SWOT Analysis and Strategic Study on the Development of Computational Archival Science in China: Postprint

Authors: Zhao Yue, Ma Xiaoyue, Zhang Jiaxin

Date: 2023-04-01T15:51:23+00:00

Abstract

[目的/意义] To review the development status of computational archival science and explore its development strategies in China, thereby providing references for advancing Chinese computational archival science under the background of new liberal arts construction. [方法/过程] Based on literature research, this study sorts out the current development status of computational archival science, employs SWOT analysis to analyze internal and external environmental elements such as opportunities, threats, strengths, and weaknesses in its domestic development, and formulates different development strategies through cross-matching of these internal and external elements. [结果/结论] The study finds that computational archival science has received continuous international attention, and preliminary directions for theoretical research and practical exploration have been established around basic theory, educational research, archival processing, archival analysis, and archivalization. However, overall, the development of computational archival science remains in the preliminary exploration stage. To promote computational archival science in China, emphasis should be placed on: leveraging interdisciplinary construction opportunities to build transdisciplinary research platforms; targeting strategic informatization needs in the field to form large-scale research directions; clarifying boundaries with related disciplines to highlight the characteristics of computational archival science; harnessing transdisciplinary research advantages while avoiding data security and privacy risks; seizing opportunities for cultivating interdisciplinary talents to integrate multiparty resources for jointly building teaching platforms; focusing on core issues in practical fields to explore operable technical solutions; strengthening institutional design and technical breakthroughs to conduct archival security risk assessment and control; and increasing basic research to clarify the theoretical, methodological, and technical systems of computational archival science.

Full Text

SWOT Analysis and Strategy Research on the Development of Computational Archival Science in China

Zhao Yue[1], **Ma Xiaoyue**[1], **Zhang Jiaxin**[2] [1]School of Public Administration, Sichuan University, Chengdu 610065 [2]School of Information Resource Management, Renmin University of China, Beijing 100872

Abstract:
[**Purpose/Significance**] This paper reviews the current state of computational archival science (CAS) development, explores strategies for advancing CAS in China, and provides a reference for its growth under the background of new liberal arts construction. [**Method/Process**] Based on literature research, this study examines the development status of CAS and employs SWOT analysis to dissect the internal and external environmental factors—including opportunities, threats, strengths, and weaknesses—of CAS development in China, forming different development strategies through cross-matching these elements. [**Result/Conclusion**] The study finds that CAS has received sustained international attention, with initial research and practice directions established around basic theory, educational research, archival processing, archival analysis, and archivalization. However, overall, CAS development remains in a preliminary exploration stage. Advancing CAS in China requires: leveraging interdisciplinary construction opportunities to build transdisciplinary research platforms; targeting strategic needs in field informatization to form large-scale research directions; clarifying boundaries with related disciplines to highlight CAS characteristics; utilizing transdisciplinary research advantages to mitigate data security and privacy risks; seizing opportunities for cultivating interdisciplinary talents and integrating resources to co-construct teaching platforms; exploring operable technical solutions around core practical problems; strengthening institutional design and technical breakthroughs for archival security risk assessment and control; and increasing basic research to clarify the theories, methods, and technical systems of CAS.

Keywords: computational archival science; new liberal arts; data transition; archival education
Classification Number: G270
DOI: 10.13266/j.issn.0252-3116.2022.04.006

Since 2006, cloud computing, big data, artificial intelligence, and other technologies have become primary drivers of social evolution. The technological updates, conceptual support, and research methods provided by these emerging technology industries have enabled computational thinking to transcend economic domains, spawning more related thinking patterns and cross-domain practices. "Computational + discipline" paradigms such as computational social science and computational linguistics have emerged as new disciplinary models

in the big data era. With the advent of the digital intelligence era, increasingly complex technological environments have also made people realize that traditional document and archival management practices require the introduction of computational theories and methods to achieve digital transformation. The emergence of new archival forms and archival problems requires not only the involvement of computer science and other disciplines but also the support of archival theories and methods. Under this new two-way demand, computational archival science, which emphasizes the integration of computational theoretical methods and archival theoretical methods, has inevitably emerged.

On the one hand, the production and consumption of emerging born-digital archives are determined by social and industrial trends and by computer and data methods that have little connection with archival methods. Understanding their production and consumption characteristics, governance points, and solving the processing, analysis, storage, long-term preservation, and access problems of large-scale born-digital archives requires the assistance of computer science and other disciplines, especially the application of computational methods and resources. At the same time, ensuring the authenticity, integrity, reliability, availability, and security of new forms of documents and archives also requires the involvement of archival science. Facing new archival forms, multidisciplinary collaboration is an inevitable trend. On the other hand, traditional archival work is accelerating its digital intelligence transformation, and archival data-driven research activities are emerging, making the data processing, analysis, association, mining, and other aspects of large-scale archival materials a challenge. At the same time, the demand for technology-enabled archival work is increasingly prominent, with goals such as automating and intellectualizing document filing, open appraisal, and sensitivity review to improve work efficiency gradually becoming clear. In short, facing new archival practices, traditional archival fields face limitations in theory, methods, and technology, and traditional archival theories, methods, and technologies cannot solve the challenges of large-scale processing and application brought by new archival practices. Contemporary archival work requires the introduction of computational thinking and methods.

To deeply explore advanced forms of integration between computation and archives in thinking, theory, and methods, the concept of computational archival science was proposed at the 2016 “Discovering New Knowledge: Archival Records in the Age of Big Data” symposium held at the University of Maryland, USA. Subsequently, the first IEEE Computational Archival Science workshop identified it as an interdisciplinary field. This discipline aims to apply computational methods and resources to large-scale document/archival processing, analysis, storage, long-term preservation, and access to improve efficiency, productivity, and accuracy, and to support appraisal, arrangement and description, preservation and access decisions, and research using archival materials [1]. Since then, foreign scholars have conducted continuous and extensive explorations of the academic fields involved in this concept. In recent years, domestic scholars have also noticed the development trend of computational archival science abroad. Fu Tianzhen published a paper in

2019 summarizing the development history, definition, and characteristics of computational archival science [2]. Subsequently, Zhou Wenhong et al. [3], Tao Yufang [4], Liu Yuenan et al. [5], and Yu Yingxiang et al. [6] have also conducted multi-angle examinations of the development of foreign computational archival science based on literature research methods. Zhao Yue et al. [7] analyzed the development prospects of computational archival science in China based on a survey of Chinese LIS (Library and Information Science) community's perception of computational archival science. Building on relevant research, this paper uses SWOT analysis to explore the development strategies of computational archival science in China, hoping to provide insights for further research and practice in this field.

