# Postprint: Multi-source Heterogeneous Big Data Fusion Based on the Physical-事理-Human Framework

**Authors:** Li Aihua, Xu Weijia, Shi Yong

**Date:** 2023-08-23T00:00:00+00:00

## Abstract

In the era of multi-source heterogeneous big data, big data exhibits novel characteristics including interdisciplinary nature, diversity, and variability, while applications across broader domains generate new demands for data fusion, thereby enriching and expanding the connotation of data fusion in this context. Broadly defined, data fusion encompasses the integration of data resources, the integration of model methodologies, and the integration of decision-makers' knowledge and experience. This article analyzes the characteristics of multi-source heterogeneous data fusion across three distinct levels: the data layer, information layer, and decision layer; explores potential challenges that data fusion may encounter in aspects such as storage, utilization, analytical technologies, data management, and value determination; and proposes corresponding countermeasures and recommendations, providing references for various entities, including enterprises and governments, to efficiently manage data resources and conduct more profound data fusion analyses.

## Full Text

## Abstract

In the era of multi-source heterogeneous big data, data exhibits new characteristics such as cross-domain interconnection, diversity, and dynamic evolution. The expanding applications across broader fields generate new demands for data fusion, enriching and extending its connotation. Generalized data fusion encompasses the integration of data resources, the synthesis of model methods, and the incorporation of decision-makers' knowledge and experience. This paper analyzes the distinctive features of multi-source heterogeneous data fusion across three hierarchical levels—the data layer, information layer, and decision layer—and examines the challenges that data fusion faces in terms of storage,

utilization, analytical technologies, data management, and value determination. Corresponding countermeasures and recommendations are proposed to assist enterprises, governments, and other entities in efficiently managing data resources and conducting more sophisticated data fusion analyses.

**Keywords:** multi-source heterogeneous, Wuli-Shili-Renli (WSR), data fusion, big data

## 1 New Characteristics of Multi-Source Heterogeneous Big Data and Emerging Demands for Data Fusion

The Internet has interconnected people' s daily lives, enterprise production, and government administration, generating vast amounts of data through countless activities across society. These data sources are extensive and structurally complex, while enhanced data availability has led various domains to increasingly emphasize the extraction of value from data resources. Consequently, the new characteristics of massive datasets and the emerging demands from diverse fields have propelled multi-source heterogeneous big data fusion to the forefront of big data research.

The new characteristics of multi-source heterogeneous big data can be summarized as cross-domain interconnection, diversity, dynamic evolution, and consensus. Content from different activities and business operations frequently overlaps, yielding large volumes of cross-industry, cross-media, and cross-database data with strong interconnectivity. Data structures are highly diverse, encompassing not only structured data such as numbers and tables but also unstructured and semi-structured data including text, images, audio, and video. Moreover, diversity extends beyond data types and structures to encompass the multi-dimensionality and multi-granularity of content and knowledge embedded within the data, reflecting complex three-dimensional relationships between data and knowledge. Dynamic evolution refers to data changing over time, while consensus indicates that relationships among data and between data and knowledge have gained widespread acceptance, possessing universal applicability that facilitates establishing associations and uncovering additional knowledge.

Extracting information and knowledge from multi-source heterogeneous data and transforming them into value necessitates data fusion. Traditionally termed information fusion, data fusion involves combining and processing data and information from multiple sources to achieve complementary advantages, eliminate noise, resolve contradictions, enhance information integrity and credibility, and obtain more accurate and reliable estimates or decisions than single-source information. Information fusion models primarily include structural models and functional models. Structural models describe the operational mechanisms of information fusion systems, with deployment architectures classified as centralized, distributed, or hybrid. Functional models primarily characterize the functions of information fusion systems and subsystems, as well as relationships among components, including the JDL (Joint Directors of Laboratories) model,

