Postprint: Research on China's Data Sovereignty Policy Based on LDA Model and Policy Tools

Authors: Qiao Han, Xu Junru

Date: 2024-05-18T00:00:00+00:00

Abstract

In the dual context of digital economy development and the holistic national security concept, data sovereignty has become an important component of national sovereignty. Major countries and regions are actively conducting strategic deployments for data sovereignty, engaging in intense competition and contestation over data resources, data technologies, and data rules. This article employs policy text analysis methods to study China' s data sovereignty policies, utilizes LDA (Latent Dirichlet Allocation) topic models and policy instruments to quantitatively analyze the process evolution and thematic characteristics of China' s data sovereignty policies, and comprehensively considering the global data sovereignty landscape, proposes four policy recommendations: actively lead and participate in international rule-making; optimize the data cross-border security assessment process; improve standard contract templates for personal information cross-border transfer; strengthen the rule-of-law guarantees for data security.

Full Text

Research on Chinese Data Sovereignty Policy Based on LDA Model and Policy Instruments

QIAO Han* XU Junru

1. School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China

2. MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation, University of Chinese Academy of Sciences, Beijing 100190, China

Abstract

Against the dual backdrop of digital economy development and the overall national security concept, data sovereignty has become an important component of national sovereignty. Major countries and regions are actively deploying data sovereignty strategies and engaging in fierce competition and 博弈 over data resources, data technologies, and data rules. This study employs policy text analysis to examine China' s data sovereignty policy, utilizing Latent Dirichlet Allocation (LDA) topic modeling and policy instruments to quantitatively analyze the evolutionary process and thematic characteristics of Chinese data sovereignty policy. Considering the global data sovereignty landscape, we propose four policy recommendations: actively leading and participating in international rule-making; optimizing data export security assessment procedures; improving standard contract templates for personal information export; and strengthening legal protection for data security.

Keywords: data sovereignty, data security, policy instruments, LDA model

DOI: 10.16418/j.issn.1000-3045.20231115001
CSTR: 32128.14.CASbulletin.20231115001

Data sovereignty refers to a nation' s dominant authority to generate, disseminate, manage, control, utilize, and protect network facilities, data subjects, data behaviors, and data resources within its jurisdictional territory, as well as related data products [1-3]. It is becoming an essential component of national sovereignty in the digital era. The United States, the European Union, China, and other countries and regions have recognized the strategic value of data resources and are actively deploying data sovereignty strategies. Due to differences in technological capabilities and economic development conditions, countries have adopted differentiated data sovereignty policies.

Data sovereignty policy plays a crucial role in safeguarding national security and protecting national interests, attracting considerable academic attention in recent years. Some scholars have proposed the concept of data sovereignty from the perspective of state actors [4-7], while others have defined data sovereignty from the standpoint of organizational and individual subjects [8,9]. Ran Congjing [10] qualitatively summarized the global landscape of data sovereignty strategic deployment. Zheng Lin et al. [11] explained the concept of national data sovereignty and analyzed European and American data sovereignty strategies. However, existing research primarily employs qualitative induction methods to analyze data sovereignty strategies at the macro national level, lacking quantitative analysis of the connotations and characteristics of data sovereignty policy texts. This study adopts policy text analysis to examine China' s data sovereignty policy, using LDA topic modeling and policy instruments to quantitatively analyze the evolution and trends of Chinese data sovereignty policy, and proposes countermeasures and recommendations based on international comparisons with global data sovereignty policies. This research not only expands the application of policy text analysis to the field of data sovereignty but also

provides methodological support for optimizing the design of China' s data sovereignty policy.

1. Construction and Design of the Analytical Framework for Data Sovereignty Policy Research

1.1 Research Methods and Framework Dimensions

This study employs LDA topic modeling to analyze and quantify China' s data sovereignty-related policies from 2010 to 2022 from the perspective of policy instruments. We have established an analytical framework for data sovereignty policy instruments [Figure 1: see original paper]. The framework involves statistical analysis of policy texts, focusing on policy timing, issuing agencies, and policy type distribution to explore the evolutionary process, institutional distribution, and policy effectiveness of data sovereignty policies.