2. Review of Current Development of Computational Archival Science

2.1 Continuous Attention as a Transdisciplinary Research Field

Liu Yuenan et al. [5] systematically traced the development process of computational archival science abroad: In 2015, Professor R. Marciano of the University of Maryland's College of Information Studies formed a small interdisciplinary research group to discuss using computational methods to solve archival problems, which is considered the origin of computational archival science internationally. The April 2016 symposium on computational archival science at the University of Maryland announced the initial formation of a cross-regional and cross-disciplinary academic community. The six conference initiators have remained core forces in this field, with invited representatives from universities, research institutions, government agencies, cultural institutions, and cooperative organizations in the UK, Canada, South Africa, and the US. After four years of development, the computational archival science community has further absorbed multiple universities in the US, state archives, as well as researchers from the University of Brasília, the Alan Turing Institute in the UK, the University of Amsterdam, the Central University of Gujarat in India, the Indian Institute of Management, Kyushu University in Japan, and the University of Canberra. The academic community has continuously expanded its scale and deepened its exploration through academic conferences, special issues, and collaborative research, promoting the development of computational archival science.

In terms of academic conferences, according to incomplete statistics, the computational archival science academic community initiated 27 academic activities in the form of workshops between 2016 and 2020, including conferences hosted by well-known computational science research institutions such as the Alan Turing Institute's 2020 computational archival science symposium. The most representative is the IEEE Big Data Computational Archival Science workshop, which began in 2016 and has been held for five consecutive sessions, reflecting the importance placed on interdisciplinary computational archival science by the big data field with computational science as its main base, and attracting more

and more scholars from computer science, archival science, information science, library science, history, art, and other fields to join, with the academic community of computational archival science continuing to expand.

The IEEE Big Data Computational Archival Science workshop has produced a total of 62 conference papers since 2016, with an increasing annual publication trend, reflecting the sustained attention trend in computational archival science. In addition, the author further searched for "computational archival science" in Google Scholar and Emerald and other academic search engines and databases (search date: February 10, 2021). After manually removing duplicate and irrelevant records, 13 additional valid foreign language documents were obtained. Through statistical analysis of document authors, it was found that 75 documents came from 243 scholars in 19 countries (regions), indicating that computational archival science has attracted the attention of researchers from many countries in its early development stage. In terms of author distribution, among the 19 countries (regions), scholars from the US were the most numerous with 163, followed by Canada (21) and the UK (9). Among the 243 scholars, those who published three or more articles were all from the library, information, and archival management disciplines, demonstrating the foundational role of these disciplines in the construction of computational archival science. In terms of author collaboration, among the 75 articles, the rate of completion by two or more scholars was 60%, reflecting a high degree of collaboration in computational archival science research. In terms of institutional collaboration, articles completed by two or more institutions accounted for 46.7%, indicating a high degree of institutional cooperation. In specific institutional cooperation patterns, 10 articles were completed through inter-university cooperation, 4 through inter-institutional cooperation outside universities, 8 through cross-department or cross-unit cooperation within universities, 8 through intra-university and extra-university inter-institutional cooperation, 3 through multi-party cooperation across universities and institutions, 1 through multi-party cooperation across departments and institutions, and 1 through multi-party cooperation across departments, universities, and institutions, once again highlighting the importance of cooperation, especially interdisciplinary cooperation, for computational archival science research.

Furthermore, the 75 documents involved a total of 138 institutions, with the University of Maryland's College of Information Studies publishing the most (17 articles), followed by the University of British Columbia's School of Information (8 articles) and King's College London's Department of Digital Humanities (5 articles). Among the 11 institutions that published more than two articles, most were university research institutions, indicating that university research institutions play a central role in promoting computational archival science. Among them, the University of Maryland's College of Information Studies, the University of British Columbia's School of Information, and King's College London's Department of Digital Humanities are core forces in computational archival science research.

The University of Maryland's College of Information Studies is committed to developing smart city technologies and creating emerging archival methods. Based on 46 research funding projects and multiple research centers (such as the Social Data Science Center, Center for Computational Linguistics and Information Processing, Human-Computer Interaction Laboratory, Center for Advanced Study of Community Information, Center for Archival Futures, and Trace R&D Center [8]), it has pioneered 14 research areas including computational archival science, digital humanities, computational linguistics, and human-computer interaction. Among them, the Digital Curation Innovation Center founded by Professor R. Marciano in 2015 [9] particularly focuses on exploring the integration of archival data and technology, developing new forms of archival analysis, and deepening the combination of historical, social, scientific, and cultural research with archives. Since its establishment, the center has collaborated with multiple parties to research more than ten interdisciplinary projects (such as the Slavery Legacy Project with the Maryland State Archives [10] and the WWII Archives Project with NARA [11-12]), becoming a core force in promoting computational archival science.

The University of British Columbia's School of Information has built the Kitimat Laboratory, Terrace Laboratory, and Greg Laboratory equipped with intelligent devices for website research or focus group discussions, and has established specialized technical consultants to provide personalized assistance covering multiple technical topics such as SQL databases, website development, programming, and prototyping. V. Lemieux, N. Payne, and their respective teams in the archival science direction have conducted innovative explorations in blockchain, artificial intelligence, and other fields, making them core forces in computational archival science. V. Lemieux's team developed a blockchain-based disposition application and "ArchContracure" smart contract [13] and applied it to land transactions, medical records, and financial document management [14]. Dr. N. Payne designed a system that emphasizes both classification accuracy and document connection [15] and developed a novel contextual information capture framework supporting automatic document classification [16].

King's College London's Department of Digital Humanities is dedicated to researching digital culture and society and advanced technological methods for humanities and social science research [17], establishing three main directions: digital culture and digital media, digital methods and digital devices, and digital community engagement platforms and channels. To compensate for deficiencies in software engineering and technology management, an independent Digital Laboratory with a software engineering team was established in 2015. Based on practical needs in different industries, the laboratory has assembled a team comprising research software analysts, engineers, UI/UX designers, project managers, and system managers, undertaking more than 100 digital humanities projects (such as the European Holocaust Research Infrastructure (EHRI) [18] and the European Big Data and Social Mining Research Infrastructure Project [19]). In addition, in 2019, King's College London's Department of Digital Humanities collaborated with the University of Maryland's College of Informa-

tion Studies, the Maryland State Archives, and the UK National Archives to establish the Computational Archival Science International Research Cooperation Network to conduct a one-year collaboration to further promote interdisciplinary exploration and practice in computational archival science [20].

