Scientific and Social Research, 2026, Volume 8, Issue 6
http://ojs.bbwpublisher.com/index.php/ssr
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Governance and Value Realization of Scientific Data in the Era of Data-Intensive Research

Yu Li*

Institute of Digital Economy, Beijing Academy of Science and Technology, Beijing 100089, China

Author to whom correspondence should be addressed.

Abstract: With the in-depth development of the data-intensive scientific research paradigm and the rise of the global data marketplace, scientific data has evolved from a by-product of academic research into a pivotal strategic resource driving scientific and technological innovation and enhancing productivity and innovation capacity. Based on the theory of public goods and transaction cost theory, this paper constructs a three-stage model for the value realization of scientific research data and reveals the phased evolutionary patterns and dynamic shifts of governance mechanisms in the process of scientific data transforming into factors of production. The research identifies that the lack of supporting infrastructure for data processing and value-adding serves as the core bottleneck. Corresponding optimization paths are further proposed, including establishing a hierarchical governance system, accelerating the construction of data processing platforms, and redefining the role of government.

Keywords: Fourth paradigm; Scientific data governance; Data marketplace; Value realization

Online publication: July 15, 2026

1. Introduction

At the beginning of the 21st century, Turing Award laureate Jim Gray put forward the concept of the Fourth Paradigm, which refers to data-intensive scientific discovery. It has profoundly reshaped the underlying logic of scientific research. Scientific research is accelerating its transformation from the traditional hypothesis-driven mode to a new data-driven paradigm, and data has become the core driver of scientific discoveries and technological innovations [1]. Meanwhile, data has been officially recognized as the fifth factor of production following land, labor force, capital, and technology. Driven by both the Open Science Movement and the commercialization of data as a production factor, the governance of scientific data has become a critical nexus connecting scientific and technological innovation with the digital economy [2].

In this context, scientific data possess dual and complex attributes. On the one hand, as outputs funded by public funding, scientific data have the characteristics of non-rivalry and non-excludability, which render them public goods and require public sharing in compliance with the principles of open science [3]. On the other hand, as key factors for industrial upgrading, scientific data need a certain degree of exclusiveness. They need to be priced, traded, and circulated via market-oriented mechanisms to incentivize the supply of high-quality data and in-depth data utilization. Such inherent structural tension has become the core dilemma in global scientific data governance [4]. This paper establishes a three-stage model for the value realization of scientific research data. Through international comparative analysis, it explores the governance logic and divergent governance pathways in the value enhancement of scientific data.

2. Literature review and theoretical framework

2.1. Evolution of scientific research paradigms

The development of scientific research has undergone four evolutionary stages: empirical science, theoretical science, computational science, and data-intensive science [5]. The core feature of the Fourth Paradigm is that scientific research is no longer fully dependent on hypothesis formulation. Instead, researchers discover underlying patterns through mining and analyzing massive datasets, thereby driving scientific discoveries and technological breakthroughs.

Scientific data are typical public goods due to their non-rivalry and limited excludability. Nevertheless, once scientific data enters the market exchange and becomes tradable data products, exclusiveness must be introduced. It is necessary to clarify property rights, define the scope of use, and establish valuation mechanisms (**Table 1**).

Table 1. Three-stage evolution and governance logic of scientific data value realization

Dimension	Stage 1: Data Aggregation and Standardization	Stage 2: Data Processing and Value Adding	Stage 3: Data Circulation and Marketization
Data Form	Raw datasets	Standardized data products	Market-oriented data assets
Core Functions	Eliminate data silos, unify formats and standards, and implement the FAIR principles	Conduct data desensitization, fusion, and annotation, match computing power, and convert raw data into highly usable data products	Realize pricing, transaction, and scenario matching, and complete value monetization and data circulation
Property Right Attribute	Pure public goods	Club goods	Factors of production
Governance Structure	Academic hierarchical governance with academic autonomy	Hybrid multi-party collaboration mechanism	Market-oriented mechanism
Core Institutions	FAIR principles, open licenses, data management plans, reputation incentives	Hierarchical classification standards, computational resource platforms, data trusts, security compliance review	Split property right system, scenario matching mechanism, income distribution system
Major Challenges	Uneven data quality, insufficient sharing incentives for researchers, lack of metadata	High processing costs, ambiguous definition of derivative property rights, prominent compliance risks	Absence of pricing models, high transaction costs, insufficient market liquidity
Participants	Researchers, academic institutions	Academic institutions, enterprises, professional platforms	Market players, investors, scenario demand sides

2.2. Three-stage model for scientific research data value realization

Combining Williamson's framework of governance structures with the theory of public goods, this paper constructs the three-stage model for scientific research data value realization (**Table 1**).

