# Data Security Governance in Open Scientific Data Sharing (Postprint)

**Authors:** Sheng Xiaoping, Guo Daosheng

**Date:** 2023-04-01T16:16:03+00:00

## Abstract

[Purpose/Significance] This study aims to identify data security issues in scientific data open sharing and propose corresponding governance countermeasures to better promote the practice of scientific data open sharing in China. [Method/Process] Employing normative analysis, this paper systematically reviews and defines data security issues in scientific data open sharing, and then explores data security governance measures from three dimensions: confidentiality, integrity, and availability. [Results/Conclusion] Scientific data open sharing faces numerous security issues regarding data confidentiality, integrity, and availability. To govern confidentiality issues, three measures are proposed: strengthening data security legislation, establishing scientific data classification and grading standards and systems, and fully utilizing privacy-enhancing technologies. To govern integrity issues, three measures are suggested: establishing a Data Protection Officer system, implementing Data Protection Impact Assessment, and employing data authentication technologies. To govern availability issues, three measures are recommended: formulating scientific data availability policies, improving scientific data quality, and constructing national scientific data centers based on data alliances.

## Full Text

## Preamble

**Research on Data Security Governance in Open Sharing of Scientific Data**

*Sheng Xiaoping, Guo Daosheng*
School of Library, Information and Archives, Shanghai University, Shanghai 200444

**Abstract:**
[Purpose/Significance] This paper reveals data security issues in the open shar-

ing of scientific data and proposes corresponding governance measures to better promote the practice of scientific data open sharing in China. [Method/Process] Using normative analysis, the paper identifies and defines data security problems in scientific data open sharing, then explores data security governance measures from three dimensions: confidentiality, integrity, and availability. [Result/Conclusion] Scientific data open sharing faces numerous security problems regarding confidentiality, integrity, and availability. Three measures can address confidentiality issues: strengthening data security legislation, establishing scientific data classification standards and systems, and fully utilizing privacy-enhancing technologies. Three measures can address integrity issues: establishing a data protection officer system, implementing data protection impact assessments, and applying data authentication technologies. Three measures can address availability issues: formulating scientific data availability policies, improving scientific data quality, and building a national scientific data center based on data alliances.

**Keywords:** scientific data; open sharing; data security; security governance
**Classification Number:** G203
**DOI:** 10.13266/j.issn.0252-3116.2020.22.003

## 1. Introduction

In the era of big data, data serves as a core asset supporting and driving industrial and innovative development across nations, receiving unprecedented attention and protection. Data open sharing and data security governance have become "two sides of the same coin," representing focal points of policy and legal attention worldwide. The EU's General Data Protection Regulation (GDPR), hailed as the "strictest data protection act in history," officially took effect on May 25, 2018, setting a global benchmark for data security protection. Building upon GDPR, the United Kingdom enacted the Data Protection Act 2018. China has implemented the National Security Law and the Cybersecurity Law, while the Personal Information Protection Law and Data Security Law have been included in the legislative agenda of the 13th National People's Congress. The Measures for Data Security Management completed public consultation on June 28, 2019. Data security has been rapidly elevated to a position of critical importance from which no individual, enterprise, or industry can remain aloof.

In recent years, scholars have extensively explored topics such as scientific data open sharing, open data protection, digital data protection, data security management, data security governance, data protection and governance, data security risk management, data privacy management, and big data privacy and security policies. However, few studies have deeply examined data security governance issues in scientific data open sharing. In fact, data security governance and data security management are distinct concepts. Data security management involves planning, developing, and executing security policies and procedures to

provide appropriate authentication, authorization, access, and auditing of data and information assets. Its fundamental objective is ensuring that the right people use and update data in the correct manner while restricting all inappropriate access and updates. Its ultimate goal is protecting data assets in compliance with privacy and confidentiality regulations and aligning with business requirements. Data security governance, by contrast, is a system for maintaining the confidentiality, integrity, and availability of organizational data assets, encompassing management commitment and leadership, organizational structure, user awareness and commitment, policies, procedures, processes, technologies, and compliance enforcement mechanisms. It is also a comprehensive governance process for data security that requires consensus on data security governance objectives across all levels—from decision-making to technical implementation, from management systems to tool support—to ensure reasonable and appropriate measures are taken to protect data assets most effectively. The primary goal of data security governance is ensuring the security of organizational data assets and achieving their preservation and appreciation.

The main business activities of data security management include understanding organizational data requirements and regulatory obligations; defining data security policies and standards; defining data security controls and measures; managing users, passwords, and access permissions; monitoring user identity authentication and access behavior; classifying data and information; and auditing data security. Data security governance's main business activities include understanding organizational data security strategic needs; developing and maintaining organizational data security strategy; establishing data security governance institutions and systems; appointing data security management officers; formulating and reviewing data security policies, standards, and procedures; coordinating data security governance activities; resolving data security issues; supervising data security management programs and services; evaluating data asset value; and monitoring compliance. While data security management and data security governance are intrinsically linked, they show clear differences in primary objectives and business activities. In summary, data security management lays the foundation for data security governance, while data security governance provides safeguards for data security management.

Since resolving data security issues is crucial for implementing scientific data open sharing in China, and given the current lack of research on data security governance in this context, this paper will define data security problems in scientific data open sharing and construct a corresponding governance model and countermeasures to better promote scientific data open sharing and open innovation.

---

## 2. Data Security Problems in Scientific Data Open Sharing

The European Commission requires that open research data pilot projects must preserve data supporting research results published in peer-reviewed publications and other defined data, preferably in research data repositories, and take measures to enable third parties to access, mine, utilize, reproduce, and freely disseminate such research data to any user. On March 17, 2018, the General Office of the State Council promulgated the Measures for Scientific Data Management to promote scientific data open sharing. In this process (including open access, open storage, open publication, and open utilization), data security becomes an unavoidable issue requiring deep understanding of both the data security concept and the main data security problems involved.

### 2.1 Data Security Concept and Connotation

Data security is a science studying how to protect data in computer and communication systems from unauthorized disclosure and modification, comprising four control activities—cryptographic control, access control, information flow control, and inference control—as well as backup and recovery processes. Data security can be divided into four dimensions: physical, personnel, procedural, and technical (see Table 1 ).