2.2 Initial Establishment of Theoretical Research and Practical Exploration Directions

Since 2016, the IEEE Big Data Computational Archival Science workshop has formed relatively stable discussion themes: application of analysis in archival materials, including text mining, data mining, sentiment analysis, and network analysis; analysis supporting archival processing, including e-discovery, personal information identification, appraisal, arrangement and description, access, digital curation, semantics, ontology, linked data, topic modeling, natural language processing, machine learning, etc.; scalable archival services, including identification, preservation, metadata generation, integrity checking, normalization, reconciliation, linked data, entity extraction, anonymization, and reduction; new archival forms, including web, social media, audio-visual archives, and blockchain; network infrastructure for archive-based research and collection development and hosting; big data and archival theory and practice; digital curation and preservation; crowdsourcing and archives; big data and memory and identity construction; specific big data technologies (e.g., NoSQL databases) and their applications; corpora and reference collections for big archival data; linked data and archives; big data and provenance; constructing big data research objects from archives; legal and ethical issues in big data archives.

These topics initially listed directions for computational archival science-related practical exploration. Later, R. Marciano et al. summarized eight typical practices that drive computational archival science research: evolutionary prototyping and computational linguistics; graphic analysis and digital humanities; computer retrieval tools; digital curation; public participation in (archival) content; authenticity; network infrastructure and records continuum; spatial and temporal analysis [21], further outlining the "territorial scope" of computational archival science and contributing to the formation of some core research areas such as archival material analysis, new form archives development, expanded services for archival processing, and big data and archival theory and practice [22]. However, it is very difficult to draw clear boundaries for computational archival science among numerous computational and archival studies. M. Lee et al. proposed a heuristic method for evaluating computational archival science research to assess whether research questions belong to the core of this field, arguing that "solving archival problems with computational thinking" does not necessarily belong to the scope of computational archival science. Computational archival science research should take the common goals of archival and computational problems as the entry point, integrating archival and computational theories to form its own expertise and new theories [23]. This method

provides some inspiration for identifying core issues in computational archival science but cannot be used to accurately delineate its research fields and boundaries.

Through analysis of domestic and foreign literature research topics in computational archival science, the author believes that the field has initially established five directions: basic theory, educational research, archival processing, archival analysis, and archivalization. Basic theory research in computational archival science is committed to explaining related concepts, characteristics, research frameworks, and disciplinary attributes. For example, in terms of concepts, in 2018, R. Marciano et al. updated the initial definition proposed at the 2016 IEEE Computational Archival Science workshop [1], changing "interdisciplinary" to "transdisciplinary" [21], emphasizing the integration of disciplinary knowledge. Later, some scholars further expanded and elaborated on the definition of computational archival science [22]. However, the definition of computational archival science is still evolving, and the current definition does not fully reflect the knowledge exchange between basic disciplines under the transdisciplinary framework and remains limited. Regarding disciplinary attributes, it is generally believed that computational archival science is a two-way interaction between computer science and archival science, a new transdisciplinary field created through the recombination and integration of their elements. However, some scholars propose that it should be based on archival science, information science, and computer science [22], and some even point out that computational archival science is not a new scientific field but merely an expanding archival science direction of information technology methods [24]. In addition, some scholars emphasize the engineering attributes of computational archival science, further proposing the concept of archival engineering, believing that the value of computational archival science can only be realized when providing products and services [25].

The educational research direction of computational archival science mainly focuses on issues related to computational thinking and archival thinking education, training, and curriculum design. H. Stancic' et al. found through investigation that under the influence of information and communication technology, archival science professional curricula in European universities expanded to information systems and digital preservation from 2003-2016. They believe that archivists need to further learn technologies such as semantic web, graph databases, and machine learning [26]. The University of Maryland's College of Information Studies proposed computational thinking composed of 22 computational practices, divided into four categories: data practices, modeling and simulation practices, computational problem-solving practices, and systems thinking practices [27]. They mapped different knowledge units in archival courses to computational thinking and constructed two ways to associate computational thinking with LIS master's education: one is to create new courses to teach computational thinking in relevant knowledge areas; the other is to introduce computational thinking into graduate course examples, exercises, and projects [28-29]. In addition, to promote in-depth professional training, the University of Maryland's College of Information Studies is committed to building a net-

work platform for computational archival science education systems to showcase, share, and teach practices for archivists and researchers, enabling educators and practitioners to learn from each other through project profiles, lesson plans, and case documents [30].

The archival processing direction of computational archival science mainly discusses issues related to archival material processing, such as digitization, e-discovery, information identification, appraisal, classification, arrangement, description and access, digital curation, semantic ontology, linked data, topic modeling, natural language processing, and machine learning. For example, in digitization, the European Data Infrastructure (EUDAT) used OCR technology to digitize plant specimen images, adopted integrated computational analysis and transferred them to trusted digital repositories to achieve sharing and long-term preservation of research data [31]. In classification, N. Payne compared methods for automatic classification of digital archives and proposed designing a system that emphasizes both classification accuracy and document connection [15], and also proposed a metadata framework that integrates judicial, historical, procedural, business, and technical elements using machine learning methods to achieve automatic document classification [16]. In appraisal, the University of Michigan Library created an evaluation tab tool for automatic identification and appraisal of sensitive information in large-scale digital archives [32]. T. Hutchinson proposed using natural language processing technology for topic modeling to help identify privacy information in review documents [33] or using supervised machine learning to identify personal information for privacy risk control [34]. In description and organization, EHRI collected trustworthy and usable hierarchical archival metadata to integrate large amounts of scattered Holocaust-related materials and created an API catalog to achieve metadata collection, association, and retrieval functions, providing cross-border access to metadata through a portal website [18, 35]. The University of Maryland processed unstructured data with automatic indexing, format conversion of data files to enable access to data from different systems, and built an annotated corpus by orchestrating conversion and extraction sequences to describe text images [36].