The first stage refers to data aggregation and standardization. Its core tasks are to break down data silos, standardize metadata, and conduct quality control, to ensure data are findable, accessible, and interoperable [6]. The main challenges include incomplete metadata, uneven data quality, and the inadequate willingness of researchers to share data.

The second stage focuses on data processing and value addition. It aims to process standardized data through desensitization, fusion, and professional annotation, and align with appropriate computational resources, to convert raw data into reusable high-value data products. High processing costs and ambiguous ownership of derivative data rights are the major obstacles at this stage, which is also the primary bottleneck restricting the value enhancement of scientific data [7].

The third stage is data exchange and market commercialization. Relying on pricing systems and trading platforms, value-added data products complete value realization and scenario-specific applications. This stage is governed by market mechanisms and relies on price signals, contractual governance, and clear property rights arrangements [8]. The prominent difficulties are the lack of mature pricing models and poor market liquidity [9].

3. Comparative analysis of four typical international governance models

3.1. European Organization for Nuclear Research (CERN)

The Open Data Platform of CERN is a benchmark for global open science. It provides free access to massive experimental data generated by the Large Hadron Collider for users worldwide. CERN has established a sound system in the data aggregation and standardization stage. However, it has long been caught in the dilemma of "open data versus commercial usability". The raw data of high-energy physics has extremely high processing barriers and requires huge professional processing costs. Most of the data cannot be processed for value addition, nor can it enter the market for circulation [10]. CERN transfers data processing technologies to fields such as medicine and energy via the OpenLab industrial alliance program, but this is essentially technology transfer rather than data circulation. The deficiency of supporting infrastructure for the second stage blocks the value realization path of raw data. This fully verifies the first proposition: phased development cannot be bypassed.

3.2. European Molecular Biology Laboratory: European Bioinformatics Institute (EMBL-EBI)

EMBL-EBI adopts a hierarchical data access mechanism and divides data into three categories: fully open data, controlled access data, and fully restricted data [11]. This hierarchical model realizes value release at different levels. For controlled data, the institute has initially established cooperation mechanisms with the industrial sector through agreements and compliance reviews, reflecting the governance characteristics of club goods. EMBL-EBI has formulated institutional arrangements for the smooth transition from the first stage to the second stage. However, market-oriented trading is not its core objective, and data delivery is mainly based on academic agreements. The market-oriented pricing and trading system at the third stage remains immature. This model provides a reference for institutional preparation before data value adding, but fails to solve the bottleneck of transition from the second stage to the third stage, which proves the second proposition: the coordinated development of all stages is a prerequisite for complementary mechanisms.

3.3. Science Data Bank (ScienceDB) of the Chinese Academy of Sciences

As China's first multidisciplinary scientific data platform open to the whole world, ScienceDB assigns Digital Object Identifiers (DOI) and Chinese Scientific and Technological Resources Identifiers (CSTR) to various datasets. It converts scattered research materials into citable, traceable, and evaluable formal academic achievements, and explores solutions to the problem of the reluctance of researchers to share data [12]. At present, the platform is actively expanding into the second stage: it formulates professional metadata standards and quality auditing rules for vertical fields via data communities, and integrates technical frameworks such as privacy computing and federated learning. It explores computing power matching and model training under the premise of "data available but invisible." Nevertheless, the platform is faced with challenges in sustainable operation and market traction. Its services are mainly public welfare-oriented, and it is difficult to recover data processing costs through market pricing. How to develop market-oriented value-added services while maintaining public welfare attributes has become a key bottleneck for its development toward the third stage. The platform is characterized by strong data aggregation capability, initially built data processing capacity, and weak circulation capacity, which confirms the third proposition of differentiated development paths.

