The Study of Privacy Protection of Scientific Data Sharing Based on Data Life Cycle

Dacheng Song \(^1\text{,}^a\text{,}^*\), Ming Chen \(^2\text{,}^b\), Sheng Fan \(^3\text{,}^c\)

\(^1\) Department of Library, Information and Archives, Shanghai University, Shanghai, 200444, China
\(^2\) Qianweichang College, Shanghai University, Shanghai, 200444, China
\(^3\) School of Mechanical Engineering, HeFei University of Technology, Anhui, 230002, China

\(^*,a\) Corresponding author e-mail: email: songdacheng1227@outlook.com,
\(^b\) email: 2268843503@qq.com, \(^c\) email: fansheng1002@163.com

Abstract. Scientific data sharing has become an important activity to promote modern research findings in which helps to reduce costs and save time for data collection, but it also brings certain privacy issues when using scientific data. First, based the network and literature survey method to described the basic conceptual of scientific data sharing, and discussed the issues of privacy violation in scientific data sharing during the data lifecycle. Second, tried to propose a privacy protection model and framework to prevent privacy violation. Finally, provided some suggestions for the current privacy protection in scientific data sharing from different perspectives. The study contributes to the scientific data sharing by shedding light on how to protect privacy of scientific data sharing through the privacy protection model and framework.

Keywords: Scientific data, data sharing, privacy protection.

1. Introduction

With the development of Internet, social media, mobile applications and it have created tremendous economic and social value, which facilitate scientific research to move towards the path of networking, digitization and openness and greatly prompting the global scientific community’s access to massive amounts of scientific data [1]. China promulgated The Measures on the Management of Scientific Data on March 17, 2018, which aims to further strengthen and standardize the management of scientific data, safeguard the safety of scientific data, and improve the level of open sharing. In addition, it clearly stipulates that Scientific data of state secrets, national security, public interests, commercial secrets and personal privacy cannot be shared with the public. In other words, privacy must be protected when scientific data is being carried out. In recent years, the conflict between data sharing and privacy protection [2], data privacy management [3], privacy and security protection [4], privacy protection policies [5], privacy protection regulations [6], and privacy protection technologies have become one of the focal points of attention. However, people rarely have explored the issue of privacy protection in scientific data sharing. In fact, revealing private information of scientific data sharing is one of the
current barriers in the path of open science [7]. Therefore, it needs to strengthen the research on privacy protection in scientific data sharing in order to better promote the sharing and use of scientific data.

2. The Concept of Privacy and Privacy Invasion

2.1. The concept of privacy
Privacy is defined as the right of people to live in a way that is reasonably free from public scrutiny, whether that scrutiny comes from neighbors, investigators, or even the news photographers under US Common Law [8]. Compared with US Common Law, the Chinese legal professionals deem privacy is a kind of private information that has nothing to do with the interest of the public or group, private activities that the person do not want others to know, private activities that the person do not want others to interfere, and private space that the parties do not want others to invade [9]. Thus, there are three forms of privacy: 1) private information, refers to personal data or documents, including all personal circumstances that he or she is living, which belongs to an intangible privacy; 2) private activities, belongs to a dynamic privacy that unrelated to the public interest, such as daily life, social interaction, etc; 3) private space or sphere, belongs to a tangible privacy that the scope of personal secrecy, personal residence, traveler's luggage, diary, address list, etc.

2.2. The concept of privacy invasion
Privacy invasion can happen at any time. When you use Google Maps, Google knows where you are; when you use Google Maps on your smartphone, it tracks your location; when you use your cell phone, the mobile company knows exactly where you are. You have nowhere to hide, thus losing your privacy. The privacy invasion means that an unwarranted intrusion into someone’s private information, private affairs, or private life without their consent [10]. There are some typical invasions of privacy: 1) invade someone’s home or private affairs ( invade secret); 2) reveal some troubling private facts in public (public disclosure); 3) reveal some facts that make someone misunderstanding (distorted image); 4) use another person's name or character for one’s own benefit (infringement of rights).