The archival analysis direction of computational archival science mainly discusses the analysis of traditional and emerging archival materials, including text mining, data mining, sentiment analysis, and network analysis. For example, in text and data mining, T. Blanke used automated analysis and topic modeling of word frequency in "distant reading" to identify changes in UK government white paper language over nearly 80 years and conducted era classification and political pattern evolution analysis of archival texts [37]. The Maryland State Archives and the University of Maryland's Digital Curation Innovation Center used crowdsourcing to code and collect multi-type, scattered documents from more than 30 archival series in the Slavery Legacy Project, and used visualization tools to analyze relationships among more than 420,000 enslavement archival data entries to reflect the reality of slavery and African Americans in Maryland [10]. The University of Limerick's "Burying Data" project transformed text content from census reports into fine-grained data to explore Irish

history from 1864-1922 and used machine learning algorithms to depict potential social structure patterns [38]. In sentiment analysis, the University of California used a three-step social media similarity mapping method to automatically identify and analyze archived Twitter records, calculating sentiment similarity with test collections to screen various emotional trends during the COVID-19 pandemic prevention and control period [39], such as using machine learning and data analysis to reveal and confirm emotional trends in the COVID-19 Hate Speech Twitter Archive (CHSTA) to provide data for crisis response or public policy formulation [40]. In addition, facing ethical challenges accompanying technology application, computational archival science emphasizes solving issues related to data security and personal privacy. For example, the Netherlands Arts and Humanities Laboratory created Jupyter Notebooks archiving tools to provide metadata archiving and visualization services, combining legal theory to anchor issues of archival information security and personal privacy protection in network environments [41]; the US Library of Congress's National Recording Preservation Board preset issues of political representation balance among different subjects and data ethics when establishing the national radio recordings database from a political science perspective [42].

The archivalization direction of computational archival science mainly discusses the archival processing issues of emerging documents or data (sets), including identification, metadata generation, integrity checking, normalization, blockchain, anonymization, etc. For example, testing the completeness of metadata schemes for European cultural heritage digital platforms through dataset operations [43]; proving that dataset identifier allocation improves dataset availability through testing applications in genomics data management [44]; summarizing preservation strategies such as mirror systems, digital records, and symbolic codes for maintaining document authenticity and security through blockchain innovation applications in land transaction data, health records, and cryptocurrency data [14]. In addition, T. Miksa et al. proposed designing feasible machine-actionable data management plans through adequate understanding of dynamic data information flows and data models, determining necessary services and infrastructure components to support automation of data management tasks [45]; H. Hamouda et al. combined archival appraisal theory and engineering methods to develop six unique video testing methods, identifying three key components—visual components, audio components, and metadata components—to check internal and external consistency of videos [46].

2.3 Computational Archival Science Development Remains in Preliminary Exploration Stage

Since 2016, scholars from multiple disciplines and countries have engaged in continuous exploration of the theory and practice of computational archival science, and its connotation has gradually become clearer with core areas being gradually identified. However, it must be acknowledged that computational archival

science development remains in a preliminary exploration stage. Through classification and statistics of research methods used in 75 foreign language documents (see Table 1), it was found that current foreign computational archival science research methods show an obvious tendency toward singularity, with non-empirical research literature accounting for as high as 73.4%. Moreover, case study literature and introduction literature account for a large proportion of non-empirical research, indicating that as an emerging disciplinary field, computational archival science has not yet formed a relatively mature research framework and theoretical system, and the academic community's focus in its initial development stage is on introducing basic issues and related practical cases in this field.

The author further conducted classified statistics on the research topics of the 75 foreign language documents (see Table 2), finding that current foreign computational archival science research topics show a large quantitative gap between basic research and applied research, with basic research literature accounting for only 24% while applied research literature accounts for 76%, indicating that exploration in the computational archival science field has strong "applicability" and is committed to solving computational technology application problems in archival practice. Specifically, in applied research, studies on archival processing account for about half, with the remainder focusing on archivalization and archival analysis themes. In basic research, studies on basic theory are more numerous, while educational issues in computational archival science are less involved.

Table 1: Classification Statistics of Research Methods in Foreign Computational Archival Science Literature

Research Method	Documents	Percentage
Non-empirical research	Introduction literature	
	Opinion literature	
	Case literature	
Empirical research	Experimental literature	
	Model literature	

Table 2: Classification Statistics of Research Topics in Foreign Computational Archival Science Literature

Research Topic	Documents	Percentage
Basic theory		
Education research		
Archives processing		
Archives analysis		
Archivalization		

Certainly, if examining computational archival science from a disciplinary perspective, the author believes that although computational archival science has the potential to develop into a discipline, it currently does not yet meet the conditions for becoming a discipline. At present, although there are professional academic conferences to discuss computational archival science issues, and the University of Maryland's College of Information Studies, the University of British Columbia's School of Information, and King's College London's Department of Digital Humanities have sporadically formed some academic teams and cooperation networks relying on relevant research institutions and computational infrastructure, and have made bold attempts in computational archival science-related curriculum construction and talent cultivation, most current computational archival science research results are conference papers, with only a small number of journal articles, lacking specialized academic monographs, and no dedicated academic journals, academic societies, or research institutions have been established. Degree education in computational archival science is still blank. The sustainable development of computational archival science still faces enormous challenges.

3. SWOT Analysis of Computational Archival Science Development in China

Since 2016, under the continuous attention of multiple countries and institutions and the promotion of core research teams, computational archival science has flourished abroad, but it also faces many difficulties in its initial development stage, such as unclear disciplinary connotation and uncertain research scope and boundaries. In China, related practices and academic research in computational archival science are mainly scattered in areas such as archival datafication, archival data governance, smart archives and smart archives construction, blockchain and document management. Some scholars from the School of Information Resource Management at Renmin University of China, the Department of Library, Information and Archives at Shanghai University, and the School of Public Administration at Sichuan University are actively tracking foreign research trends in computational archival science, but domestic research in computational archival science tends to focus on basic theory, with applied research and practical exploration obviously lagging behind. Currently, no relevant computational archival science research centers or laboratories have been established in China, no interdisciplinary cooperative research teams have been formed, and no national-level research projects on computational archival science have been approved. The domestic academic community and industry still lack clear answers to questions such as whether computational archival science can adapt to China's disciplinary construction and academic research environment and how to integrate into practical development. Therefore, the author attempts to explore development strategies for computational archival science in China through SWOT analysis, referencing foreign practical experience in promoting computational archival science development and combining domestic policy background, disciplinary construction, and practical needs to examine

the internal and external environments of computational archival science development in China, as shown in Table 3 .