Classic data security requirements include data confidentiality, integrity, and availability, aiming to prevent data leakage or destruction during transmission, storage, and other stages. Data confidentiality means a secure system only allows individuals to see data they are permitted to see, including ensuring private and secure storage of sensitive data, verifying legitimate users, and implementing granular access control. Data integrity refers to data consistency, correctness, validity, and compatibility, meaning data is protected from deletion and damage when stored in databases or transmitted through networks. Data availability means authorized users of a secure system can access data without delay. Since data or information is the core asset of modern organizations, their confidentiality, integrity, and availability form the foundation for any organization's long-term survival in the 21st century. Organizations that fail to adopt comprehensive and systematic approaches to protecting these attributes will remain vulnerable to various threats, including hard drive damage, human error or operational mistakes, hacking, virus infections, information theft, natural disasters, power failures, and magnetic interference.

However, data security needs to clarify and correct several security myths: (1) Hackers cause most security vulnerabilities. In fact, 80% of data loss is caused by insiders. (2) Encryption makes your data secure. In reality, encryption is only one method of protecting data; security also requires access control, data integrity, system availability, and auditing. (3) Firewalls make your data secure. In fact, 40% of internet intrusions occur despite firewall protection.

### 2.2 Data Security Problems in Scientific Data Open Sharing

Data security problems in scientific data open sharing similarly manifest in three aspects: confidentiality, integrity, and availability.

**2.2.1 Confidentiality-Related Security Problems** Scientific data open sharing must guarantee data confidentiality. Current confidentiality-related issues mainly include: (1) Privacy leakage due to ineffective privacy protection. For example, much research in public health involves medical records and histories, making it very difficult to protect patient privacy while openly sharing research results. (2) Anonymous data is not entirely secure. Open sharing causes data controllers to lose control over who can access data. Even anonymous data can reveal private information about data subjects or may still contain sensitive information related to individuals, allowing re-identification through linkage with other publicly available information. Thus, anonymous data is not completely secure. (3) Conflicts between scientific data open sharing and personal data protection. Open sharing requires storing scientific data containing personal information in open data repositories for unrestricted user access, mining, copying, dissemination, and utilization, which clearly conflicts with personal data protection principles. (4) Lack of data classification and grading standards and norms. China currently lacks top-level design for data open sharing and has not established government or scientific data classification standards, making it impossible to effectively identify important, sensitive, and private data or provide guidance principles for different data types. (5) Imperfect intellectual property protection mechanisms, including: difficulty in defining intellectual property rights for digital scientific data; lack of a legal framework for scientific data open sharing; and contradictory legal provisions regarding data sharing and usage rights. (6) Failure to adopt effective data security protection technologies, such as data encryption or privacy-enhancing technologies.

**2.2.2 Integrity-Related Security Problems** Scientific data open sharing must ensure data integrity but faces key challenges: (1) Non-standard or inconsistent data formats, incomplete or overly complex data, and incompatible software. (2) Conflicting scientific data results, where research results produced using the same data under identical conditions contradict each other, or similar data stored in different systems produce different outcomes. (3) Data contamination, including data distortion, fabrication, and overload. (4) Data theft and tampering, where individuals or institutions may steal openly shared scientific data for commercial gain or other malicious purposes without attribution or citation, compromising data reliability. (5) Data misuse, where open shared data may be abused to leak sensitive personal information, commercial secrets, or national intelligence for commercial reward. (6) Data loss, such as missing or incomplete auxiliary data, severe loss of historical data, loss of original data due to discarded research notebooks, data corruption from hard drive crashes, and degradation of digital media over time.

**2.2.3 Availability-Related Security Problems** Scientific data open sharing must guarantee data availability, with security problems mainly including: (1) Inadequate documentation and processing of scientific data. Many datasets may not be properly recorded from the outset, rendering them unusable. A Finnish social science data archive survey found that 54% of respondents considered concerns about data availability (e.g., incomplete documentation) a major reason for non-reuse of data in their field. (2) Strong restrictions on using personal data for scientific research. Scientific research exemptions cannot legitimize processing of personal data, only longer storage periods or further processing. Researchers must obtain data owners' consent when processing personal data. (3) Lack of 完善的 scientific data open sharing platforms. China's scientific data sharing platforms generally suffer from poor website functionality, limited browsable, searchable, and accessible data resources. (4) Data isolation and lack of update guarantees, resulting in poor data availability. (5) Ambiguous data rights and lack of effective authorization. Rights to data such as right to know, collection, ownership, preservation, and usage are currently vague and lack effective legal definition, with unresolved issues of co-ownership among multiple authors. A Finnish survey found that 47% of respondents considered lack of ownership agreements an important reason for data non-reuse; two-thirds (66%) considered "lack of informed consent" a major barrier to open access research data; and 48% believed open access increased risks related to confidentiality, research ethics, and data protection.

---

## 3. Data Security Governance Model for Scientific Data Open Sharing

To address these data security problems in scientific data open sharing, strengthening data security governance is an imperative choice. Since these problems involve many major activities of scientific data open sharing (e.g., open access, open storage, open publication, open utilization such as open citation) and many key governance links (including data property rights protection, data privacy protection, data security monitoring, data quality monitoring, data infrastructure construction, data personnel management, and data security technology development and application), this paper draws on Michael E. Porter's value chain model to construct a scientific data open sharing data security governance model, as shown in Figure 1 [Figure 1: see original paper].

The model's main features and value are: (1) It implements value chain thinking. The model summarizes the main activities of the scientific data open sharing value chain as open access, open storage, open publication, and open utilization, and treats various data security governance countermeasures as supporting activities, integrating these main and supporting activities from the data security requirements dimension to fuse data security problems and governance countermeasures. (2) It emphasizes problem orientation. The model constructs data security governance countermeasures oriented toward different data security

requirements—confidentiality, integrity, and availability—so that various data security problems in scientific data open sharing can find appropriate solutions. (3) It emphasizes correlation, interaction, and integration. The model forms an interconnected data security governance system covering main scientific data open sharing activities, applicable to various scientific data open sharing environments.