3. Types of Privacy Violations in Scientific Data Sharing
The National Institutes of Health (NIH), one of the first international organizations to engage in data sharing and practices, believes that all scientific data should be opened for all of us, and it defines the scope and manner of data sharing in The Data Sharing Policy, in which make the rules to the process of data collection methods, storage locations, usage, and open access [11]. The National Science Foundation (NSF) as well as sets out some requirements about the project that stipulate the sharing of scientific data sources, sharing formats, archiving and preservation, reuse, dissemination, and derivative use in The Data Sharing and Management for Research [12]. Consequently, the issues about privacy violations of scientific data sharing from the data lifecycle perspective can be summarized in four areas: data collection, data storage, data use, and data disclosure.

3.1. Privacy violations in data collection
Data collection is a systematic process of collecting and analyzing specific information to provide solutions to relevant problems or evaluating the results. It aims to analysis some particular topics and further hypothesis testing to explain a certain phenomenon [13]. The metadata was settled at the initial stage of data collection, some of the raw data such as basic personal information, medical data and other sensitive data, has been malicious attacked by internal and external, making the metadata set easily subject to unlawful deletion, falsify, dissemination or exploitation. Data leakages occur frequently during the data collection process and it often connected with data loss [14]. Data loss means that the data no longer exists or unusable that it has been corrupted, while data is still present and unaffected when data leakages occur. Data leakages can come from a variety of sources, including primarily print formats and digital media. In general, privacy violations resulting from the data collection process can be divided into two categories—the visible privacy and the invisible privacy [15].
The visible privacy refers to large-scale collection of publicly available data that results in excessive data mining that infringes the rights of the data subject. For example, the researchers of University of Texas found it relatively easy to re-identify users in the Netflix database and confirmed all of the movies that users had evaluated when scientific data was released within two weeks [16]. According to the study, using the data about evaluation of six movies, an individual could be identified within 84% accuracy; if add the approximate date and rate of those six movies, the 84% accuracy can be raised 99%.

The invisible privacy refers to the collection of personal information by using various data collection tools, but the data subject is not aware of it. In March 2018, a major data breach scandal broke out at Facebook, and the Cambridge Analytica stole more than 87 million Facebook user’s data without get data subject permission [17]. It was said that those data were used to support US presidential candidate Donald Trump during the 2016 election, which found that it was also misused to influence the outcome of the UK’s EU referendum in favor of the Leave campaign.

3.2. Privacy violations in data storage

Some public research institutions units or government databases accumulate a large amount of research data, but it’s usually lack of clear monitoring mechanisms during data storage, sometimes external environment also can leads to data leakage in the databases, resulting in privacy violations in data storage. Thus, individuals have little control over the data stored in research institutions, government or other databases. According to The Insider Threat Report in 2019, it classifies the risks of scientific data breach into three areas [18]: 1) accidental data breach; 2) inadvertent data breach; 3) and malicious data breach.

Accidental data breach mainly refers to the fact that the data in the database is obtained by others due to some internal or external reasons, which causes certain sensitive data to be published, causing privacy violations. In May 2019, a data breach occurred in a system operated by Indian state-owned utility company named Indene. Anyone can download the private information of all Aadhaar users, exposing their names and unique 12-digit identification numbers, and even the services they are used, such as their bank information, other private information and etc. [19]. The government database contains the identity and biometric information of more than 1.1 billion registered citizens of India, such as fingerprints and iris scans. Anyone in the database can use their data (or thumbprints) to open a bank account, buy a mobile phone SIM card, participate in public utilities, and even receive state aid or financial aid. Some companies, like Amazon, can also use the Aadhaar database to identify their customers.

Inadvertent data leakage mainly refers to the data, especially sensitive data, in the database being illegally stolen, modified, copied, etc. Because internal personnel operating errors, violating security policies, or failing to clearly inform the collector of the specific circumstances during the data storage, giving rise to information leakage and privacy violation [20]. A typical example is Facebook’s large-scale scientific experiment [21]. When users access a new APP through Facebook, the personal information in the account is inadvertently stolen, causing the information of 50 million people to be leaked. When Facebook learned of the data leakage after the news, it did not take any action, but waited a few months before issuing an order to delete all data. The most important thing is that the experiment was carried out without clearly informing the participants.