The classification of supply-based, environment-based, and demand-based policy instruments is most common in existing research [12]. This study adopts this dimensional division of policy instruments to conduct coding analysis of data sovereignty policies. LDA topic modeling is used to mine core thematic terms in data sovereignty policy texts for visualization. Policy text analysis represents a primary paradigm in policy research, focusing on policy connotations and instruments, using quantitative analysis to provide in-depth interpretation and analysis of textual provisions, which is significant for policy interpretation and implementation.

LDA topic modeling is a natural language processing model for probabilistic topic modeling of massive long-text data, which can be abstracted as a three-layer Bayesian model of "document-topic-word." As a typical method for identifying document topic distribution, LDA topic modeling has wide applications in text topic mining and text sentiment analysis. Existing research has classified policy instruments from different perspectives, including the definition of supply-based, environment-based, and demand-based policy instruments, as well as the division between market-based and administrative instruments. The classification of supply-based, environment-based, and demand-based policy instruments is most common in existing research, as it comprehensively considers government intervention methods and degrees.

1.2 Data Sources and Coding

This study selects the "Peking University Law Database" as the source for collecting data sovereignty-related policy texts, conducting full-text searches using keyword combinations such as "data sovereignty," "data security," "data cross-border," and "national security." We screened "central regulations" as the analytical texts, collecting a total of 45 policy documents (Table 1 shows partial data sovereignty policy texts). We extracted and coded the 45 central regulation policy documents included in the sample (partial encoding of data sovereignty

policy text content is shown in Table 2) to more clearly label the policy instruments contained in different policies and specific provisions, transforming unstructured text data for quantitative analysis.

2. Statistical Analysis of Data Sovereignty Policy Texts

2.1 Temporal Evolution Distribution of Policies

According to the statistical results of policy text publication dates, China' s data sovereignty-related policies began with the "Twelfth Five-Year Development Plan for the Internet Industry" issued by the Ministry of Industry and Information Technology in May 2012. The term "data sovereignty" first appeared in the "Outline for Promoting Big Data Development" issued by the State Council in 2015, which stated that China should "enhance its capacity to protect data sovereignty in cyberspace, safeguard national security, and effectively enhance national competitiveness." However, the concept of internet sovereignty had already emerged as early as 2010. The number of data sovereignty policy releases saw two significant increases in 2016 and 2021 [Figure 2: see original paper]. Combined with the emergence and rapid development of new technologies such as big data, cloud computing, and blockchain, the evolution of data sovereignty policy can be divided into four stages:

(1) **Internet Sovereignty Stage (2010-2013).** In 2010, the State Council Information Office released the white paper "The Internet in China," stating that "the Internet within the territory of the People's Republic of China is under China' s sovereign jurisdiction, and China' s Internet sovereignty should be respected and safeguarded." China gradually improved Internet laws and regulations, actively promoted Internet infrastructure construction, strengthened Internet management and supervision, ensured Internet security, and promoted the development of the Internet industry.

(2) **Information Sovereignty Stage (2014-2015).** On July 16, 2014, President Xi Jinping delivered a speech titled "Carrying Forward Traditional Friendship and Jointly Creating a New Chapter of Cooperation" at the Brazilian Congress, emphasizing that no country' s sovereignty and interests in the information field should be violated. During this stage, relevant policies emphasized strengthening information autonomy and control, building information infrastructure, promoting research and development of information security technologies and standards, and enhancing China' s actual control capabilities and international discourse power in the field of information sovereignty.

(3) **Cyberspace Sovereignty Stage (2015-2018).** The National Security Law of the People' s Republic of China, promulgated and implemented in July 2015, first explicitly defined the concept of cyberspace sovereignty. In the "International Cooperation Strategy on Cyberspace" released in March 2017, "firmly safeguarding China' s cyberspace sovereignty, security, and development interests" was identified as one of China' s strategic

goals for participating in international cyberspace cooperation. In May 2018, the Cyberspace Administration of China released the "Digital China Construction and Development Report (2017)," which listed "safeguarding cyberspace sovereignty" as a situation and direction for Digital China construction. During this stage, relevant policies emphasized strengthening cybersecurity protection, establishing cyberspace sovereignty management mechanisms, and advancing cyberspace governance to ensure that the state has the rights and capabilities to protect national security and safeguard national interests in cyberspace.