3.4. Shanghai Data Exchange

Shanghai Data Exchange is a landmark infrastructure for the market-based allocation of data as a production factor in China. Its data corpus marketplace introduces academic literature, research reports, and other structured knowledge products into trading services to meet the training demands of large-scale AI models. This practice tries to build a market circulation framework in advance through institutional arrangements and government endorsement, so as to drive the upstream upgrading through reverse incentives. However, this model also reveals deep-seated contradictions consistent with the first proposition. Currently, transactions on the platform are mostly driven by policies. Product pricing mainly relies on cost-plus pricing and negotiated pricing, without mature market bidding mechanisms, resulting in insufficient market liquidity. Although the trading framework has been completed, the supply of high-quality tradable data products is inadequate, leading to the phenomenon of "low transaction volumes despite well-developed infrastructure." The root cause lies in the weakness of the data processing and value-adding links in the upstream, rather than defects in the circulation link. To build a sustainable market closed loop, the priority is to activate the upstream processing links and foster a sufficient supply of high-quality data products [13].

3.5. Differences between Chinese and foreign development paths

Comparative analysis reveals that European and American models are rooted in the tradition of academic autonomy. Governments mainly provide financial support instead of direct intervention, following a bottom-up progressive evolution path. They first improve academic norms, and then gradually expand to commercial applications. The advantages lie in endogenous institutional development of the academic community and low trust costs; the disadvantages are fragmented market exploration and difficulty in achieving economies of scale.

In contrast, China adopts a top-down development model with active government planning. Relying on strategic planning, large-scale investment, and policy pilots, China has rapidly completed the construction of standardized infrastructure and tried to deploy market-oriented platforms through leapfrog development. Its strengths include strong resource integration capacity and fast progress; the weaknesses are insufficient endogenous market momentum and difficulties in transforming from policy-driven to market-driven operation. The above differences stem from different initial institutional conditions: Europe and America feature academic autonomy and decentralized governance, while China features a strong national governance capacity and centralized resource allocation mechanism [14].

4. Discussion and policy implications

4.1. Common bottlenecks and breakthrough directions

The three-stage model for scientific research data value realization demonstrates that open science and the data marketplace are not a zero-sum game, but mutually complementary collaborative systems. This conclusion challenges the prevailing view that simply regards the two as opposites. The three core propositions have been verified through cross-case comparative analysis: development cannot be bypassed; different governance mechanisms apply to corresponding stages; and development paths vary under different institutional backgrounds.

The four typical cases all point to the universal core bottleneck worldwide: the deficiency of the second stage (data processing and value adding). The transformation from raw data to tradable products involves multiple costs, including data rights clarification, pricing, cleaning, and compliance review. Such work has significant externalities and indivisibility, which cannot be undertaken by a single participant alone. This results in the dilemma that "open data does not equal usable data, and usable data does not equal tradable data." In practice, CERN is restricted by ultra-high data decoding costs; ScienceDB struggles to balance public welfare and commercial interests; Shanghai Data Exchange suffers from the shortage of high-quality data products. Despite different manifestations, their root causes are the same: the absence of supporting infrastructure and institutional arrangements for data processing. The key to breaking the deadlock is to cultivate hybrid governance organizations. Such organizations are neither purely public welfare platforms nor fully commercial enterprises. They can integrate multiple participants, share various costs, and align diverse incentives, serving as intermediate organizations connecting all links.

4.2. Optimization paths

First, establish a hierarchical and dynamically evolving governance system. Learning from EMBL-EBI's experience, build a comprehensive standard framework covering all three stages. Basic scientific data, as pure public goods, should be fully open and shared. Industrial standard datasets, categorized as club goods, are processed and operated by authorized institutions or data trusts through paid licensing. Value-added data products, as factors of production, are developed by market players and circulated on trading platforms. The core is to clearly define the attributes and property right boundaries of different types of data, and align governance strategies with data types.

Second, strengthen the foundation of the second stage and build national-level scientific data processing platforms. Rely on national scientific data centers, integrate high-performance computing, privacy-preserving computation, professional annotation, and intelligent cleaning technologies to build public data processing platforms that also serve as test bases. Encourage enterprises to conduct secondary development and algorithm competitions based on open raw data. Bring mature use cases to market, so as to fill the infrastructure gap between open sharing and market circulation [15].