Malicious data leakage mainly refers to external targeted attacks, such as hacker attacks, computer viruses, information espionage, which lead to data information leakage and the identification of sensitive information of relevant stakeholders in the database. For example, the US Natural Health Chain Community Health System (CHS) stated that millions of copies of non-medical patient identification data related to doctors’ practice were stolen [22]. Although there is no medical information or credit card numbers were stolen, it included millions of individual patient's name, address, date of birth, telephone number and social security number have brought serious hidden dangers to personal privacy.
3.3. Privacy violations in data use
In modern society, personal privacy is threatened. With the development of modern information technology, people can use search engines, social networks and modern technologies or tools, such as data mining and machine learning, to analyze in-depth personal information recorded in various forms. In the era of big data, privacy issues regarding the collection, storage and use of personal information are a recurring topic in public discourse. In general, erasing data sets can be re-identified in three ways: 1) insufficient identification; 2) pseudonym reversal; 3) and combing of datasets. These technologies are not mutually exclusive, and all three methods can be used to re-identify erased data.

Insufficient identification can cause some privacy risks. Inadequate de-identification occurs when direct or indirect identification is inadvertently retained among the public in the dataset, structured and unstructured data have caused re-identify, because direct or indirect identifier may reveal a person’s identity. For example, America Online (AOL) publicly released 20 million search requests to AOL search engine users. Before to the release of these data, AOL anonymized the data by deleting identifying information, such as user names and IP addresses [23]. However, these identifiers were replaced with unique identification numbers, so researchers could still use these data to be able to track user queries to specific individuals in a relatively short period of time.

Pseudonymization is an effective mechanism for removing sensitive data only if it cannot be reversed. In general, there are several ways to reverse the process of pseudonyms [24]. First, some pseudonyms are designed to be reversible and retain a key to reverse the process, yet this hinders their security function. Second, the longer you use the same pseudonym for a particular person, the less safe it is and easier to re-identify the person. Finally, if the method used to assign pseudonyms is discovered, the data can be redefined too.

Combine of datasets can re-identify sensitive data. The most powerful tool for re-identifying erased data is to comb through two datasets, both of which contain the same individuals. Dr. Sweeney reconfirmed a set of medical data that may have been anonymized by connecting the two databases together, she combined the electoral rolls that she purchased with the database of the local hospital. She bypassed the cleaning procedure in this way and reconfirmed the anonymous data [25].

3.4. Privacy violations in data disclosure
Data disclosure refers to the voluntary sharing of any data deemed relevant to a specific situation. Depending on the specific circumstances, the disclosure ways may be different. The specific phenomena of privacy violations in data disclosure are mainly divided into four categories [26]: 1) public disclosure; 2) accidental or malicious disclosure; 3) mandatory disclosure; 4) Open government information.

Public disclosure refers to the data available to the public at any time by publishing relevant information on the Internet. It is obvious that the disclosure of Personally Identifiable Information (PII) related data can cause privacy risks [27]. The more common risks are publicly available network tracking and activity logs, which can easily bring safety or security to individuals and organizations.

Accidental or malicious disclosure refers to the act of providing data to a third party due to insufficient data protection. In 2006, America Online (AOL) released a set of anonymous search query data. Once there is enough public metadata, these data will be linked back to the users who performed the search. These users will then be reported in the New York Times.

Mandatory disclosure refers to risks that accompany the obligation to own data. For example, you must respond to a subpoena requesting disclosure of data onto litigation. The Recording Industry Association of America (RIAA) sued Internet service providers (ISPs) for failing to effectively prevent piracy, raising concerns about the wisdom of handing over user identities and addresses to institutions other than law enforcement agencies [28]. ISPs are increasingly worried that special subpoena powers may be abused, and providers may be found to disclose personal information to copyright owners in bad faith. Many entities (including research organizations) choose not to keep traffic statistics on business and research interests to avoid any such enforcement.

Government information disclosure refers to the information that is produced or obtained by administrative agencies in the course of performing their duties, recorded and preserved in a certain
form, and publicly releases them in a timely and accurate manner. A typical example is that major telecom operators disclosed call data records of the National Security Agency around 2007, which created another level of risk of civil rights and freedom, such as imprisonment and restrictions on speech [29].