(4) **Data Sovereignty Stage (2019-present).** In July 2019, the "Guiding Opinions on Strengthening Industrial Internet Security Work" stated that an industrial Internet data classification and grading management system should be established based on industrial categories, data types, and data values, and that security assessments and monitoring of important data export should be carried out. In September 2020, China launched the "Global Data Security Initiative," calling on countries not to directly access data located in other countries from enterprises or individuals without the permission of other countries' laws. The Data Security Law of the People' s Republic of China, which came into effect in September 2021, stipulates that safeguarding data security should adhere to the overall national security concept, establish and improve the data security governance system, and enhance data security protection capabilities. In the same year, the Personal Information Protection Law of the People' s Republic of China also made provisions for the cross-border flow of personal information, data localization storage, and extraterritorial effects. During this stage, relevant policies emphasized establishing data classification and grading systems, clarifying data export security assessment requirements, and advancing data protection capability certification mechanisms to ensure effective protection and compliant use of data during cross-border flow.

In summary, China' s data sovereignty strategy has evolved through the stages of "Internet sovereignty—information sovereignty—cyberspace sovereignty—data sovereignty," with overall policy characteristics focusing on security as the foundation, promoting regulation and institutional development, facilitating orderly data flow, and strengthening risk assessment and supervision of data cross-border flow processes and post-export scenarios.

2.2 Distribution of Policy-Issuing Agencies

Policy-issuing agencies reflect the effectiveness level of various policies and regulations. A total of 33 agencies, including the State Council, Ministry of Industry and Information Technology, Ministry of Commerce, and Standing Committee of the National People' s Congress, have participated in data sovereignty policy formulation. Table 3 shows the distribution of core agencies issuing policies. China' s data sovereignty policies are issued in two forms: joint issuance and

individual issuance. Among the 45 central regulation policy documents, 8 were jointly issued and 37 were individually issued, with individual issuance accounting for 82.2%.

Based on statistical results, agencies that issued more policies include the State Council, Ministry of Industry and Information Technology, and Ministry of Commerce. Joint issuances mainly involved the National Development and Reform Commission and the Ministry of Industry and Information Technology. There are five core issuing agencies, accounting for 15.2%, while non-core agencies number 28, accounting for 84.8%. The distribution of policy subjects is relatively dispersed overall. In general, numerous entities participate in data sovereignty policy formulation, but the core agencies account for 54.2% of total issuances, showing relatively high concentration. The overall structure features five central issuing agencies: the State Council, Ministry of Industry and Information Technology, Ministry of Commerce, Standing Committee of the National People' s Congress, and Cyberspace Administration of China .

2.3 Policy Type Distribution

The types of documents used for policy issuance vary according to policy nature and objectives, resulting in different policy effectiveness [13,14]. Among the policy samples collected in this study, data sovereignty policies were issued in nine different forms, including plans, opinions, laws, notices, outlines, and measures, demonstrating formal diversity . Among them, plans, opinions, and outlines account for relatively high proportions, indicating that existing policies contain numerous viewpoints and handling opinions from relevant departments. However, measures, which are more targeted and specific, are less common in the overall policy distribution. The National People's Congress and its Standing Committee have promulgated four laws, accounting for 8.9%.

3. Two-Dimensional Analysis of Data Sovereignty Policy

Policy instruments are specific measures and means adopted by policy subjects to achieve policy objectives. Based on existing research [15-17], this study constructs an analytical dimension for data sovereignty policy instruments from three aspects: demand-based, environment-based, and supply-based instruments. Through policy instrument identification, we identify important pathways and safeguards for policy implementation. Based on LDA topic modeling, we conduct policy thematic analysis to reveal the core content and main objectives of policies, providing decision-making references for further optimizing and improving national data sovereignty policy formulation.

Considering that one policy analysis unit may simultaneously apply multiple policy instruments, the total number of policy instruments used for analysis in this study exceeds the total number of coding units. In terms of policy instrument application types, China' s data sovereignty policy employs a relatively comprehensive set of instruments, with demand-based, environment-based, and

supply-based policy instruments accounting for 21.6%, 59.6%, and 18.8% respectively [Figure 3: see original paper]. Among them, environment-based policy instruments are most commonly used, while demand-based and supply-based policy instruments are applied less frequently. China employs multiple policy instruments to jointly promote secure cross-border data flow and data sovereignty protection.