Table 3: SWOT Matrix for Computational Archival Science Development in China

	Strengths (S)	Weaknesses (W)
Internal Environment	• Two approaches to solving practical problems have been established • Preliminary disciplinary research directions have been formed	• Severe shortage of resource investment in computational archival science • Extremely weak research capacity in computational archival science
External Environment	**Opportunities (O)**	**Threats (T)**
	• Aligns with higher education and talent cultivation trends • Aligns with archival informatization development strategy requirements	• Faces impact from related emerging interdisciplinary disciplines • Significantly affected by data security risks

3.1 Analysis of External Environment for Computational Archival Science Development in China

3.1.1 Opportunity: Aligns with Higher Education and Talent Cultivation Trends In May 2019, 13 ministries and commissions including the Ministry of Education officially launched the "Six Excellence and One Top" Plan 2.0. The construction of new liberal arts, as an important component of this plan, adheres to problem orientation, responds to social needs, emphasizes breaking disciplinary barriers, integrating arts and sciences, and explores new majors, new directions, and new models [47]. This not only reflects new trends in current Chinese disciplinary development and higher education talent cultivation but also calls for new research methods and tools in the digital twin world [48]. Under new liberal arts construction, there is a call for deep integration of disciplinary problems with digital technology to enhance the wisdom level of humanities and social science data resources. Computational archival science is precisely data-driven archival practice and research problem-oriented, responding to social needs regarding archives and history, culture, society, and science in the digital intelligence era, and conforming to the trends of new liberal arts and liberal arts laboratory construction. This trend will promote the development and construction of corresponding platforms and tools, such as general

liberal arts experimental platforms for data collection, processing, long-term preservation, and visualization, and targeted system tools for semantic understanding and fine-grained knowledge extraction, which will lay a solid foundation for computational archival science development. In addition, the research and development of integrated experimental platforms will provide experience for applying computational methods and resources to large-scale document or archival processing, analysis, storage, long-term preservation, and access, facilitating the integration of computational thinking and archival thinking to shape a completely new transdisciplinary field.

3.1.2 Opportunity: Aligns with Archival Informatization Development Strategy Requirements The newly revised Archives Law of the People's Republic of China in 2020 added a special chapter on "archival informatization construction," clarifying the overall principles and work priorities of archival informatization construction and highlighting new requirements for archival informatization. During the "14th Five-Year Plan" period, China's archival informatization strategy will further focus on planning and design around three tasks: digital archival resource systems, application systems and utilization systems, and infrastructure and security systems, developing toward datafication, networking, automation, and intelligence. The construction of computational archival science and corresponding computational archival laboratories precisely meets the requirements of archival informatization strategy: on the one hand, computational archival science is committed to improving data processing efficiency, productivity, and accuracy, achieving structured processing and data association of archival data through computational methods, which will provide strong support for archival data resource development, governance, sharing, and application, and boost the construction of smart archival application platforms; on the other hand, computational archival science is committed to solving the archival governance, long-term preservation, and maintenance of emerging digital documents and (big) data resources in government, enterprise, scientific research, and cyberspace, providing methodological guidance and systematic, automated application solutions for digital document single-track operation and single-set preservation and data continuity assurance.

3.1.3 Threat: Faces Impact from Related Emerging Interdisciplinary Disciplines In recent years, data science and digital humanities have developed rapidly in China, with research institutions springing up like mushrooms. According to Ministry of Education statistics, as of June 30, 2021, 12 universities including Peking University, Tsinghua University, University of Science and Technology of China, and Wuhan University have established data science as an interdisciplinary discipline, involving first-level disciplines including computer science and technology, software engineering, management science and engineering, library, information and archival management, mathematics, statistics, and information and communication engineering. At the same time, universities such as Fudan University, East China Normal University, Yunnan University, and

Renmin University of China have also added data science-related second-level disciplines under first-level disciplines such as computer science and technology, software engineering, management science and engineering, and statistics. Renmin University of China has added digital humanities as a second-level discipline under the first-level discipline of library, information and archival management, integrating faculty from the School of Information Resource Management, School of History, School of International Studies, School of Arts, School of Law, and School of Environment to explore innovative paths for interdisciplinary cultivation of new liberal arts talents in digital humanities. Compared with the popularity of data science and digital humanities discipline construction, computational archival science has not received much attention. Coupled with unclear boundaries with related disciplines, computational archival science, which has just emerged in terms of disciplinary construction foundation, resources, and direction aggregation, is easily overlooked and impacted.

3.1.4 Threat: Significantly Affected by Data Security Risks Currently, the state attaches great importance to archival data security issues, but the archival data security regulation system and top-level design are still not sound, and archival data faces technical risks such as hacker attacks, leading to hidden dangers in archival data security [49]. Archival institutions' concerns about archival data security restrict the large-scale opening and development of archival data, limiting access channels and processing effectiveness of archival data resources, resulting in insufficient data resources and lack of research prerequisites for computational archival science research. Currently, the forms of archival data opening in China are limited, mostly in the form of catalog data. Large-scale opening of archival content data requires comprehensive advancement from policy, institutional, technical, platform, format, data governance maturity, and readiness assessment levels. Due to archivists' lack of technical knowledge and insufficient reserves of interdisciplinary talents in archival institutions, the processing, analysis, and application development of large-scale historical archival materials all depend on third parties, leading to many security risks for archival data in the outsourcing process. Most archival institutions are afraid to bear such risks, unwilling to let their collection resources leave their custody and control scope, and maintain a very cautious attitude toward third-party involvement. These security risks and archival institutions' concerns about them will to some extent hinder the development of computational archival science, which is dedicated to large-scale document or archival material processing and research.

3.2 Analysis of Internal Environment for Computational Archival Science Development in China

3.2.1 Strength: Two Approaches to Solving Practical Problems Have Been Established Currently, the practical approaches of computational archival science have gradually become clear: on the one hand, it is committed to solving problems encountered in the datafication, networking, automation,

and intelligence transformation process of archival departments. For example, sensitivity review, privacy, and open appraisal issues for large-scale archival materials; value appraisal and preservation decision issues for emerging born-digital and data-state archival materials; and large-scale archival material mining and research issues oriented by historical, social, scientific, and cultural research needs, involving the application of computational methods and resources to large-scale document or archival material processing, analysis, storage, long-term preservation, and access. On the other hand, it is committed to solving problems encountered in the governance process of emerging digital documents and (big) data in various social fields. For example, long-term preservation issues for government data resources, archival preservation issues for scientific big data, document preservation issues in data analysis and governance activity processes, value appraisal and long-term preservation selection criteria issues for data heritage, and big data security governance and personal privacy protection issues [50], ensuring that electronic archives meet requirements of reliable sources, standardized procedures, and compliant elements, and ensuring that data resources meet requirements of continuity, traceability, trustworthiness, reliability, and security, involving the application of archival theories and methods to big data governance and preservation of emerging digital documents in various departments.

3.2.2 Strength: Preliminary Disciplinary Research Directions Have Been Formed The emerging transdisciplinary field of computational archival science has enormous potential, forming some preliminary research directions, with related practices as support. For example, foreign scholars demonstrated through eight cases how different interdisciplinary efforts have addressed changes in archival practice environments, proposing application methods for constructing eight areas of computational archival science [22], initially forming five directions for theoretical research and practical exploration around archival material analysis, development of new form archives, expanded services for archival processing, big data and archival theory and practice, and network infrastructure for archive and collection-based research [22]. Based on domestic practical approaches, computational archival science will integrate scattered research content based on directions such as archival informatization, archival datafication, smart archives and smart archives, archival knowledge discovery and knowledge services, archival data governance, archival data infrastructure construction, archival data opening, archival data preservation, blockchain and document management, artificial intelligence and document management, trusted digital documents, and archivalization and digital curation, forming new unified and holistic understandings.