4. Models and Frameworks for Privacy Protection in Scientific Data Sharing

4.1. Models for privacy protection in scientific data sharing

The Australian Computer Association (ACS) released a technical white paper——Privacy in Data Sharing, which discussed the challenges of privacy in the scientific data sharing [30]. It also introduced several concepts for Data Sharing Models that used by several agencies around the world, such as the Australian Bureau of Statistics, to help them make decisions about how to effectively use confidential or sensitive data. Meanwhile, this report emphasized that a basic challenge is to solve the problem that a set of datasets contain personal information in the data sharing, because the data or the combining datasets will identify personal information. In this article, we proposes "Five Security Model" based on the white paper, which provide a common model for scientific data sharing without considering its specific situation. This model attempts to quantify different security thresholds, and aims to provide a reference model that can be applied to all privacy issues in the scientific data sharing, as shown as figure 1.

![Five Security Model](image)

**Fig.1 Five Security Model**

Based on Privacy in Data Sharing, this article further proposes a data sharing model to protect privacy, which solves the broader data sharing issues and can be applied to all as a reference model for privacy issues in the scientific data sharing. 1) Safe data mainly refers to the degree of privacy risk in scientific data sharing. It can also refer to the quality of the data, the conditions for collecting the data, the accuracy of the data, the percentage of coverage (completeness), the number of features contained in the data (richness) or the sensitivity of the data. 2) Safe project refers to legal, ethical and ethical considerations surrounding the use of shared data. Data privacy protection is usually clearly stipulated in regulations or legislation, including the EU’s General Data Protection Regulation (GDPR), Australia’s Privacy Act and Canada’s Personal Information Protection and Electronic Documents Act (PIPEDA). 3) Safe people refer to reviewing the knowledge, skills, and motivations of data users so that they can better store and use data. The basic premise is that the data owner can trust that those who will access the data will use it appropriately. 4) Safe setting refers to the actual control of data access. For security settings, the possibility of deliberate and accidental data leakage also needs to be clearly considered. 5) Safe output refers to the residual risk of publications derived from sensitive data, because no matter how hard-
working and capable data researchers are, errors may occur in all types of organizations. Therefore, facility managers need to check all statistics being released and prevent them from being released publicly if necessary.

This Five Security Model following characteristics: 1) a series of values of the model embodied in the five dimensions. The five concepts are measurement standards, not kinds of state, such as, security data is a dimension for evaluating data security, which does not mean that data is not public. 2) This model not only considers the personal information factors in the data, the sensitivity of the data, but also takes into account the sensitivity of the output data used in the data analysis. 3) This model explores five different and quantifiable security levels of people, projects, settings, data, and output. These five different security levels can interact with each other in different situations. At the same time, it needs to be pointed out that the biggest challenge of the model is the interaction between different dimensions, such as data can affect people, settings, projects and outputs. As time goes on, the scope of five risk dimensions is not enough to address the risks of scientific data sharing. It also needs to be supplemented with other dimensions, such as security organization, safe use, and security, data lifecycle etc.

4.2. Building the balanced framework of scientific data sharing
The ideal situation of privacy protection in scientific data sharing is to find a balance between ensuring the effectiveness of data sharing and the privacy protection, but there is a contradictory relationship between privacy protection and data sharing [31]. Based on Canadian data sharing policies and practices, this article attempts to build a balanced framework of scientific data sharing and privacy risks. It takes into account the different data privacy risk levels, the degree of data sharing and the weight of influencing factors in the scientific data sharing.

4.2.1. Privacy risk levels of different data types. The data sharing creates different levels of privacy risk depending on the type of data [32]. In order to balance the relationship between privacy and the utility of data sharing, four types of data with different degrees of privacy risks can be distinguished: a) raw personal data, which is data that has not been processed or simplified, such as names, social security numbers, or personal email addresses. For this data, no attempt is made to reduce the risk of re-identification. b) Pseudonymized data, which means that a person’s identification information is replaced by a random unique identifier. c) Anonymized data, which is data that is no longer identifiable, removes all personally identifiable information from the dataset and converts identifiable data into anonymous. However, as large amounts of data accumulate, data miners can discover hidden personal information from the datasets, which can lead to re-identification of the data. d) Non-personal data refers to datasets that do not contain personal information at all, such as public transportation time, weather conditions, ocean tides, route maps, public sector budgets, or environmental pollution. These datasets have little to do with personal information.