the Omnibus model, and the OODA (Observation, Orientation, Decision, Action) model and its variants. The improved JDL model constructs a six-level functional framework for multi-source information fusion tasks: sub-object estimation, object assessment, situation evaluation, impact assessment, process optimization, and cognitive optimization. The OODA model comprises four components—observation, orientation, decision, and action—with extended versions capable of processing interdependent information. Information fusion patterns can be abstracted into three hierarchical levels: data-level fusion, feature-level fusion, and decision-level fusion. Data-level fusion directly integrates data collected by identical sensor types; feature-level fusion extracts features from raw data before integration; decision-level fusion performs higher-level synthesis on features or preliminary results to derive more comprehensive and systematic decisions. Numerous methods and technologies exist across these levels, including principal component analysis, Kalman filtering, Bayesian estimation, machine learning, D-S evidence theory, and intelligent computing.

Multi-source heterogeneous big data has introduced new demands that challenge data fusion in theoretical research, methodological development, and practical applications. Current applications of data mining and fusion have extended to socio-economic issues such as enterprise management, government governance, and banking risk prevention—domains that differ from traditional military-oriented sensor fusion applications. In social, economic, and management fields, data sources exhibit greater complexity and openness, research problems typically involve multiple stakeholders and strong systemic interconnections, and substantial directly or implicitly correlated data exists. Compared with traditional sensor data, the importance of pre-defined data sources may diminish while the discovery and identification of new data sources become critical. Additionally, since social activities invariably involve human participation, adequately incorporating "soft factors" and "soft data" related to "people" and integrating them with "hard data" has emerged as a new requirement. Regarding fusion patterns, combining multiple hierarchical levels rather than confining fusion to a single level, thereby enabling data fusion throughout the entire data mining process, represents a future development direction. Furthermore, application scenarios in social, economic, and management domains require strengthening the integration of cutting-edge technologies with domain expert knowledge, enhancing the interpretability of methodological tools and their connection to practice.

## 2 Research Framework for Multi-Source Heterogeneous Big Data Fusion Based on WSR

As previously discussed, given the cross-media, cross-industry characteristics and the interconnectivity, diversity, dynamic evolution, and consensus of current multi-source heterogeneous data, unified analysis and mining of structurally diverse data necessitate data fusion. Li et al. conducted a comparative analysis of the three hierarchical levels of information fusion and their relationships with

the "data, information, knowledge" hierarchy in business intelligence. Building upon this foundation and drawing from the Wuli-Shili-Renli (WSR) systems methodology, they proposed the concept of generalized data fusion to be implemented throughout the business intelligence analysis process.

The WSR systems methodology comprehensively considers three dimensions—"Wuli" (physical), "Shili" (logical/principle), and "Renli" (human)—in systematic practice, emphasizing the dynamic unity and close interconnection among the objective world, organizational systems, and human actors, all of which are indispensable. "Wuli" concerns the composition, attributes, and objective laws of the real world; "Shili" addresses problem-solving methodologies; and "Renli" focuses on human dynamic activities, thought processes, and interactions with the environment. In social, economic, and management domains, humans are crucial participants in various activities, and practical problem-solving and decision-making depend on objective conditions, solution approaches, and human-related factors. These correspond to the "physical," "logical," and "human" dimensions of data fusion. Based on WSR, this paper proposes that generalized data fusion involves comprehensively employing multiple methods to mine multi-source heterogeneous raw data, then synthetically and holistically processing and analyzing the extracted patterns, decisions, and other "soft factors" to ultimately achieve efficient fusion outcomes that support decision-making. Generalized multi-source heterogeneous big data fusion encompasses the integration of data resources, model methods, and decision-makers' knowledge and experience. In business intelligence, "data" represents raw, unprocessed resources obtained through various channels; "information" comprises potential features, associations, and patterns discovered through preliminary analysis; and "knowledge" consists of more valuable conclusions derived through further reasoning. Data provides raw materials for problem-solving, while information and knowledge furnish the basis for decision-making. Since WSR-based generalized data fusion permeates the entire business intelligence analysis process of "data—information—knowledge," it can be divided into three hierarchical levels: data layer fusion, information layer fusion, and knowledge layer fusion [Figure 1: see original paper].

