3D Visualization of Geological Data. Geological Bulletin of China, 2011, 30(5): 642-649.

18. Wu Chonglong, Liu Gang, Zhang Xialin, et al. Discussion on Several Issues Concerning Big Data and Its Utilization in Geoscience. Chinese Science Bulletin, 2014, 59(12): 1047-1054.
19. Wu Chonglong, Liu Gang. Digital Mine and Its Implementation. Mine Surveying, 2015, (2): 3-5.

Note: Figure translations are in progress. See original paper for figures.