4.2.2. Degree of scientific data sharing. Scientific data sharing can be interpreted as providing other researchers with reasonable data access or supporting unique research materials for published articles. According to the degree of scientific sharing, it can be divided into three types [32]: a) Unshared data, which means that target group does not provide research data for specific purposes because of confidentiality of data. b) Conditional sharing data, where data is shared with a specific group for a specific purpose rather than open access. On the one hand, it blocks data sharing completely; on the other hand, it accelerates data as open data, so conditional sharing data can be seen as two extremes on a continuum. c) Open data sharing, it is an act in which an individual or organization provides data to society in a network environment and authorizes others to use it for free. At this time, the shared data is released as completely open data, with no restrictions on access or reuse.

4.2.3. Influencing factors of scientific data sharing. According to the Canadian Data Privacy Protection Guidelines, it is believed that the datasets should be made public depending on the specific circumstances. Each dataset has its own risk-benefit, in which the damage caused by the leaked must be
weighed against the expected benefit. Two factors should be considered [33]: a) The weight of social interests, that is, the data sharing should be related to the goal it is pursuing. It mainly should answer the following questions: What is the main goal of releasing data? How important is the goal? Whether the data may be used primarily for scientific research purposes? b) The weight of privacy benefit, that is, considering the degree of privacy violation in the data sharing. If the datasets contain the names of people with HIV and the public knows they are infected, they may be discriminated against, as shown as figure 2.

![Diagram](image)

**Fig.2** The balanced framework of scientific data sharing based on privacy risks

5. Privacy Protection Measures of Scientific Data Sharing

To address the issues of privacy invasion in scientific data sharing, countermeasures are proposed from different perspectives in order to make a balance between scientific data sharing and privacy protection.

5.1. Refining scientific data sharing privacy policies and legal issues

The law is the basis for realizing the privacy protection of scientific data sharing, and it is also the prerequisite for preventing data leakage. In the 1980s, the OECD formulated the earliest formal international instrument data privacy protection framework The Code of Conduct for the Protection of Privacy, and the Council of Europe’s Convention for The Protection of Individuals with regard to the automatic processing of personal data, these two documents form the core of many countries' data sharing privacy protection laws [34]. The core of the foreign legal system mainly revolves around the use and protection of data, including the EU’s General Data Protection Regulation (GDPR), Australia’s Privacy Act, Canada’s Personal Information Protection and Electronic Documents Act (PIPEDA), and California Consumer Privacy Act (CCPA). These regulations determine the principles and supervision methods for data protection in the form of statutory regulations [35]. With the further development of society, some special industries have higher standards and requirements for data privacy protection. The United States has formulated more refined laws and regulations for different industries, such as Financial Privacy Act, Cable TV Privacy Act, Video Privacy Protection Act, Freedom of Information Act, etc.

China enacted the Network Security Law of the People's Republic of China in 2017, it discuss the protection of personal information and rules that prevent network data from being leaked or stolen, but it hardly involved with the specific operation of data privacy protection in data sharing. Therefore, China should refine the corresponding legal provisions on specific protection measures for data privacy, formulate more detailed data sharing privacy protection laws, in order to strengthen the protection of privacy in scientific data sharing to ensure that the privacy of data subjects is not violated through specific legislative measures.
5.2. Protect data privacy from different stages of scientific data sharing

It is an important way to prevent data disclosure that strengthen the protection of data privacy at different stages of data sharing [36]. The stages of data sharing can be divided into planning, construction, operation, preservation, etc. The privacy protection in data sharing does not formulate corresponding laws or regulations, but also needs to be implemented in the specific implementation process stage.

First, in the planning stage, the corresponding data privacy protection framework and model should be formulated according to the relevant industry data privacy protection laws and regulations combined with its current situation, which formulate the corresponding data privacy protection framework and propose the further solutions. Secondly, we should formulate data privacy protection framework and model in accordance with the data management process in the construction phase, and select corresponding technology to strengthen the privacy protection in data sharing in line with the enterprise. In addition, it should discover problems during the operation process in the operation phase, so as to achieve the realization of the data sharing security goal during the feedback process. Finally, the preservation phase refers to the definition of data access system standards and the control and implementation of access rights. The standards defined by the data access system should be unified, in other words, it should be security, simplicity, integrity, availability, etc. Meanwhile, the purpose, conditions, time, responsibility, and privacy level of data access should be clearly defined in the access control.