At the data layer, WSR manifests primarily in data source identification and data collection. In scenarios such as social governance, enterprise management, economic development, and risk management, human behavioral data collected through mobile devices and networks play increasingly important roles. The cross-industry, cross-domain, and cross-disciplinary nature of research problems, combined with the cross-media characteristics of multi-source heterogeneous data, substantially increases the number of data sources while simultaneously complicating their selection and determination. Data selection involves subjectivity—different data choices for the same problem may yield discoveries from different perspectives. Data selection must comprehensively consider practical problems and domain expert experience. Therefore, based on WSR, data can be selected from three dimensions: objective data, data generated from behavioral activities, and data closely related to "people" such as evaluations, opinions,

emotions, judgments, and expectations. Additionally, multi-source heterogeneous data transformation and comprehensive indicator construction constitute essential components of data layer fusion. Structured, semi-structured, unstructured, multi-granularity, dynamic, and static data are difficult to model directly. Consequently, multi-source data must be transformed through aggregation, association, feature extraction, text mining, and new variable computation to enable unified analysis and provide a foundation for building comprehensive models and mining deep information at the information layer. The "Renli" dimension in WSR is reflected in the interpretability and practical relevance of data transformation and indicator establishment processes.

At the information layer, WSR application is manifested not only in the comprehensive integration of multiple model methods for analyzing data layer fusion results but also in the selection of model methods and the combination of data science technologies with knowledge, principles, and methods from social and economic domains. Social governance, economic development, and enterprise management possess their own characteristics and theoretical foundations; data fusion technological tools cannot be separated from these disciplinary cornerstones. Therefore, method selection requires continuous exploration of pathways for integrating traditional and emerging technologies, seeking a balance between accuracy and interpretability. Common methods for model establishment at the information layer include classification, clustering, association rule mining, and other machine learning, deep learning, and artificial intelligence approaches, as well as integrated models combining multiple methods. Moreover, method selection at the information layer is closely related to "Renli." When addressing practical social management problems, appropriate methods and models must be adopted based on comprehensive consideration of all stakeholders rather than blindly pursuing methodological complexity and precision.

At the knowledge layer, WSR application is primarily reflected in the need to organically integrate "Renli" into final decision-making, combining human-related factors with objective data through higher-level reasoning and mining methods to fuse lower-level fusion results, thereby obtaining deep relationships and comprehensible knowledge to fulfill requirements or support decisions. Decision-making and knowledge cognition are intimately related to humans; expert opinions, decision-maker preferences, and social environments may all influence final outcomes, making the "Renli" dimension critically important at the knowledge layer of data fusion.

## 3 Challenges in Multi-Source Heterogeneous Big Data Fusion During Data Application

The new characteristics of multi-source heterogeneous big data fusion in emerging application scenarios enrich and extend its connotation while simultaneously introducing novel challenges and difficulties in storage, integration, analysis, and management.

**(1) High-quality data storage remains an urgent issue.** Data storage constitutes the foundational and front-end work of data analysis; more effective storage facilitates more convenient and efficient subsequent data extraction, preprocessing, and analysis. However, data storage itself is cumbersome and complicated, and the explosive growth of multi-source heterogeneous data further increases its difficulty. Confronted with complex data sources, storage must address two key questions: first, which data to store—cleaning and removing historical data can save space but may also discard valuable resources, requiring a trade-off between data importance and space occupation; second, how to store data—organizing structurally diverse data in a clear and logical manner presents an unavoidable challenge. Data quality is also paramount, as it significantly impacts analysis results. Neglecting the storage stage, leading to non-standard preservation or errors, severely affects subsequent analysis efficiency and accuracy.

**(2) Data silos and usage barriers hinder integration and fusion.** Although massive amounts of data are generated continuously, effective utilization remains challenging, with most data existing in "data silos" that are mutually isolated, creating obvious barriers to utilization. The difficulty of data layer fusion and integration stems from two aspects: subjective factors, namely data usage permissions—many internal data are not externally accessible, making acquisition difficult; and objective factors, namely data generated from different business activities. Even within the same enterprise or institution, data suffer from fragmentation issues, disparate structural formats, and varying storage standards, complicating cross-departmental data usage. These factors increase the difficulty of data resource fusion and impede full value extraction.