---

## 4. Data Security Governance Measures in Scientific Data Open Sharing

Based on the above governance model, data security governance countermeasures can be analyzed from three dimensions—confidentiality, integrity, and availability—to promote scientific data open sharing.

### 4.1 Confidentiality-Oriented Governance Measures

Since confidentiality problems in scientific data open sharing involve both inadequate management mechanisms (e.g., lack of data protection laws, lack of data classification standards) and insufficient data protection technologies (e.g., privacy leakage, failure to adopt privacy-enhancing technologies), three governance measures are proposed from legislative, management, and technical perspectives.

**4.1.1 Strengthen Data Security Legislation to Consolidate the Legal Foundation** To effectively protect data security, the EU enacted GDPR in 2016, Germany passed the new Federal Data Protection Act in 2017, and the UK passed the new Data Protection Act in 2018. Although China has formulated and implemented the National Security Law, Cybersecurity Law, and Measures for Scientific Data Management, it has not yet promulgated a dedicated Data Security Law or Data Protection Law. The National Security Law, effective July 1, 2015, establishes a national security system integrating political, territorial, military, economic, cultural, social, technological, cyber and information, ecological, resource, space, deep sea, polar, and nuclear security, but does not specify how government, institutions, or individuals should safeguard scientific data security. The Cybersecurity Law, effective June 1, 2017, regulates network-level security requirements, explicitly requiring maintenance of network data integrity, confidentiality, and availability; network operators must protect networks from interference, damage, or unauthorized access, prevent network data leakage or theft, and must not disclose, alter, or destroy collected personal information without consent. However, it cannot systematically resolve data security problems in scientific data open sharing.

The Measures for Scientific Data Management, promulgated March 17, 2018, clearly stipulate that scientific data involving state secrets, national security, social public interests, commercial secrets, and personal privacy must not be opened for sharing; if opening is necessary, review of utilization purpose, user

qualifications, and confidentiality conditions must be conducted with strict control over access scope. Competent departments and legal entities must establish sound management systems for scientific data involving state secrets, strictly managing creation, review, registration, copying, transmission, and destruction. They must strengthen full lifecycle security management, develop protection measures, enhance authentication and authorization management for data downloads, prevent malicious use, establish security review systems for data to be published or provided, build cybersecurity protection systems, and establish emergency management and disaster recovery backup mechanisms with offsite backups for important data. However, the Measures do not specify how to implement effective data security management at different lifecycle stages, nor do they plan data security governance measures for scientific data open sharing behavior. Consequently, data security problems in scientific data open sharing still lack adequate legal protection.

This situation has attracted attention from legislative bodies, government departments, and experts. On May 28, 2019, the Cyberspace Administration of China released the Measures for Data Security Management (Draft for Comment), which mainly regulates network operators' data collection, storage, transmission, processing, and usage behaviors within China and data security supervision requirements, helping to strengthen and clarify network operators' responsibilities in ensuring data security. However, it cannot comprehensively solve data security problems in scientific data open sharing, primarily because it does not regulate other stakeholders' (e.g., data producers, organizers, users) responsibilities and supervision mechanisms for ensuring data security. Fortunately, from June 28-30, 2020, the 20th meeting of the Standing Committee of the 13th National People's Congress reviewed the Data Security Law (Draft), which was published on July 2, 2020 for public comment. The draft contains 7 chapters and 51 articles covering general provisions, data security and development, data security systems, data security protection obligations, government data security and openness, and legal liabilities. Notably, "establishing and improving the data security governance system to enhance data security protection capabilities" and "the state establishes and improves a collaborative data security governance system" are included in Articles 4 and 9. It is hoped that the Data Security Law will be enacted soon to consolidate the legal foundation for scientific data security governance in China.

**4.1.2 Establish Scientific Data Classification Standards and Systems for Reasonable Security Control** Data classification is significant in data security governance. It is the process of grouping and organizing data according to common characteristics, such as sensitivity levels, risks, and compliance rules for their protection. The 2015 Outline for Promoting Big Data Development issued by the State Council explicitly requires promoting the formulation and implementation of key common standards for data collection, government data opening, indicators, classification catalogs, exchange interfaces, access interfaces, data quality, data trading, technical products, and security confiden-

tiality. China's Measures for Scientific Data Management stipulate that scientific data centers must be responsible for classification, processing, and analysis of scientific data; legal entities must classify scientific data, specify confidentiality levels and periods, opening conditions, objects, and review procedures, and publish open catalogs to provide social sharing through online download, offline sharing, or customized services. In this context, scientific data classification management has become a key issue requiring urgent resolution.

Recently, the Ministry of Industry and Information Technology issued the Guidelines for Industrial Data Classification (Trial), which 率先 domestically classifies industrial data into four levels, encouraging enterprises to appropriately share first- and second-level data while restricting second-level data to authorized institutions and personnel, and 原则上 not sharing third-level data. Although industrial data is not identical to scientific data, as an important type of scientific data, these guidelines can provide reference for formulating scientific data classification standards.

However, scientific data encompasses numerous types, including but not limited to: any data generated during research; any recorded data significant to researchers; source or primary materials needed to verify research results; digital object sets obtained and generated during research; application content (e.g., analysis software, simulation software, model inputs/outputs); database content (video, audio, text, images); research project supervision data; design portfolios and physical models; research logs; experimental results and lab notes; field notes and diary content; bibliographies and reading materials; spreadsheets; metadata; methods and workflows; models, algorithms, scripts; notes, audio tapes, video tapes; music score drafts; human, animal, geological materials; images or data visualizations; photos, films; plant materials, cell, bacterial, or virus samples; clinical records and test results; protein or gene sequences; questionnaires, transcripts, codebooks; various interview records; survey responses; test results; spectra; standard operating procedures and protocols; trade secrets, commercial information, pre-publication materials, or similar information protected by law. Therefore, dedicated scientific data classification standards are needed.