In some specific industries during the data sharing, such as biology, medicine, geography, the privacy protection of organizations are different. More specifically, the specific areas of the different industries affect the specific implementation content and methods of privacy protection.

5.3. Use information technology to solve the re-identification of risks in scientific data sharing

Information technology is an important way for realizing privacy protection in scientific data sharing, and it is also helpful to solve the risk of re-identification in scientific data sharing. Innovations and applications in emerging technologies such as big data, cloud computing, and the Internet of Things have created tremendous economic and social value, but collecting relevant data, especially in cloud storage capacity, also makes privacy issues more visible in scientific data sharing. From the technical point of view, it is mainly possible to choose three different angles for privacy protection: 1) statistical disclosure control (SDC); 2) privacy protection in data mining (PPDM); 3) privacy enhancement technology (PET).

As the growing need for researchers to access useful microdata, microdata can be published in the form of statistical disclosure in order to balance the validity of data with privacy protection. PPDM use the technology of the generalization and the concealment, the loss of connectivity of the split technology—ANGEL, and micro-aggregation technology to solve the privacy problems in the scientific data sharing. Finally, PET can ensure the security of non-sensitive data without excessive demands on the processing of data, and organizations can continue to meet the high expectations of the public in terms of services and processing of personal data by applying PET to simplify the processing of personal data.

The choice of three technologies depend on different aspects of IT capabilities, including infrastructure, organizational systems and specific cultural environments [37]. Meanwhile, these different aspects impact privacy protection, such as data sharing procedures. The three technologies mentioned above have similar interests in related fields in terms of data privacy protection, it is to avoid the disclosure of sensitive or non-sensitive data to third parties. We must know that information technologies are not growth model and need not to climb to the top. If an organization applies a certain privacy enhancement technology, this does not mean that it must continue to climb to a higher level of privacy enhancement technology. The suitability of different privacy enhancement technology choices depends on specific circumstances.
5.4. Improve the information literacy of data subjects in scientific data sharing, and establish a sense of privacy protection

Governments, organizations, research institutions and individuals need to improve their data security literacy, and establish a sense of privacy protection. They should clarify their responsibilities and obligations for data security in the process of participating in scientific data sharing, and protect data privacy [38]. Especially as the main scientific source, they should do the following:

First, they should improve their privacy security awareness and need to improve personal data risk prevention awareness during using social media or browsing the web. It is necessary to have a clear and objective understanding of privacy disputes in scientific data sharing, and try to maximize data disputes. Second, they should improve their own privacy protection ability, the relevant interest groups will use specific scenarios to induce individuals to disclose information. Some individuals will give up control of privacy during the competitive pressure, the interests driven, the system's default settings, the obscure text provisions, or even the complicated information processing. Hence, it is particularly important to improve our information literacy. It is only possible to prevent privacy violations in data sharing if the data security literacy and privacy awareness of the stakeholders are raised.

5.5. Formulate privacy protection principles for scientific data sharing

Developing privacy principles and norms for data sharing is the key to solving data abuse problems. Organizations are publish, store, and openly share scientific data that collected through researchers or other sources. As intermediaries for the release of scientific data, they need to assume certain responsibilities and formulate corresponding privacy protection terms and rules [39]. The relevant institutions should provide privacy protection principles to ensure the security of data for the protection of data privacy, such as the principles of informed consent, reasonable moderation, maximization and minimization.

The principle of informed consent refers to the right of data sources to determine whether their own information is collected by others, data collectors have the responsibility to provide the necessary information in an appropriate manner so that the data subject can make decisions; the principle of reasonable moderation refers to the reasonable and moderate collection of data under the premise of the law, the use of personal data should not be infringing on the privacy of others as the standard, it is to maintain the balance between the data subject and the interests of society; the principle of maximization and minimization means that data utility or social interests should be maximized in data sharing, while ensuring that privacy infringement is minimized, but the interests of others should not be infringed under the guise of realizing the public interest.