**(3) Multi-source, heterogeneous big data increases analysis and mining difficulty across scenarios.** Multi-source heterogeneous big data presents numerous new challenges for data fusion technologies. Data fusion must not only achieve transformation and unified integration of multi-source heterogeneous data but also focus on implicit knowledge behind the data, strengthening understanding of data meaning and organically combining consensual knowledge with digital analysis. Cross-domain, cross-media, cross-language, and multi-disciplinary fusion remains at the forefront of research challenges. Fusion objects coexist in multiple forms—including numbers, tables, text, images, video, audio, knowledge, patterns, and models—spanning different domains and potentially different languages, requiring full consideration of diverse data resource characteristics and inter-domain differences and commonalities. Cross-language fusion depends on cross-language data association and large-scale knowledge bases. Furthermore, current massive, multi-source, heterogeneous data impose new requirements on processing and analysis speed, as much value is embedded in high-frequency data or data streams that demand efficient real-time processing technologies. Simultaneously, data fusion methods and technologies require continuous optimization to handle increased data volumes.

**(4) Data maintenance, security, and privacy leakage are current**

**management priorities.** Multi-source heterogeneous big data requires high-performance network architectures and robust data center support, making data warehouse and data center operation and maintenance significant challenges. Large data volumes and dynamic evolution substantially increase the difficulty of incremental updates and error recovery operations in databases and knowledge bases. Ensuring data stability and high concurrency support while reducing server underload has become a key focus of data center maintenance. Data fusion analysis necessitates heightened attention to data security. Hardware failures and cyberattacks can lead to data loss, requiring continuous strengthening of multi-replica and disaster recovery mechanisms. Information security has also gained widespread attention, with increasing emphasis on privacy protection. While data fusion enhances data interconnectivity, it also increases leakage risks and threats to personal privacy, enterprise, and national security information. Therefore, protecting sensitive information during analysis and safeguarding data security while enabling flexible data utilization constitute important future research topics.

**(5) Data openness and sharing, data exchange, and data asset pricing require further attention.** Realizing the potential value of data is closely related to its degree of openness—generally, more openly available data can be mined for greater value and applied across more scenarios and domains. However, data openness faces numerous complex issues. Commercial interests, industry monopolies, and information security concerns severely restrict data openness. Clear definition of data rights and responsibilities presents difficulties; for instance, individual users are often both data producers and beneficiaries. In practice, data ownership and rights continuously change, with no clear consensus yet reached on ownership definition and rights allocation. The lack of comprehensive policies and regulations for data sharing also constrains data openness. As data value receives increasing recognition, data exchange, trading, and related markets have emerged, posing new challenges for the big data era: how to define data transaction value, maintain secure and healthy trading practices, and protect the legitimate rights and interests of individuals, organizations, and nations.

## 4 Reflections on the Development of Multi-Source Heterogeneous Big Data Fusion

From a complex systems perspective, data represents the objective "physical" reality, methods for fusing different data constitute the "logical/principle" dimension, and management of multi-source heterogeneous big data relates closely to the "human" dimension. Therefore, addressing challenges in storage, utilization, analysis, and maintenance of multi-source heterogeneous data, this paper draws upon the WSR methodology' s integrative thinking across its three dimensions to propose three development reflections from the perspectives of data, methods, and management [Figure 2: see original paper].