3.1 Empirical Analysis of Policy Instrument Dimensions

(1) Demand-based policy instruments aim to introduce various forces for exchange and cooperation, actively conduct pilots in data security management, cross-border flow, and international cooperation rules, highlight important areas and key links, thereby promoting secure and orderly data flow . Demand-based data sovereignty policy instruments focus on pilot demonstrations and international cooperation, emphasizing the accumulation of effective experience in data security management and data sovereignty protection through pilot demonstrations and international cooperation development models, which are then promoted and popularized throughout society. They focus on guiding social forces to participate, conducting promotional activities, thereby promoting industry-standard development and raising data sovereignty protection awareness.

(2) Environment-based policy instruments aim to guide and supervise various subjects to adopt more proactive data security management and data sovereignty protection measures by optimizing environmental conditions, policy orientation, standard specifications, and accountability mechanisms for implementation by various subjects and links . Overall, environment-based policy instruments account for more than half of the total, making them a preferred indirect regulatory means [19] for policy subjects. Regulation is a commonly used method. With the development of new technologies and the continuous expansion of data scale, issues such as international data sovereignty threats have become increasingly prominent, requiring strengthened regulation to actively respond to extraterritorial effects and other issues. Standard specifications, security norms, safeguard measures, and review assessments are policy documents that are gradually being improved to guide the standardized development of emerging digital industries and ensure secure and orderly data flow, including cross-border circulation guidelines, security assessments, and safeguard mechanisms. Strategic measures play a supplementary role in the policy system. Currently, intellectual property policy instruments are rarely applied, representing a clear gap in data sovereignty policy.

(3) Supply-based policy instruments aim to provide various resource supports for data security management, data sovereignty protection, information infrastructure construction, technology research and development, talent cultivation, and overall coordination of various subjects . Among supply-based policy instruments, information infrastructure construction, technical support, and organizational construction account for relatively high proportions. With the emergence of new technologies such as big data, cloud computing, and

blockchain, information facilities have become important foundational conditions for digital economy development. Both the "14th Five-Year Plan for National Economic and Social Development and Long-Range Objectives Through the Year 2035" and the Cybersecurity Law of the People' s Republic of China propose establishing and improving critical information infrastructure protection systems. Technical support is an important force for ensuring data security and should build a secure and reliable data circulation environment through developing core technologies. Under the overall coordination of organizational construction, information infrastructure construction and technical support complement each other, providing continuous momentum for secure data flow and data sovereignty protection. Currently, talent support and financial support policy instruments are less applied, accounting for 6.67% and 8.89% respectively. Financial support has directional advantages, while talent cultivation has long-term benefits, and the application of these two policy instruments should be appropriately increased.

3.2 LDA Topic Model Analysis

3.2.1 Topic Number and Coherence Analysis of the LDA Model The LDA topic model can be used to mine potential thematic features of large volumes of text. It is an unsupervised, unstructured probabilistic model that utilizes the regularity of semantic associations in word co-occurrence, discovering potential topics in document corpora without pre-setting dictionaries or topic categories [20], thereby reducing the influence of researchers' subjective judgments on topic classification [21]. This study treats each coding unit as a document and selects the topic coherence index model to determine the optimal number of topics. If coherence is at a high level, the topic structure of the model will be more stable. This study uses the "Harbin Institute of Technology Stopword List" to preprocess documents (Chinese word segmentation, stopword removal). The results of topic number and coherence are shown in [Figure 4: see original paper]. Through topic evaluation, this study selects 3 as the number of topics and uses the pyLDAvis algorithm for visualization analysis and presentation of LDA topic model results.

3.2.2 Word Cloud Visualization [Figure 5: see original paper] shows the bag of words for data sovereignty policy mined based on the LDA topic model. According to the visualization results of the LDA topic model, current data sovereignty policy is divided into three categories: data security and personal information protection; data cross-border and international cooperation; and data security assessment and data export.

(1) Data Security and Personal Information Protection [Figure 5a: see original paper]. Promote the construction of a comprehensive data security protection system, including improving data classification and grading systems and personal information authorized use systems. To form a benign development pattern of data resources convergence and sharing and secure and orderly data

flow, the relationship among national security, data cross-border circulation, and personal privacy protection should be properly handled.