3.2.3 Weakness: Severe Shortage of Resource Investment in Computational Archival Science Currently, it is the fifth year since the rise of computational archival science abroad. Research institutions led by the University of Maryland's College of Information Studies and its Digital Curation Inno-

vation Center have invested a certain amount of resources in the construction and development of computational archival science, including: network infrastructure and other computational resources, launching the DRAStic program for developing and maintaining large-scale document data management digital repositories [51]; human, financial, and material resources, collaborating with archival institutions such as the US National Archives and Records Administration and the Maryland State Archives, with cooperative archival institutions providing analyzable archival resources, and personnel including the University of Maryland's College of Information Studies and its student teams and archival institution staff conducting cooperative projects funded by the Institute of Museum and Library Services, the US National Science Foundation, and other foundations. In contrast, related explorations in computational archival science in China have just begun. Currently, only a very few institutions such as the Zhejiang Provincial Archives and Qingdao Municipal Archives have made bold attempts in this field. For example, the Zhejiang Provincial Archives proposed building an archival data center and, taking this opportunity, signed a smart archives research cooperation framework agreement with Alibaba Cloud Computing Co., Ltd. to tackle difficult problems such as archival open appraisal and archival data governance. Overall, China still has a severe shortage of investment in funds and computational resources for computational archival science, which is not conducive to its construction and development.

3.2.4 Weakness: Extremely Weak Research Capacity in Computational Archival Science Ideally, computational archival science researchers need to possess both archival thinking and computational thinking. However, currently, most archival science researchers lack computational thinking and skills to proficiently apply relevant computational methods, while computer science researchers lack archival thinking. Currently, foreign computational archival science research capacity mainly comes from university research institutions and archival institutions, forming cross-border cooperation models to achieve complementary resources and technologies. In contrast, domestic computational archival science research capacity mainly comes from university research institutions, lacking interdisciplinary research teams. On the one hand, the archival academic community has not clearly unified the connotation and boundaries of computational archival science, and interdisciplinary exchange and cooperation face great obstacles. Different disciplines attach different importance to computational archival science, and their participation positioning and role motivations are not clear enough. On the other hand, the archival academic community has not proposed a universal and systematic interdisciplinary archival talent cultivation system. Facing the requirement of possessing both archival thinking and computational thinking, China needs to form interdisciplinary teams with a cross-border cooperation concept to jointly conduct computational archival science research and practice, exploring research methods that integrate computational thinking methods and archival thinking methods around the core purpose of applying computational methods and resources to

process and analyze large-scale documents/archives to improve efficiency and accuracy.

4. Strategy Research for Computational Archival Science Development in China

4.1 Strengths-Opportunities (SO) Strategy

First, we must rely on interdisciplinary construction opportunities to build transdisciplinary research platforms. Cultivating computer and archival interdisciplinary talents requires not only the participation of archival science researchers but also strengthened cross-field cooperation with other disciplines such as computer science and information science to form research teams. We should seize the policy advantages of national encouragement for disciplinary integration and new discipline establishment under new liberal arts construction, actively explore methodological innovations in building transdisciplinary fields, and establish cross-college, cross-disciplinary transdisciplinary research platforms. Second, we must target strategic needs in field informatization to form large-scale research directions. The research topics and directions proposed by foreign scholars in computational archival science cover a wide range, such as big data, graphic analysis, and digital humanities network infrastructure construction. For China, we need to examine the informatization strategic needs and practical problems in government informatization, archival informatization, and cultural informatization, align with the theoretical, thinking, and methodological outputs of computational archival science, focus on issues such as archival data development and utilization, data continuity assurance, and archival data curation and governance, and gather and strengthen characteristic computational archival science research directions with domestic features based on practical needs.

4.2 Strengths-Threats (ST) Strategy

First, we must clarify boundaries with related disciplines and highlight the characteristics of computational archival science. As a new member of the "computational +" discipline array, computational archival science is a transdisciplinary field created through the recombination and integration of elements from archival science, information science, computer science, and other disciplines to create new knowledge. Since the disciplinary framework of computational archival science is not yet fully formed and its boundaries with digital humanities, data science, and other related disciplines are still unclear, it is necessary to first clarify the disciplinary system and research framework of computational archival science. On this basis, we can explore disciplinary development scope and fields by grasping disciplinary connotations, find clear differences from related disciplines such as digital humanities and data science, form an independent identity and characteristics distinct from computational social science, computational linguistics, and computational information science, and

highlight the characteristics of computational archival science. Second, we must leverage transdisciplinary research advantages to avoid data security, privacy risks, and technological ethical risks. Although the digital intelligence era has greatly promoted social transformation, technology application has also intensified data security risks and accompanying ethical issues, posing severe challenges to national and field data security assurance. Computational archival science is an organic integration of computational theoretical methods and archival theoretical methods, with inherent advantages in solving data security and privacy issues. Therefore, in the development of computational archival science, we must leverage this transdisciplinary research advantage to provide theoretical, methodological, and technical support for avoiding and preventing data security and privacy risks.

4.3 Weaknesses-Opportunities (WO) Strategy

First, we must seize the opportunity for cultivating interdisciplinary talents and integrate multiple resources to co-construct teaching platforms. Since the 21st century, driven by information technology, the demand for interdisciplinary talents who understand both technology and management has risen sharply, especially in government informatization and archival informatization. However, China's archival science or public management discipline talent cultivation has not effectively solved the problem of interdisciplinary talent cultivation. The emergence of computational archival science brings opportunities to bridge the gap in cultivating interdisciplinary talents with both computational thinking and archival thinking. We should first build interdisciplinary teaching platforms, concentrate multidisciplinary teaching staff, discuss and formulate computational archival science curriculum systems, integrate knowledge from different disciplines into teaching courses [22], comprehensively transform the content of traditional core archival science courses, and develop applied archival technology courses that cultivate computational thinking. At the same time, we should integrate various resources including disciplinary literature resources, technical equipment resources, and intra- and extra-school cooperation resources to establish disciplinary practice platforms, learn from foreign iSchool project-based training models, build industry-university-research innovation bases, and allow students to participate in computational archival science project practice to master relevant knowledge and methods from practice. Second, we must explore operable technical solutions around core problems in practical fields. We should transform practical needs into problem orientation for disciplinary development, and around strategic needs in field informatization and core problems in field data management, explore operable technical solutions relying on computational archival science. For example, exploring solutions for building trusted digital archives around blockchain technology applications, exploring integrated four-character detection tools around electronic archives' four-character detection issues, and exploring solutions for intelligent appraisal of open archives around artificial intelligence technology applications.