In 2017, Washington, D.C. adopted a 5-level data classification model: Level 0 (open data), Level 1 (public data), Level 2 (data for local government use), Level 3 (confidential data), and Level 4 (restricted confidential data), widely praised by open data advocates. The University of California, Berkeley classifies research data into: Level 1 (least sensitive, i.e., public information), Level 2 (low sensitivity, i.e., non-public, non-sensitive personally identifiable information), Level 3 (moderately sensitive personally identifiable information), and Level 4 (highly sensitive personally identifiable information). Similarly, the University of New South Wales classifies data into public, private, sensitive, and highly sensitive levels. Drawing on these perspectives, scientific data can be classified into four levels as shown in Table 2 .

To effectively implement scientific data classification, "data tags" can be used

to build a scientific data classification system. This system essentially uses data tags to associate data levels, security attributes, and access conditions, establishing a data tag knowledge base to store and share data files and implement classification management according to different security and access requirements. Based on the above classification table, a 4-level data tag model can be constructed (see Table 3 ). This model uses four different colors to represent different tag categories corresponding to different data levels. As levels increase, transmission, storage, and access requirements and security attributes also increase. For example, at the lowest level, blue data tags require no access credentials. Green data tags require verification of the requester's email address, possibly sending a link via email that the requester must respond to, or using password credentials. Starting from yellow data tags, requesters must sign data use agreements and use passwords or authentication. Red data tags require two-factor authorization, such as verifying both email and mobile phone numbers. Using computers, data tag processing can be automated, helping achieve reasonable control of scientific data security.

**4.1.3 Fully Utilize Privacy-Enhancing Technologies to Strengthen Confidentiality** Scientific data open sharing makes data more transparent and reusable by researchers, greatly promoting a favorable research environment but also exposing hidden privacy threats. Potential privacy violations inevitably increase. Privacy-enhancing technologies can reduce data privacy risks and strengthen confidentiality protection. These technologies include differential privacy, federated analysis, homomorphic encryption, zero-knowledge proof, functional encryption, secure multi-party computation, searchable encryption, private information retrieval, smart contracts, etc. Among them, differential privacy, homomorphic encryption, zero-knowledge proof, and secure multi-party computation are particularly noteworthy.

While anonymization can protect sensitive data, it relies on background knowledge assumptions, often only ensuring privacy protection on single datasets and failing to meet privacy protection needs in massive data environments. Differential privacy overcomes anonymization's limitation of requiring external information knowledge, applicable to scientific data sharing. It adds random noise to real data, causing protected data to be distorted while maintaining specific data or attributes (e.g., statistical characteristics), ensuring data remains usable after interference to achieve privacy protection.

Homomorphic encryption allows computations on encrypted text, generating encrypted results identical to those obtained using unencrypted original data. This avoids data leakage from interception during scientific data transmission and enables real-time collaboration among researchers who can share data without exposing original data. Zero-knowledge proof, similar to homomorphic encryption, does not reveal any original data. It can verify information validity without exposing the data proving it, providing a way to determine whether scientific data use remains consistent with the initial purpose for requesting

sensitive data, preventing misuse.

Secure multi-party computation is a cryptographic protocol distributing computation among multiple parties, allowing mutually distrustful parties to collaboratively compute on private data. Its implementation involves homomorphic encryption, garbled circuits, oblivious transfer, etc. For scientific data open sharing, its greatest benefit is meeting and exceeding GDPR requirements for cross-border data transmission, as it enables data scientists and researchers to conduct compliant, secure, and private computations on distributed data without exposing or moving them. In summary, privacy-enhancing technologies have great potential and have developed rapidly in recent years, providing support for confidentiality protection in scientific data open sharing.

### 4.2 Integrity-Oriented Governance Measures

In scientific data open sharing, main threats to data integrity include: (1) Hardware failure: storage devices or other computer hardware failures may cause corruption. (2) Configuration problems: errors in computing systems (e.g., software or security applications) may corrupt data. (3) Human error: people make mistakes that may accidentally damage data. (4) Corruption in transmission: data may be corrupted when transmitted to storage devices or through networks. (5) Intentional sabotage: people or software may intrude into computers and alter data. These five threats stem from poor data management, inadequate data protection, and lack of relevant technologies. Therefore, integrity problems can be addressed through establishing a data protection officer system, implementing data protection impact assessments, and applying data authentication technologies.

**4.2.1 Establish a Data Protection Officer System** To strengthen data security management and protection, GDPR explicitly requires government departments or public institutions conducting data processing, institutions whose core business is large-scale data processing (including regular, normal, systematic monitoring and processing), and enterprises with 250 or more employees to establish a Data Protection Officer (DPO) position. The DPO is responsible for supervising an organization's data protection strategy and implementation to ensure compliance with data protection laws and regulations. DPO responsibilities include but are not limited to: (1) drafting, reviewing, and updating data protection policies; (2) providing focus for decisions affecting personal data use across multiple departments, including conducting data protection (or privacy) impact assessments; (3) coordinating with other appropriate personnel responsible for relevant affairs and functions; (4) managing risks that may arise from personal data processing operations; (5) continuously conducting control assessments to ensure compliance with key data protection procedures; (6) handling and managing queries and complaints from data subjects regarding personal data protection; (7) developing, reviewing, and revising policies, processes, and procedures for processing personal data; (8) promoting a data protection cul-

ture and accountability among employees; (9) ensuring compliance with data protection laws and implementing regulatory feedback; (10) reporting directly to the board and cooperating with data regulatory authorities.

To fulfill these duties, DPOs must possess relevant legal knowledge and professional skills: (1) familiarity with data protection laws and practices, especially legal protection requirements for sensitive data; (2) understanding of data controllers' or processors' business processes and content; (3) knowledge of data information systems and security technologies; (4) ability to advocate and cultivate an organizational culture of data protection.

Data controllers or processors may hire internal staff or external institutions/individuals as DPOs, but must sign data protection service contracts. The same DPO can hold positions in multiple institutions if competent and easily contactable by regulators, employing units, and data subjects. To enable DPOs to perform their duties effectively, employing institutions must provide necessary support: funding and basic working conditions, establishing DPO teams if needed; guaranteeing non-dismissal during tenure; requiring functional departments to support DPO work; ensuring adequate time for duty fulfillment; and authorizing access to organizational or personal databases.