**(1) From the data perspective, continuously optimize collection and**

**storage.** For massive and structurally complex data, storage and database construction are complex engineering tasks. First, business requirements must be clarified, leveraging the combined efforts of data engineers, domain experts, and business personnel—this depends on in-depth analysis of objective data characteristics (the "physical"dimension), full understanding of the "human"dimension, and coordination between data and human needs. Data storage should not be limited to current requirements; new and potential demands will continuously emerge with technological advancement and business evolution. Data storage resources can be determined based on the three WSR dimensions: "physical" (objective data), "logical" (behavioral activity data), and "human" (evaluation/opinion/emotion data). Furthermore, collecting and storing cross-media, multi-source heterogeneous big data requires more advanced "logical"approaches. Under new data fusion demands, database construction and maintenance must be further strengthened, considering multi-source heterogeneity during storage to achieve compatibility with structured, semi-structured, and unstructured data, and establishing data fusion traceability mechanisms to enhance flexibility and simplicity in database incremental updates and partial modifications. Efficient, high-quality data storage is the cornerstone of big data fusion analysis; storage must maximize convenience for data utilization, with clear formats and unified standards facilitating efficient data invocation, processing, analysis, updating, and maintenance, thereby substantially saving resources and costs.

**(2) From the methods perspective, enhance data fusion effectiveness multi-dimensionally.** Improving fusion effectiveness for massive multi-source heterogeneous data depends on the joint progress of hardware equipment and technologies. Complex structures and enormous data volumes impose higher requirements on hardware equipment; improving hardware performance and perfecting relevant infrastructure lays a solid foundation for future big data fusion development. Regarding fusion methods, continuous technological innovation is needed, with improvements, integration, and fusion of original algorithms and models tailored to the characteristics, distinctions, and demands of data layer, information layer, and decision layer fusion. Fully drawing upon multi-disciplinary thinking provides inspiration for processing multi-source data and fusing multi-dimensional knowledge from different perspectives. Additionally, strengthening interdisciplinary talent cultivation enables data scientists, domain experts, and domain knowledge bases to jointly exert complementary advantages that achieve "1+1>2" synergies in data fusion theoretical research and practical application.

**(3) From the management perspective, establish sharing mechanisms to safeguard data openness and security.** As an emerging factor of production, data can generate increasing value. Both enterprises and governments are attaching greater importance to data, continuously enhancing big data management and proposing digital development strategies that keep pace with the times. Consequently, how to fully, efficiently, and securely realize data value has become a critical issue. Data value realization and potential release depend on data openness and sharing, yet openness inevitably impacts data security.

Therefore, it is necessary to comprehensively consider the interests of all parties, establish and improve data sharing mechanisms, and continuously refine relevant laws and policies to provide strong regulatory guarantees for data sharing and security. This will curb data abuse while enabling data sharing, establishing a sustainable and healthy data sharing ecosystem. Data and information security can also be protected through physical isolation combined with access control, preventing illegal access through isolation; researching strategies and assessment models to reduce privacy leakage risks, enabling timely risk warnings and protection strategy updates; and strengthening big data network security construction. In multi-source heterogeneous big data fusion management, the "human" dimension is crucial—connecting data silos and breaking data barriers requires efficient communication and collaboration among departments. A big data sharing ecosystem requires the participation and co-governance of all societal stakeholders to achieve a healthy environment of data sharing, interest protection, and security safeguarding, providing the foundation for future data fusion development and data value growth.

## 5 Conclusion

In emerging application scenarios, multi-source heterogeneous big data fusion has developed new characteristics and connotations across the data, information, and knowledge layers. Drawing upon the WSR systems methodology to analyze and research each hierarchical level of data fusion through the integrated dimensions of physical, logical, and human factors facilitates better solutions to multi-source heterogeneous data fusion problems and provides more comprehensive support for decision-making. Data fusion presents new challenges to humanity' s ability to harness data, creating novel difficulties in data storage, utilization, and management, yet it also offers tremendous space and potential for gaining deeper, more systematic, and comprehensive insights and for achieving more comprehensive mining and utilization of data value.