4.4 Weaknesses-Threats (WT) Strategy

First, we must strengthen institutional design and technical breakthroughs, and do well in archival security risk assessment and control. Since the 21st century, with rapid economic and social development, the internal and external environments for archival work have become increasingly complex, with traditional and non-traditional risks threatening archival security increasing daily, making ensuring archival security an important content of China's archival undertaking development. Currently, against the background of opening movements in various fields, moving from archival opening to archival data opening requires top-down opening policies and institutional design and bottom-up data opening technical breakthroughs according to the provisions of the newly revised Archives Law, focusing on resolving the contradiction between opening and confidentiality in archival data opening and handling the relationship between data opening and privacy protection. We should strengthen technical breakthroughs in opening and security appraisal, using natural language processing, machine learning, and other technologies as entry points, targeting different types such as text, images, audio, and video, and different forms such as emails, electronic documents, web files, and social media files. In addition, we must do well in archival data security risk assessment, identify potential and possible archival data security risk elements, prevent risky behaviors by archival formation institutions and their staff, and timely eliminate archival security hazards to compensate for or reduce losses. At the same time, while recognizing that computational thinking and technology enhance data processing capabilities, we must pay attention to the risk of the digital divide and strengthen digital inclusion. The advantages of facing large-scale, multi-type documents and data processing can help bridge the gap between different groups in sharing resources while solving different practical needs. Second, we must increase investment and support for basic research to clarify the theoretical, methodological, and technical systems of computational archival science. Since the beginning of the 21st century, China's large-scale archival digitization projects have laid a good foundation for computational archival science. On the one hand, we need to use data methods to 剥离 valuable information from cumbersome and redundant low-value-density data; on the other hand, we need to avoid the negative impacts of technology on archives and understand documents in the new technological environment. Solving these problems requires in-depth cooperation between archivists and computer scientists. However, computational archival science has a short development time and immature disciplinary development, so we should increase investment in its basic research, clarify the basic theories and methodologies of computational archival science, thereby overcoming potential problems in computational archival science practical application.

In the digital intelligence era, the LIS discipline needs new social contributions and urgently needs to make its voice heard in the new era. The arrival of the digital intelligence era and accompanying changes bring new challenges to LIS, and the archival discipline and archival profession must pay close attention to

and actively respond to major social challenges. Faced with practical needs in informatization and big data management, relying on archival science knowledge alone can no longer respond accurately and efficiently. Archival scholars have realized that the archival discipline must cooperate with other disciplines to jointly address new challenges. The development of computational archival science provides a platform for interdisciplinary exchange, cooperation, and integration. Computational archival science is not a one-way output of methods and technologies from computer science to archival science, but a multi-way output between archival science and other disciplines. The archival discipline should seize the opportunity of computational archival science development to export archival disciplinary knowledge, theories, and methods, and expand the influence of the archival discipline. The rise of computational archival science internationally is not accidental. The continuous emergence of new archival forms calls for joint responses from multiple disciplinary fields; the accelerated transformation of traditional archival work also requires the involvement of computational thinking and methods; the large gap in interdisciplinary archival talents 更需要 innovative methods in archival higher education. Currently, through explorations by domestic and foreign scholars, the transdisciplinary connotation of computational archival science has gradually become clearer, and core areas are being gradually identified, but issues such as disciplinary boundaries, research frameworks, technical systems, and practical paths remain unclear and require further discussion. This paper has proposed development strategies for computational archival science in China to a certain extent based on SWOT analysis. However, due to the preliminary development stage of computational archival science, especially since China's computational archival science construction has not been comprehensively launched at the practical level, the conclusions obtained from this qualitative SWOT analysis are macroscopic and subjective, with limitations. Future research could further combine methods such as multidisciplinary expert consultation or in-depth interviews to understand the attitudes and suggestions of the Chinese LIS community and industry regarding the promotion of computational archival science.

References

[1] CASWORKSHOP. IEEE big data 2016: CAS#1 [EB/OL]. [2021-01-20]. https://ai-collaboratory.net/cas/cas-workshops/ieee-big-data-2016-1st-cas-workshop/.

[2] Fu Tianzhen, Zheng Jiangping. The rise, exploration, and enlightenment of computational archival science [J]. Archives Science Bulletin, 2019(4): 28-33.

[3] Zhou Wenhong, Dai Linxu, He Tantao, et al. Analysis and prospect of computational archival science connotation [J]. Archives Science Study, 2021(1): 49-57.

[4] Tao Yufang. Research status and future prospects of computational archival science [J]. Zhejiang Archives, 2021(2): 53-56.

[5] Liu Yuenan, Yang Jianliang, He Siyuan, et al. Computational archival science: A new development of archival discipline [J]. Library and Information Knowledge, 2021, 38(3): 4-13.

[6] Yu Yingxiang, Liu Qian. On the emergence logic of computational archival science [J]. Archives Science Bulletin, 2021(5): 22-31.

[7] Zhao Yue, Zhang Jiaxin. Analysis of the development prospects of computational archival science in China—Based on a survey of computational archival science perception in China's LIS community [J]. Archives Science Bulletin, 2021(5): 32-39.

[8] UMD ISCHOOL. Research centers & labs [EB/OL]. [2021-11-24]. https://www.ischool.umd.edu/research/centers-and-labs.

[9] UMD ISCHOOL. Directory of richard marciano [EB/OL]. [2021-11-24]. https://www.ischool.umd.edu/about/directory/richard-marciano.

[10] Cox R, Shah S, Frederick W, et al. A case study in creating transparency using cultural big data: The legacy of slavery project [C]//2018 IEEE international conference on big data. Piscataway: IEEE, 2018: 2689-2695.

[11] Underwood W, Marciano R, Liabs, et al. Computational curation of digitized records series of WWII Japanese-American internment [C]//2017 IEEE international conference on big data. Piscataway: IEEE, 2017: 2309-2313.

[12] Marciano R, Lee M, Underwood W, et al. Digital curation of a WWII Japanese-American internment collection: Implications for sociotechnical archival systems [C]//2018 IEEE international conference on big data. Piscataway: IEEE, 2018: 1-4.