Establishing the DPO system can strengthen internal data supervision, reduce data infringement risks, and enhance data security governance. Although GDPR requires EU members to establish DPOs, this has not yet been written into Chinese law. Chinese public institutions, research organizations, and large/medium enterprises should learn from international experience, align with international standards, and establish DPO systems to better implement scientific data security governance.

**4.2.2 Implement Data Protection Impact Assessment** A Data Protection Impact Assessment (DPIA) is a systematic method for analyzing, identifying, and minimizing data protection risks in projects or plans. It helps determine the most effective ways to comply with data protection law obligations and meet privacy protection expectations. DPIA targets data processing behaviors that pose high risks to natural persons' rights and freedoms, including data collection, recording, organization, structuring, storage, modification, recovery, querying, disclosure, dissemination, distribution, use, clearance, or destruction. Since these behaviors involve confidentiality, integrity, and availability issues, implementing DPIA helps achieve data security governance.

The DPIA process includes three phases:

**(1) Preparation Phase.** Main tasks include: (a) Considering whether DPIA is necessary. GDPR requires data controllers to implement DPIA when data processing poses high risks to natural persons' rights and freedoms. DPIA should also be implemented for scientific data open sharing, especially when data leakage, infringement, or security risks exist. (b) Planning DPIA, including defining scope and establishing the DPIA team. (c) Identifying data pro-

cessing requirements and details: processing objectives; covered disciplines and geographic areas; data types; whether special data or sensitive information is included; collection methods and sources; data usage; processing methods; data formats, standards, and applicable software/systems; whether data is anonymous or pseudonymous; storage and destruction methods; retention periods; sharing methods, scope, and objects; user informed consent; relevant industry standards and guidelines; and prominent security issues. (d) Identifying relevant actors: data controllers, developers, organizers, processors, users, and other stakeholders. (e) Identifying relevant legal requirements such as GDPR, China's Cybersecurity Law, and Measures for Scientific Data Management. (f) Documenting preparation phase results in standardized procedures.

**(2) Assessment Phase.** Main tasks include: (a) Establishing evaluation criteria based on data security protection objectives—confidentiality, integrity, and availability. For confidentiality: unauthorized access must be prevented. For integrity: processed data must be complete, current, unmodified, authentic, and correct. For availability: relevant data must be accessible, understandable, and processable in a timely manner. (b) Identifying data security risk sources and types from data itself, processing procedures, systems, and methods. (c) Determining intervention levels and protection standards: normal, high, and very high. (d) Assessing security risks to confidentiality, integrity, and availability. (e) Determining appropriate security measures: encryption, write permission restrictions, hash value comparison, regular integrity checks, minimum/maximum reference values. (f) Implementing determined measures (after confirming compliance with laws like GDPR). (g) Testing and documenting evaluation results. (h) Producing the DPIA report.

**(3) Review Phase.** After completing the DPIA report, data regulatory authorities should evaluate and audit it to ensure implemented safeguards. When processing risks change, DPIA should be reviewed to ensure safeguards adapt to changes and receive continuous supervision.

**4.2.3 Apply Data Authentication Technologies** Data authentication technologies can address integrity issues like distortion, fabrication, damage, tampering, and loss. Main authentication technologies include traditional cryptography-based and digital watermarking-based authentication.

Traditional cryptographic methods generate digital signatures using hash functions for authentication. Digital signatures (or electronic signatures) are defined by ISO 7498-2 as "data attached to a data unit or cryptographic transformation of a data unit that allows recipients to confirm the data unit's source and integrity and protect against forgery." In scientific data sharing, digital signatures authenticate identities to ensure original data senders remain unchanged. They are portable, non-replicable, and can be automatically timestamped, preventing senders from easily modifying messages afterward. However, when necessary modifications occur, original signatures must be discarded and recalculated, consuming considerable time.

Digital watermarking-based authentication offers greater inclusiveness and anti-interference capability. Digital watermarking embeds information (e.g., authorship, timestamp, product attributes) into carriers in forms such as text, graphics, or sequences. When suspected infringement occurs, watermark information can be extracted via algorithms to prove whether digital products have been tampered with or forged. Watermarks are classified as fragile or semi-fragile. Fragile watermarks are highly sensitive for precise authentication, applicable to sharing very sensitive multimedia files where authentication fails if even one bit changes. Semi-fragile watermarks are more applicable, allowing conventional processing operations as long as content remains authentic and complete, distinguishing normal signal processing from malicious tampering. Therefore, semi-fragile watermarks can be used for copyright protection and content verification to ensure data integrity.

### 4.3 Availability-Oriented Governance Measures

A European study found that measuring health equity across Europe largely depends on reliable and comparable data availability at the regional level, and eliminating "data gaps" is a condition for eliminating "health gaps" among and within EU countries. Scientific data availability is equally crucial for other industries like aviation and pharmaceuticals. In scientific data open sharing, data management policies, data quality, and open sharing platforms all affect availability. Therefore, multiple measures can enhance availability: formulating scientific data availability policies or publishing data availability statements, improving scientific data quality, and building unified open sharing platforms.

**4.3.1 Formulate Scientific Data Availability Policies or Publish Data Availability Statements** In the open data movement, many government agencies, research institutions, and publishers have formulated data availability policies or published statements to promote scientific data open sharing and utilization. On December 23, 2019, the U.S. Office of Management and Budget (OMB) released the Federal Data Strategy and 2020 Action Plan, with "leveraging data as strategic assets" as its core goal. The strategy requires implementing 40 data management practices across three categories based on 10 principles including accountability, transparency, and relevance: (1) building a culture that values data and promotes public use; (2) controlling, managing, and protecting data; (3) promoting effective and appropriate data use. This strategy provides national-level guidance for managing and using federal government data, ensures federal data availability, and promotes data sharing.

Publishing data availability statements has become many publishers' primary choice for ensuring availability. Some research funders like UK Research Councils require data availability statements in publications. *Nature* encourages statements such as: datasets generated/analyzed in the current study are available in designated repositories; available from corresponding authors upon reasonable request; all data included in the article and supplementary documents; or

available from corresponding authors upon reasonable request due to reasons preventing public disclosure. Other top journals also provide data availability statements to support open science and ensure scientific data is discoverable, verifiable, and reusable. In summary, formulating availability policies or publishing statements helps enhance scientific data availability and improve security governance.