In addition, the International Organization for Economic Cooperation and Development to develop the eight basic principles of privacy protection: the principle of information collection restrictions, the principle of data quality, the principle of purpose, the principle of use restrictions, the principle of security protection, the principle of openness, the principle of individual participation, the principle of responsibility, these principles are also the values of organizations and countries.

6. Conclusion

Scientific data sharing increases opportunities to use scientific research data, but it also brings certain privacy protection issues. This article preliminarily explains the basic concepts of scientific data sharing, discusses and summarizes the privacy violations of the data scientific sharing, then establishes a privacy protection model and framework, last some suggestions are put forward based on identifies and analyzes the problems in order to provide reference for the privacy protection in the current scientific data sharing. In the process of scientific data sharing, how to grasp the boundaries of scientific data sharing and privacy protection, how to define the right to use scientific data sharing, and how to improve the early warning of privacy issues. These issues require further theoretical and practical research to solve. At present, China still has certain issues between scientific data sharing and privacy protection, this requires localized research on the basis of learning from foreign laws, policies and practices.
Acknowledgments
Thank you very much to my tutor, Mr. Xiangping Sheng, for putting forward the outline of this paper and suggestions for revision, and thanks Miss Lihua Wang for pointing out the mistakes for this paper.

References
[1] D. S. Sayogo, and T. A. Pardo, “Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data,” Government information quarterly, vol. 30, pp. 19-31, March 2013.
[2] G. Livraga, V. Torra and A. Aldinl, Data privacy management and security assurance, 3rd ed., vol. 2. Cham: Springer International Publishing AG, 2016 pp.18-36.
[3] A. B. Makulilo, African Data Privacy Laws, 2rd ed., vol. 33. Cham: Springer, 2016, pp.22–23.
[4] A. Tamò-Larrieux, Tamò-Larrieux, and Seyfried, Designing for privacy and its legal framework, 2d ed., vol. 12. Cham: Springer, 2018, pp.111–113.
[5] A. Solanas, and A. Martinez-Ballesté, Advances in artificial intelligence for privacy protection and security, 1d ed., vol. 1. World Scientific, 2009, pp.22–23.
[6] S. D. C. di Vimercati, S. D. C., S. Foresti, G. Livraga, and P. Samarati, Protecting privacy in data release, 2d ed., vol. 33. Cham: Springer International Publishing AG, 2011, pp.1-34.
[7] R. Caso, "The Darkest Hour: Private Information Control and the End of Democratic Science," Trento Law and Technology Research Group Research Papers, vol. A4, pp. 35–55, April 2018.
[8] W L. Prosser, “Privacy,” California Law Review, vol. 48, pp. 383–423, May 1960.
[9] A. Kirby, "The production of private space and its implications for urban social relations," Political Geography, vol. 27, pp. 74–95, February 2008.
[10] E. F. Stone - Romero, D. L. Stone, and D. Hyatt, “Personnel selection procedures and invasion of privacy,” Journal of Social Issues, vol. 59, pp. 343-368, March 2003.
[11] S. O. Dyke, and T. J. Hubbard, “Developing and implementing an institute-wide data sharing policy,” Genome medicine, vol. 39, pp. 1–8, July 2011.
[12] L. M. Jahneke, and A. Asher, “Dilemmas of Digital Stewardship: Research Ethics and the Problems of Data Sharing,” Research Data Management, vol. 80, pp. 21–28, October 2013.
[13] J. K. Goodman, C. E. Cryder, and A. Cheema, “Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples,” Journal of Behavioral Decision Making, vol. 26, pp. 213–224, December 2013.
[14] S. Vavilis, M. Petković, and N. Zannone, “A severity-based quantification of data leakages in database systems,” Journal of Computer Security, vol. 24, pp. 321–345, March 2016.
[15] S. Cho, J. Kim, and B. C. Ha, “A Solution to Privacy Violation Problem in Korean Ubiquitous Government Service,” Indian Journal of Science and Technology, vol. 8, pp. 1–68, December 2015.
[16] S. Berry, S. Fazzio, Y. Zhou, B. Scott, and L. Francisco - Revilla, “Netflix recommendations for groups,” Proceedings of the American Society for Information Science and Technology, vol. 47, pp. 1–3, March 2010.
[17] J. Isaak, and M. J. Hanna, “User data privacy: Facebook, Cambridge Analytica, and privacy protection,” Computer, vol. 51, pp. 56–59, April 2018.
[18] C. Colwill, “Human factors in information security: The insider threat–Who can you trust these days?,“ Information security technical report, vol. 14, pp. 186–196, January 2009.
[19] S. Chakravarty, and R. Ray, “Is it Heterogeneous Investor Beliefs or Private Information that Affects Prices and Trading,” Sugato Chakravarty, vol. 33, pp. 1–49, July 2010.
[20] J. Bendix, J. Nieschulze, and W. K. Michener, “Data platforms in integrative biodiversity research,” Ecological Informatics, vol. 11, pp. 1–4, February 2012.
[21] L. Akers, and J. S. Gordon, “Using Facebook for large-scale online randomized clinical trial recruitment: effective advertising strategies,” Journal of medical Internet research, vol. 20, pp. 21–29, October 2018.
[22] J. Boullata, “Natural health product interactions with medication,” Nutrition in clinical practice,
[23] H. Postigo, “America Online volunteers: Lessons from an early co-production community,” International Journal of Cultural Studies, vol. 12, pp. 451–469, December 2009.