## References

[1] Meng X F, Du Z J. Research on the big data fusion: Issues and challenges. Journal of Computer Research and Development, 2016, 53(2): 231-246. (in Chinese)

[2] Chen K W, Zhang Z P, Long J. Multisource information fusion: Key issues, research progress and new trends. Computer Science, 2013, 40(8): 6-13. (in Chinese)

[3] Peng D L, Weng C L, Xue A K. Theory and application of multi-sensor and multi-source information fusion. Beijing: Science Press, 2010. (in Chinese)

[4] White J F E. A model for data fusion// Proceedings of the First National Symposium on Sensor Fusion. 1988, 2: 149-164.

[5] Bedworth M, O' Brien J. The Omnibus model: A new model of data fusion?. IEEE Aerospace and Electronic Systems Magazine, 2000, 15(4): 30-36.

[6] Blasch E, Breton R, Valin P, et al. User information fusion decision making analysis with the C-OODA model// 14th International Conference on Information Fusion. Chicago: IEEE, 2011: 1-8.

[7] Shahbazian E, Blodgett D, Labbé P. The extended OODA model for data fusion systems// Proceedings of 2001 International Conference on Information Fusion. Montreal: International Conference on Information Fusion, 2001: 8.

[8] Liggins M E, Hall D L, Llinas J. Handbook of Multisensor Data Fusion: Theory and Practice (Second Edition). New York: CRC Press, 2008.

[9] Kuznetsova P, Ordonez V, Berg T L, et al. TreeTalk: Composition and compression of trees for image descriptions. Transactions of the Association for Computational Linguistics, 2014, 2: 351-362.

[10] Rövid A, Remeli V. Towards raw sensor fusion in 3D object detection// 2019 IEEE 17th World Symposium on Applied Machine Intelligence and Informatics (SAMI). Herlany: IEEE, 2019: 293-298.

[11] Lathuilière S, Massé B, Mesejo P, et al. Deep reinforcement learning for audio-visual gaze control// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2019: 1555-1562.

[12] Liang X C, Hu P Y, Zhang L G, et al. MCFNet: Multi-layer concatenation fusion network for medical images fusion. IEEE Sensors Journal, 2019, 19(16): 7107-7119.

[13] Moratuwage D, Wang S, Wang D, et al. Belief functions for sensor data fusion for multiple object association using belief functions. Information Fusion, 2020, 57: 44-58.

[14] Li A H, Xu W J, Shi Y. Framework of business intelligence and analysis based on data fusion. Computer Science, 2022, 49(12): 185-194. (in Chinese)

[15] Gu J F, Zhu Z C. Knowing Wuli, sensing Shili, caring for Renli: Methodology of the WSR approach. Systemic Practice and Action Research, 2000, 13(1): 11-20.

[16] Zhu Z C. The WSR approach in the international systems community// Proceeding of 11th Annual Conference of Systems Engineering Society of China. Yichang: Systems Engineering Society of China, 2000: 149-164. (in Chinese)

[17] Gu J F, Tang X J. From ancient system thoughts to modern oriental systems methodology. Systems Engineering - Theory & Practice, 2000, (1): 90-93. (in Chinese)

[18] Li A H, Xu W J, Shi Y. A new data fusion framework of business intelligence and analytics in economy, finance and management// 2020 IEEE/WIC/ACM

International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). New York: IEEE, 2020: 940-945.

[19] Peng Y, Kou G. Research on data mining theory framework based on domain knowledge// The 3rd (2018) Chinese Academy Of Management Annual Conference- Collected Papers of Information Management Branch. Changsha: Chinese Academy of Management, 2008: 43-51. (in Chinese)

---

**LI Aihua** Ph.D. in management science and engineering, Professor of School of Management Science and Engineering, Central University of Finance and Economics (CUFE). Her research focuses on big data management & application, optimization & management decision, fintech and risk management, etc. E-mail: aihuali@cufe.edu.cn

**SHI Yong** Professor and Doctoral Supervisor of University of Chinese Academy of Sciences. Director of the Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences (CAS), Director of the CAS Research Center on Fictitious Economy and Data Science, and Fellow of the World Academy of Sciences for the advancement of science in developing countries (TWAS). The main research directions are data mining and knowledge management. E-mail: yshi@ucas.cn

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