**4.3.2 Improve Scientific Data Quality** Research confirms positive correlation between data quality and availability—improving data quality effectively enhances availability. Data quality attributes include accuracy, confidentiality, integrity, availability, consistency, timeliness, relevance, and validity. While availability is only one attribute, it has intrinsic connections with others. Enhancing scientific data availability can thus start by improving multiple quality attributes. Data quality is also closely linked to data production, collection, organization, storage, and publication. Sharing is an important link in the data lifecycle, encompassing collection, organization, publication, dissemination, and utilization. Scientific data open sharing mainly includes open publication, open access, open storage, and open utilization. Therefore, improving data quality requires attention to the open sharing process. Moreover, scientific data is generated from research, production, and management practices, directly related to producers, organizers, publishers, disseminators, managers, and users. Improving data quality requires full stakeholder participation. In summary, implementing Total Data Quality Management (TDQM) is urgently needed to enhance availability.

TDQM applies total quality management thinking to effectively manage data or data products to improve their quality and utility. Drawing on existing perspectives, a TDQM process can be constructed (see Table 4 ) to improve data quality. Critical aspects include: (1) defining scientific data quality requirements, especially accuracy, confidentiality, integrity, availability, consistency, timeliness, relevance, and validity; (2) identifying TDQM stakeholders (producers, suppliers, organizers, disseminators, managers, users) for full management; (3) executing the TDQM process iteratively to continuously improve quality; (4) establishing a data governance framework for data quality management.

**4.3.3 Build a National Scientific Data Center Based on Data Alliance** Scientific data open sharing platforms play an irreplaceable role in providing usable data. China has built eight national scientific data sharing platforms including the National Population and Health Scientific Data Sharing Service Platform (2017) and proposed 20 national scientific data centers including the National High Energy Physics Scientific Data Center (2019) as strategic choices for optimizing national science and technology resource sharing platforms and improving service systems. However, some platforms suffer from low data accessibility and citation rates. Current national scientific data centers provide registered users with access within specific disciplines but have not achieved true open sharing. More importantly, these centers have not established inter-

connections or integration to form a unified "National Scientific Data Center," resulting in limited data availability.

Unlike domestic centers, the Australian National Data Service (ANDS), led by Monash University and jointly established with the Australian National University and Commonwealth Scientific and Industrial Research Organisation (CSIRO), not only manages Australia's scientific data but also openly shares research data from over 100 Australian research institutions, government agencies, and universities through its "Research Data Australia" portal, covering natural sciences, social sciences, arts, and humanities. This data alliance-based operation model has set a successful example in improving scientific data availability and openness. The model essentially involves national scientific data centers (or platforms) uniting with data producers, providers, organizers, and managers to form scientific data alliances jointly participating in data sharing and utilization. Users can access not only the center's data but also use the platform's alliance member dataset index to access scientific data stored in institutional repositories elsewhere, greatly promoting open sharing. Therefore, building a national scientific data center based on data alliance can help solve China's low scientific data availability and improve security governance.

---

## Conclusion

Ensuring data security is an unavoidable key issue in implementing scientific data open sharing. Data security problems in scientific data open sharing concentrate in three aspects: confidentiality, integrity, and availability, requiring multi-dimensional governance measures from legal, policy, institutional, management, technical, and platform perspectives to construct a scientific data open sharing data security governance system. However, this paper's proposals remain theoretical and require further validation and refinement in practice to effectively govern data security problems in China's scientific data open sharing and improve data governance levels and national governance capabilities.

---

## References

[1] Shi Yingcun. Analysis of global data security governance trends and industrial trends [J]. Information Security and Communications Privacy, 2019, 41(4): 35-37.

[2] Zhang Hanqing. Multi-level protection needed for data security in the big data era [N]. Economic Information Daily, 2019-05-09(7).

[3] Sheng Xiaoping, Wu Tong. Review of research on scientific data open sharing at home and abroad [J]. Library and Information Service, 2019, 63(17): 6-14.

[4] Pereira S, Gibbs RA, McGuire AL. Open access data sharing in genomic research [J]. Genes, 2014, 5(3): 739-747.

[5] Wiebe A, Dietrich N. Open data protection: study on legal barriers to open data sharing—data protection and PSI [M]. Göttingen: Universitätsverlag Göttingen, 2017.

[6] Littled DD B, Farmers, El-Hilali O. Digital data integrity: the evolution from passive protection to active management [M]. West Sussex: John Wiley & Sons Ltd, 2007.

[7] Li Shanqing, Zheng Yanning, Xing Xiaozhao, et al. Research on security management issues in scientific data sharing [J]. China Science and Technology Resources Review, 2019, 51(3): 11-17.

[8] Du Yuejin. Several basic issues of data security governance [J]. Big Data, 2018, 4(6): 85-91.

[9] Wang Shixuan, Zhang Liang, Li Jiaojiao. Data security protection in the big data era—centered on data security governance [J]. Information Security and Communications Privacy, 2020, 42(2): 82-88.

[10] Hill DG. Data protection: governance, risk management, and compliance [M]. Boca Raton: CRC Press, 2010.

[11] Fu Xia, Fu Cai. Legal governance of data security risks in the new era [J]. Journal of Yangtze University (Social Sciences Edition), 2019, 42(2): 58-61.

[12] Livraga G, Torra V, Aldini A, et al. Data privacy management and security assurance [M]. Cham: Springer International Publishing AG, 2016.

[13] Tamanee S, Solanki VK, Dey N. Privacy and security policies in big data [M]. Hershey: IGI Global, 2017.

[14] Mosley M, Brackett M, Earley S, et al. The DAMA guide to the data management body of knowledge (DAMA-DMBOK) [M]. Bradley Beach: Technics Publications, 2009: 151.

[15] Solms SH, Solms R. Information security governance [M]. New York: Springer, 2009: 24.

[16] Chen Lei. Seeing the light through the clouds—data security governance system [J]. Security Monthly, 2019(10): 4-10.