[24] R. Noumeir, A. Lemay, and J. M. Lina, “Pseudonymization of radiology data for research purposes,” Journal of digital imaging, vol. 20, pp. 284–295, October 2007.

[25] S. Yu, “Big privacy: Challenges and opportunities of privacy study in the age of big data,” IEEE access, vol. 4, pp. 2751–2763, March 2016.

[26] K. Nissim, and A. Wood, “Is Privacy Privacy?,” Philosophical Transactions of the Royal Society A Mathematical Physical & Engineering Sciences, vol. 376, pp. 201–228, December 2018.

[27] M. H Joo, S. Yoon, H. Y. Kwon, and J. I. Lim, “De-identification policy and risk distribution framework for securing personal information,” Information Polity, vol. 23, pp. 195–219, March 2018.

[28] B. Boliek, “ISPs question RIAA suits,” Hollywood Reporter-International Edition, vol. 2, pp. 38–48, December 2003.

[29] Z. Sardar, “Postnormal times revisited,” Futures, vol. 67, pp. 26–39, May 2015.

[30] B. Malin, and K. Goodman, “Between access and privacy: challenges in sharing health data,” Yearbook of medical informatics, vol. 27, pp. 54–55, May 2018.

[31] M. Chen, Y. Qian, J. Chen, K. Hwang, S. Mao, and L. Hu, “Privacy protection and intrusion avoidance for cloudlet-based medical data sharing,” IEEE transactions on Cloud computing, vol. 22, pp. 1237–1265, September 2016.

[32] G. K. Gupta, “Computer literacy: Essential in today's computer-centric world,” ACM SIGCSE Bulletin, vol. 38, pp. 115–119, November 2006.

[33] A. Ali-Eldin, J. van den Berg, and H. A. Ali, “A risk evaluation approach for authorization decisions in social pervasive applications,” Computers & Electrical Engineering, vol. 55, pp. 59–72, December 2016.

[34] C. P. Garrison, and M. Ncube, “A longitudinal analysis of data breaches,” Information Management & Computer Security, vol. 19, pp. 105–111, February 2011.

[35] P. De Hert, and V. Papakonstantinou, “The proposed data protection Regulation replacing Directive 95/46/EC: A sound system for the protection of individuals,” Computer law & security review, vol. 28, pp. 130–142, April 2012.

[36] N. Swartz, “Canada reviews PIPEDA,” Information Management, vol. 41, pp. 7–8, April 2007.

[37] E. J. Huth, “Prepublication release of scientific data and the right of the public to know: adapting to our times,” Journal of infectious diseases, vol. 159, pp. 407–411, June 1989.

[38] S. E. Fienberg, and J. Jin, “Statistical Disclosure Limitation for Data Access,” Encyclopedia of Database Systems, vol. 22, pp. 33–48, May 2018.

[39] M. F. Gan, H. N. Chua, and S. F. Wong, “Privacy Enhancing Technologies implementation: An investigation of its impact on work processes and employee perception,” Telematics and Informatics, vol. 38, pp. 13–29, October 2019.