(2) Data Cross-Border and International Cooperation [Figure 5b: see original paper]. Clarify the extraterritorial application effect of the Data Security Law, further improve data cross-border circulation norms, and implement pilots for data cross-border transmission and security control. Explore participation in the formulation of regional international data cross-border circulation rules, promote the formation of a global collaborative mechanism for data cross-border circulation, and strengthen secure collaboration and information resource sharing between China and other countries and regions worldwide.

(3) Data Security Assessment and Data Export [Figure 5c: see original paper]. Improve the data classification and grading management system and data export security review mechanism. Ensure that data can undergo comprehensive pre-export assessment, continuous supervision, and risk self-assessment to effectively identify and prevent security risks brought by data export and protect national and personal sensitive information. Explore the design of data protection capability certification mechanisms to provide objective and credible assessment standards for data export, ensuring data legality, security, and controllability.

In summary, China' s data sovereignty policy primarily employs environment-based policy instruments, supplemented by demand-based and supply-based policy instruments, to promote secure cross-border data flow and data sovereignty protection. Policy text content exhibits three major thematic characteristics: "data security and personal information protection," "data cross-border and international cooperation," and "data security assessment and data export." China has successively formulated laws such as the Personal Information Protection Law and the Data Security Law, issued regulations including the Measures for Data Export Security Assessment, and implemented strict data export and information protection management models to ensure the protection of China' s data sovereignty.

4. Implications and Recommendations

Based on the analysis of policies in the data sovereignty field, this study proposes four policy recommendations for China' s secure cross-border data flow and data sovereignty protection:

(1) Actively lead and participate in international rule-making to establish mutual trust mechanisms and enhance international discourse power. China should actively lead and constructively participate in the formulation and improvement of international rules and standards in the digital era, establishing multi-form data cooperation platforms with a broader range of countries to provide technical support and capacity building in data infrastructure construction, data resource development and utilization, and data security protection, promoting data interconnection and sharing for mutual benefits. China

should promote cooperation and discussions with other countries or regions on data protection levels, standards, and norms, striving for more consensus and agreements to achieve mutual recognition or reciprocal appropriateness of data protection standards and norms.

(2) Optimize data export security assessment procedures to improve assessment efficiency and accuracy. Establish a risk-oriented data classification management system, adopt differentiated control measures for different types and levels of data, strengthen the formulation and implementation of data security assessment standards, clarify requirements and guidance for data security assessment, and ensure security and credibility during data export processes. Utilize advanced technical means to improve the efficiency and accuracy of data export security assessment. For example, artificial intelligence and big data analysis technologies can be introduced to automate assessment processes, quickly identifying high-risk data export behaviors to improve the accuracy and reliability of assessment results.

(3) Improve standard contract templates for personal information export to enhance compliance operation efficiency. The Measures for Standard Contracts for Personal Information Export came into effect on June 1, 2023, aiming to ensure the legal, secure, and orderly cross-border transmission of personal information. By using standard contract templates, China can "extend" its domestic jurisdiction to foreign territories through the legal binding force granted by the Measures, achieving a certain "extraterritorial application of domestic law" effect and realizing cross-border personal information flow protection. In the future, standard contract templates can be modularly expanded to develop more optional modules based on the business needs of organizations, enterprises, or individuals, reducing compliance operation costs for relevant entities and improving efficiency.

(4) Strengthen legal protection for data security to build a data sovereignty defense line. Guided by the overall national security concept, China should improve laws and regulations to ensure data security and strengthen the data security governance system. It should clarify legal responsibilities for data security, protect critical data infrastructure, establish data security risk assessment and emergency response mechanisms, conduct data sovereignty and security publicity and education to raise public data security awareness and capabilities, and promote international cooperation. China should strengthen exchanges and mutual learning with other countries and regions on data security laws and regulations, jointly promote the formulation of international data security standards, and enhance China' s discourse power and influence in global data governance.