[17] European Commission. Guidelines on open access to scientific publications and research data in Horizon 2020, Version 3.2 [EB/OL]. [2020-06-06]. https://ec.europa.eu/research/participants/data/ref/h2020/grants_{manual}/hi/oa_{pilot}/h2020-hi-oa-pilot-guide_{en}.pdf.

[18] Denning DE. Cryptography and data security [M]. Massachusetts: Addison-Wesley Publishing Company, 1982: V, 7.

[19] Moran R, Levinger J. Oracle security overview 10g release 1 (10.1) [R/OL]. [2020-06-06]. https://docs.oracle.com/cd/B12037_{01}/network.101/b10777.pdf.

[20] Feng Dengguo. Big data security and privacy protection [M]. Beijing: Tsinghua University Press, 2018: 5.

[21] Calder A, Watkins S. IT governance: an international guide to data security and ISO27001/ISO27002 [M]. 6th ed. London: Kogan Page Limited, 2015: 10.

[22] National Academies of Sciences, Engineering, and Medicine. Open science by design: realizing a vision for 21st century research [M]. Washington, DC: The National Academies Press, 2018: 50-51.

[23] Ye Runguo, Chen Xiuxiu. Problems and suggestions on security protection for government data open sharing [J]. Information Technology and Standardization, 2016, 58(6): 22-25, 34.

[24] Peng C, Song X, Jiang H, et al. Towards a paradigm for open and free sharing of scientific data on global change science in China [J/OL]. Ecosystem health and sustainability, 2016, 2(5): e01225. [2020-06-06]. https://esajournals.onlinelibrary.wiley.com/doi/epdf/10.1002/ehs2.1225.

[25] Liu Runda, Sun Jiulin, Liao Shunbao. Preliminary study on data authorization in scientific data sharing [J]. Journal of Intelligence, 2010, 29(12): 15-18.

[26] Stagars M. Open data in Southeast Asia [M]. Singapore: Palgrave Macmillan, 2016: 17-20.

[27] Janssen M, Charalabidis Y, Zuiderwijk A. Benefits, adoption barriers and myths of open data and open government [J]. Information systems management, 2012, 29(4): 258-268.

[28] Wen Liangming, Zhang Lili, Li Jianhui. Ethical issues in scientific data sharing in the big data era [J]. Information and Documentation Services, 2019, 40(2): 38-44.

[29] Zhang Yiming. Brief analysis of data governance process [J]. China Information World, 2012, 10(9): 15-17.

[30] Committee on Science, Engineering, and Public Policy (U.S.), Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. Ensuring the integrity, accessibility, and stewardship of research data in the digital age [M]. Washington, DC: The National Academies Press, 2009: 96.

[31] Kuula A, Borg S. Open access to and reuse of research data—state of the art in Finland [M]. Tampere: Finnish Social Science Data Archive, 2008: 11-12.

[32] Xin Yi. Comparative study on website construction of scientific data sharing platforms in nine provinces [J]. China Science and Technology Resources Review, 2019, 51(3): 18-23.

[33] Sedransk N, Young LJ, Kelner KL, et al. Make research data public? Not always so simple: a dialogue for statisticians and science editors [J]. Statistical

science, 2010, 25(1): 41-50.

[34] Porter ME. Competitive advantage: creating and sustaining superior performance [M]. New York: The Free Press, 1985: 36-43.

[35] National Security Law of the People's Republic of China (Full Text) [EB/OL]. [2020-06-06]. http://news.sina.com.cn/c/2015-07-01/220132055212.shtml.

[36] Cybersecurity Law of the People's Republic of China [EB/OL]. [2020-06-06]. http://www.xinhuanet.com//zgjx/2016-11/08/c_{135813275}.htm.

[37] Zhou Yu. NPC deputy Wei Ming: accelerate formulation of Data Security Law [EB/OL]. [2020-06-06]. https://www.sohu.com/a/397151630_{362042}.

[38] General Office of the State Council. Notice on issuing Measures for Scientific Data Management [EB/OL]. [2020-06-06]. http://www.gov.cn/zhengce/content/2018-04/02/content_{5279272}.htm.

[39] Cyberspace Administration of China. Notice on public consultation for Measures for Data Security Management (Draft for Comment) [EB/OL]. [2020-06-06]. http://www.gov.cn/xinwen/2019-05/28/content_{5395524}.htm.

[40] Data classification guide [EB/OL]. [2020-06-06]. https://www.spirion.com/data-classification/.

[41] State Council. Notice on issuing Outline for Promoting Big Data Development [EB/OL]. [2020-06-06]. http://www.gov.cn/zhengce/content/2015-09/05/content_{10137}.htm?url_{type}=39&object_{type}=webpage&pos=1.

[42] General Office of Ministry of Industry and Information Technology. Notice on issuing Guidelines for Industrial Data Classification (Trial) [EB/OL]. [2020-06-06]. http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757020/c7772152/content.html.

[43] UNSW. Research data governance & material handling policy [EB/OL]. [2020-06-06]. https://www.gs.unsw.edu.au/policy/documents/researchdatagovernancepolicy.pdf.

[44] AWS. Data classification: secure cloud adoption [EB/OL]. [2020-06-06]. https://d1.awsstatic.com/whitepapers/compliance/AWS_{{{Data}}_{{Classification}}}.pdf.

[45] Berkeley Information Security Office. How to classify research data [EB/OL]. [2020-06-06]. https://security.berkeley.edu/education-awareness/best-practices-how-tos/how-classify-research-data.

[46] UNSW. Data classification standard [EB/OL]. [2020-06-06]. https://www.gs.unsw.edu.au/policy/document

[47] Sweeney L, Crosas M, Bar-Sinai M. Sharing sensitive data with confidence: the data tags system [EB/OL]. [2020-06-06]. https://techscience.org/a/2015101601/download.pdf.

[48] Fu Yu, Yu Yihan, Wu Xiaoping. Differential privacy protection technology and application in big data environment [J]. Journal on Communications, 2019, 40(10): 157-168.