Third, reposition government roles from direct transaction coordination to rule-making and ecological cultivation. In terms of standard formulation, elevate data processing capability to the same strategic priority as computing capability, and unify standards for data product format, quality grading, and security compliance. In terms of incentive mechanisms, provide initial liquidity support for data processing enterprises via government procurement and industrial funds. In terms of ecological development, foster professional service institutions engaged in data asset evaluation, compliance audit, and data trust to strengthen the institutional foundation for market maturity. In the early stage of market development, the government should fill institutional gaps and gradually withdraw after the market matures, realizing a smooth transition from policy-driven to market-driven operation.

5. Conclusion

Against the backdrop of the paradigm shift toward data-intensive scientific research, releasing the value of scientific data has become a key factor in global scientific and technological competition and digital economic development. This research demonstrates that the transformation of scientific data from public goods to factors of production is not a simple linear process. It is a complex evolution constrained by institutional arrangements and path dependence. Although European, American, and Chinese models follow different development paths, they all face the common problem of inadequate infrastructure and institutional arrangements in the data processing and value-adding stage.

China should leverage its advantages in institutional integration and resource mobilization to achieve breakthroughs in the data processing stage. By constructing hierarchical governance rules and national-level public data processing platforms, the industry can bridge the key gap between open sharing and market-oriented transactions. The implementation of the above measures will convert dormant data resources into circulating data assets, realize the full potential of the multiplier effect of scientific data on productivity and innovation capacity, and provide sustainable support for high-level scientific and technological self-sufficiency.

Disclosure statement

The author declares no conflict of interest.

References

[1] Liu GF, Wu YQ, Han MZ, et al., 2024, Research Progress on Data Factors in China: Connotation Interpretation, Operational Mechanism, Governance System, and Practical Application. Library and Information Work, 68(23): 139–152. https://doi.org/10.13266/j.issn.0252-3116.2024.0252-3116.2024.23.011

[2] Central People's Government of the People's Republic of China, 2023, Opinions on Building a Basic Data System to Better Play the Role of Data Factors. Gazette of the State Council of the People's Republic of China, 2023(1): 28–33.

[3] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al., 2016, The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data, 2016(3): 160018.

[4] Huang RH, Li BY, Li N, 2022, Research on Collaborative Governance of Scientific Data under the Open Science

Background. Journal of Library Science in China, 48(1): 45–60.
[5] Hey T, Tansley S, Tolle K, 2009, The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research, Redmond.
[6] Gu LP, Chen XL, Zhang XY, et al., 2020, Strategies for Value Enhancement of Open Scientific Research Data. Library Forum, 40(9): 115–124.
[7] Leonelli S, 2016, Data-centric Biology: A Philosophical Study. University of Chicago Press, Chicago.
[8] Jones CI, Tonetti C, 2020, Nonrivalry and the Economics of Data. American Economic Review, 110(9): 2819–2858. https://doi.org/10.1257/aer.20191330
[9] Delacroix S, Lawrence ND, 2019, Bottom-up Data Trusts: Disturbing the "One Size Fits All" Approach to Data Governance. International Data Privacy Law, 2019(4): 236–252. https://doi.org/10.1093/idpl/ipz014
[10] CERN, 2024, CERN Open Data. https://opendata.cern.ch
[11] EMBL-EBI, 2024, Terms of Use. https://www.ebi.ac.uk/about/terms-of-use
[12] Jiang LL, Wang PY, Li ZW, et al., 2024, Construction of Security System for Shared Data Platforms — A Case Study of ScienceDB. Journal of Agricultural Big Data, 6(3): 351–362. https://doi.org/10.19788/j.issn.2096-6369.000030
[13] People's Government of Shanghai, 2023, Action Plan for Promoting Industrial Innovation in the New Digital Economy Track (2023–2025).
[14] Guo HD, Zou ZM, Chen G, et al., 2024, Reflections on Scientific Data Supporting Major National Strategies. Chinese Science Bulletin, 69(9): 1116–1122.
[15] Sun LL, Wang WJ, Sheng J, 2023, System Dynamics Simulation on Influencing Factors of Scientific Data Value Addition. Journal of Agricultural Library and Information Science, 35(9): 28–42.
[16] Williamson OE, 1975, Markets and Hierarchies: Analysis and Antitrust Implications. Free Press, New York.
[17] OECD, 2015, Making Open Science a Reality. OECD Science, Technology and Industry Policy Papers, No. 25. Paris: OECD Publishing. https://doi.org/10.1787/5jrs2g63zs1-en
[18] OECD, 2021, Recommendation of the Council on Enhancing Access to and Sharing of Data. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463