References

1. Hong Y M, Zhang M, Liu Y. Promoting safe and orderly flow of cross-border data to lead development of globalization of digital economy. Bul-

letin of Chinese Academy of Sciences, 2022, 37(10): 1418-1425. (in Chinese)

2. Ran C J, Liu Y. Outline of data sovereignty genealogies. Wuhan University Journal (Philosophy & Social Science), 2022, 75(6): 19-29. (in Chinese)
3. He B. The development, challenges, and countermeasures of data sovereignty. Journal of Cyber and Information Law, 2019, (1): 201-216. (in Chinese)
4. Rainie S C, Schultz J L, Briggs E, et al. Data as a strategic resource: Self-determination, governance, and the data challenge for indigenous nations in the United States. International Indigenous Policy Journal, 2017, 8(2), doi: 10.18584/iipj.2017.8.2.1.
5. Hippelainen L, Oliver I, Lal S. Towards dependably detecting geolocation of cloud servers// International Conference on Network and System Security. Cham: Springer, 2017: 643-656.
6. Sun N X, Zhang X J. On data security—An examination of game and cooperation in cyberspace. Pacific Journal, 2015, 23(2): 63-71. (in Chinese)
7. Qi A M, Pan J. Right to data, data sovereignty and the basic principle of big data protection. Journal of Soochow University (Philosophy & Social Science Edition), 2015, 36(1): 64-70. (in Chinese)
8. Jarke M, Otto B, Ram S. Data sovereignty and data space ecosystems. Business & Information Systems Engineering, 2019, 61(5): 549-550.
9. Alboaie S, Cosovan D. Private data system enabling self-sovereign storage managed by executable choreographies// IFIP International Conference on Distributed Applications and Interoperable Systems. Cham: Springer, 2017: 83-98.
10. Ran C J. Global situation of data sovereignty governance and China' s response. People' s Tribune, 2022, (4): 24-27. (in Chinese)
11. Zheng L, Li Y, Wang Y F. Research on national data sovereignty strategies in the new era. Information Studies (Theory & Application), 2022, 45(6): 55-60. (in Chinese)
12. Lv W J, Chen J, Liu J. A quantitative analysis of China' s Artificial Intelligence industry policy from the perspective of policy tools. Studies in Science of Science, 2019, 37(10): 1765-1774. (in Chinese)
13. Chen L, Duan Y Q. Analyzing implementation of the Chinese government open data policy using government bulletin text as example. Journal of the China Society for Scientific and Technical Information, 2020, 39(7): 698-709. (in Chinese)
14. Huang R H, Wen F F. Policy framework and content of opening and sharing government data in China: A content analysis of policy documents

at the national level. Library and Information Service, 2017, 61(20): 12-25. (in Chinese)

15. Rothwell R. Reindustrialization and technology: Towards a national policy framework. Science and Public Policy, 1985, 12(3): 113-130.
16. Huang C, Xu L, Zhong X T, et al. A framework of integrated policy-technology roadmap (P-TRM) and its use: Examples of wind turbine and solar PV industries. China Soft Science, 2014, (5): 76-84. (in Chinese)
17. Fu L, Zhang D Y, Yan H B, et al. Research on China' s policy instruments of carbon emission mitigation through policy content analysis method. Studies in Science of Science, 2023, 41(3): 435-443. (in Chinese)
18. Li X, Chen Q, Jia H M. Research on evolution of the smart city policy system in China. Science Research Management, 2022, 43(7): 1-10. (in Chinese)
19. Shi Y H, Xue X. Research on information security policy in China from the perspective of policy tools—Based on the quantitative analysis of 81 policy texts. Journal of Modern Information, 2022, 42(1): 130-138. (in Chinese)
20. Zhang C, Leng F H. An analysis of the machine-learning-assisted intelligent decision-making—A study by taking "green innovation"as an example. Science Research Management, 2022, 43(9): 32-40. (in Chinese)
21. Zhang C, Leng F H. Discovery and interpretation roles: A topic modeling approach. Management Science, 2018, 64(6): 2833-2855.

QIAO Han is a Professor and Doctoral Supervisor at the School of Economics and Management, University of Chinese Academy of Sciences (UCAS), and Executive Associate Editor-in-Chief of *Management Review*. Her research focuses on digital economy, scientific and technological innovation, business model innovation, venture capital, and entrepreneurship. E-mail: qiaohan@ucas.ac.cn

*Corresponding author

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