[49] The Royal Society. Israel-UK privacy and technology workshop note of discussions [EB/OL]. [2020-06-06]. https://royalsociety.org/-

/media/policy/projects/privacy-enhancing-technologies/israel-uk-privacy-and-technology-workshop-note.pdf?la=en-GB&hash=218915A3D5AA244D333A22D104882551.

[50] Alameda T. What are PET technologies? How to maximize data value while preserving privacy [EB/OL]. [2020-06-01]. https://www.bbva.com/en/what-are-pet-technologies-how-to-maximize-data-value-while-preserving-privacy/.

[51] INPHER. What is secure multi-party computation? [EB/OL]. [2020-06-02]. https://www.inpher.io/technology/what-is-secure-multi-party-computation.

[52] Dobran B. What is data integrity? Why your business needs to maintain it [EB/OL]. [2020-06-03]. https://phoenixnap.com/blog/what-data-integrity.

[53] Montezuma LA. Why should a data protection officer be global? [EB/OL]. [2020-06-06]. https://iapp.org/news/a/why-should-a-data-protection-officer-be-global/.

[54] Lambert P. The data protection officer: profession, rules, and role [M]. Boca Raton: CRC Press, 2017: 45-46.

[55] Liu Jiangshan. Data protection officer system in EU General Data Protection Regulation [J]. China Science and Technology Forum, 2019(12): 173-179.

[56] Freeprivacypolicy. GDPR data protection impact assessments [EB/OL]. [2020-06-06]. https://www.freeprivacypolicy.com/blog/gdpr-data-protection-impact-assessment/.

[57] Central London Healthcare. Data protection impact assessment [EB/OL]. [2020-06-06]. https://clch.nhs.uk/about-us/publications/data-protection-impact-assessment-dpia-summaries.

[58] Bieker F, Friedewald M, Hansen M, et al. A process for data protection impact assessment under the European General Data Protection Regulation [J]. Lecture notes in computer science, 2016, 9857: 21-37.

[59] Bieker F, Martin N, Friedewald M, et al. Data protection impact assessment: a hands-on tour of the GDPR's most practical tool [C]// Hansen M, Kost E, Nai-Fovino I, et al. Privacy and identity management: the smart revolution. Cham: Springer International Publishing AG, 2018: 207-220.

[60] Zhang Xingang, Yan Haowen, Zhang Liming. A perceptual hashing algorithm for DEM data authentication and tampering localization [J]. Journal of Geo-information Science, 2020, 22(3): 379-388.

[61] Li Shuanbao. Information security fundamentals [M]. Beijing: Tsinghua University Press, 2014.

[62] Shoeb ZH, Sobban MA. Authentication and authorization: security issues for institutional digital repositories [J]. Library philosophy and practice, 2010, 12(5): 1-6.

[63] Tan Hui. Digital watermarking technology and its application [J]. Information & Computer (Theoretical Edition), 2018, 12(13): 221-222, 225.

[64] Costa C, Freitas A, Stefanidi I, et al. Evaluation of data availability on population health indicators at the regional level across the European Union [EB/OL]. [2020-06-04]. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6188-6.

[65] Shehab E, Bouin-Poitevet M, Hole R, et al. Enhancing digital design data availability in the aerospace industry [J]. CIRP journal of manufacturing science and technology, 2010, 2(4): 240-246.

[66] Hopkins AM, Rowland A, Sorich MJ. Data sharing from pharmaceutical industry sponsored clinical studies: audit of data availability [EB/OL]. [2020-06-04]. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-018-1154-z.

[67] The President's Management Agenda Team. Federal data strategy 2020 action plan [EB/OL]. [2020-06-04]. https://strategy.data.gov/assets/docs/2020-federal-data-strategy-action-plan.pdf.

[68] Springer Nature. Data availability statements [EB/OL]. [2020-06-04]. https://www.springernature.com/gp/authors/research-data-policy/data-availability-statements/12330880.

[69] Editorial. On data availability, reproducibility and reuse [J]. Nature cell biology, 2017, 19(4): 259-259.

[70] Ding Xiaoou, Wang Hongzhi, Zhang Xiaoying, et al. Research on correlation relationships among multiple properties of data quality [J]. Journal of Software, 2016, 27(7): 1626-1644.

[71] Bi Datian, Cao Ran, Du Xiaomin. Current status and prospects of scientific data sharing research [J]. Library and Information Service, 2019, 63(24): 69-77.

[72] Wijnhoven F, Boelens R, Middel R, et al. Total data quality management: a study of bridging rigor and relevance [EB/OL]. [2020-06-05]. https://ris.utwente.nl/ws/portalfiles/portal/47275011/Wijnhoven07total.pdf.

[73] Boelens R. A product-attribute approach to total data quality management [EB/OL]. [2020-06-06]. http://essay.utwente.nl/57694/1/scriptie_{Boelens}.pdf.

[74] Si Li, Hua Xiaoqin. Analysis of service effectiveness of China's scientific data sharing platforms [J]. Library Work and Study, 2014, 36(4): 24-26.

[75] Australian National Data Service. Research data Australia [EB/OL]. [2020-06-08]. https://www.ands.org.au/online-services/research-data-australia.

[76] Liu Runda, Zhao Hui, Li Daling. Preliminary study on data alliance model for scientific data sharing platforms [J]. China Basic Science, 2010, 12(6): 27-32.

**Author Contributions:**
Sheng Xiaoping: Conceptualization, writing and revision;
Guo Daosheng: Initial draft preparation.

**Abstract:**
[Purpose/Significance] This paper reveals data security problems in scientific data open sharing and proposes corresponding governance countermeasures to better promote China's scientific data open sharing practice. [Method/Process] Using normative analysis, the paper identifies and defines data security problems in scientific data open sharing, then discusses governance measures from confidentiality, integrity, and availability dimensions. [Result/Conclusion] Scientific data open sharing faces many security problems in confidentiality, integrity, and availability. Confidentiality problems can be governed by strengthening legislation, establishing classification standards and systems, and utilizing privacy-enhancing technologies. Integrity problems can be governed by establishing a data protection officer system, implementing data protection impact assessments, and using data authentication technologies. Availability problems can be governed by formulating availability policies, improving data quality, and building a national scientific data center based on data alliance.

**Keywords:** scientific data; open sharing; data security; security governance

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