[13] Batista D, Weingaertner T. ArchContract: Using smart contracts for disposition [C]//2019 IEEE international conference on big data. Piscataway: IEEE, 2019: 3060-3065.

[14] Lemieux V. A typology of blockchain record keeping solutions and some reflections on their implications for the future of archival preservation [EB/OL]. [2021-09-30]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Lemieux.pdf.

[15] Payne N. Auto-categorization & future access to digital archives [EB/OL]. [2021-04-13]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Payne.pdf.

[16] Payne N. An intelligent class: The development of a novel contextual information capturing framework for the functional classification of records [EB/OL]. [2021-04-13]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2019/11/Payne.pdf.

[17] King's College London. About the department of digital humanities [EB/OL]. [2021-11-24]. https://www.kcl.ac.uk/ddh/about/about.

[18] Bryant M. GraphQL for archival metadata: An overview of the EHRI GraphQL API [EB/OL]. [2021-04-13]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/8.Bryant.pdf.

[19] SoBigData. European research infrastructure for big data and social mining [EB/OL]. [2021-11-24]. http://www.sobigdata.eu/index.

[20] Noah D. Computational archival science international network to be launched [EB/OL]. [2021-11-24]. https://ischool.umd.edu/news/computational-archival-science-international-network-to-be-launched.

[21] Marciano R, Lemieux V, Hedges M, et al. Archival records and training in the age of big data [M]//Percell J, Sa R. Re-envisioning the MLS: Perspectives on the future of library and information science education. Bingley: Emerald, 2018: 179-199.

[22] Payne N. Stirring the cauldron: Redefining computational archival science (CAS) for the big data domain [C]//2018 IEEE international conference on big data. Piscataway: IEEE, 2018: 2743-2752.

[23] Lee M, Zhang Y, Chen S, et al. Heuristics for assessing computational archival science (CAS) research: The case of the human face of big data project [C]//2017 IEEE international conference on big data. Piscataway: IEEE, 2017: 2262-2270.

[24] Stancic H. Computational archival science [EB/OL]. [2021-09-30]. http://bib.irb.hr/datoteka/994072.Stancic_H._{{{Computational}}*{{Archival}}}*{Science}.pdf.

[25] Thibodeau K. Computational archival practice: Towards a theory for archival engineering [EB/OL]. [2021-04-12]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/3.Thibodeau.pdf.

[26] Stancic H, Rajh A, Jamic M. Impact of ICT on archival practice from the 2000s onwards and the necessary changes of archival science curriculum [C]//Proceedings of the 40th Jubilee international convention on information and communication technology, Electronics and microelectronics MIOP 2017. Bjelanovic: Petar, 2017: 812-817.

[27] Weintrop D, Beheshti E, Horn M, et al. Defining computational thinking for mathematics and science classrooms [J]. Journal of science education and technology, 2015, 25(1): 127-147.

[28] Underwood W, Weintrop D, Kurtz M, et al. Introducing computational thinking into archival science education [C]//2018 IEEE international conference on big data. Piscataway: IEEE, 2018: 2761-2765.

[29] Underwood W, Marciano R. Computational thinking in archival science research and education [C]//2019 IEEE international conference on big data. Piscataway: IEEE, 2019: 3146-3152.

[30] Marciano R, Jansen G, Underwood W. Developing a framework to enable

collaboration in computational archival science education [EB/OL]. [2021-04-13]. https://www2.archivists.org/sites/all/files/Marciano,%20Richard.pdf.

[31] Dugenie P, Freire N, Broeder D. Building new knowledge from distributed scientific corpus: HERBADROP & EUROPEANA: Two concrete case studies for exploring big archival data [C]//IEEE international conference on big data. Piscataway: IEEE, 2017: 2231-2239.

[32] Shallcross M. Appraising digital archives with archivematica [EB/OL]. [2021-04-13]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/9.pdf.

[33] Hutchinson T. Protecting privacy in the archives: Preliminary explorations of topic modeling for born-digital collections [EB/OL]. [2021-04-11]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Hutchinson.pdf.

[34] Hutchinson T. Protecting privacy in the archives: Supervised machine learning and born-digital records [EB/OL]. [2021-04-13]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/5.Hutchinson.pdf.

[35] Bryant M. In-place synchronisation of hierarchical archival descriptions [EB/OL]. [2021-04-13]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/8.Bryant.pdf.

[36] Thomas W R. Petabytes in practice: Working with collections as data at Scale [J]. Data and information management, 2019; 3(1): 18-25.

[37] Blanke T. Identifying epochs in text archives [EB/OL]. [2021-04-13]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2017/06/Blanke.pdf.

[38] O'Shea E, Khan R, Breathnach C, et al. Towards automatic data cleansing and classification of valid historical data an incremental approach based on MDD [C]//IEEE international conference on big data. Piscataway: IEEE, 2020: 1914-1923.

[39] Yin Z, Fan L, Yu H, et al. Using a three-step social media similarity (TSMS) mapping method to analyze controversial speech relating to COVID-19 in twitter collections [C]//IEEE international conference on big data. Piscataway: IEEE, 2020: 1949-1953.

[40] Fan L, Yin Z, Yu H, et al. Using data-driven analytics to enhance archival processing of the COVID-19 hate speech twitter archive (CHSTA) [EB/OL]. [2021-04-13]. https://www.researchgate.net/publication/349071485_{{{Using}}*{{machine}}*{{{learning}}*{

[41] Wigham M. Jupyter notebooks for generous archive interfaces [EB/OL]. [2021-04-13]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/13.Wigham.pdf.

[42] Goodman N E, Matienzo M A, Vancour S, et al. Building the national radio recordings database: A big data approach to documenting audio heritage [C]//2019 IEEE international conference on big data. Piscataway: IEEE, 2019: 3080-3086.

[43] Kiraly P. Measuring completeness as metadata quality metric in europeana [EB/OL]. [2021-09-13]. https://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/7.Kiraly.pdf.

[44] Xu W, Hhuang R, Este M, et al. Content-based comparison of common data model for machine-actionable data management plans [C]//IEEE international conference on big data. Piscataway: IEEE, 2016: 3283-3289.

[45] Miksa T, Cardoso J, Borbina J. Framing the scope of a common data model for machine-actionable data management plans for collection identification [C]//IEEE international conference on big data. Piscataway: IEEE, 2018: 2733-2742.

[46] Hamouda H, Bushey J, Lemieux V, et al. Extending the scope of computational archival science: A case study on leveraging archival and engineering approaches to develop a framework to detect and prevent "fake video" [C]//2019 IEEE international conference on big data. Piscataway: IEEE, 2019: 3087-3097.

[47] Fan Liming. "New liberal arts": Era demands and construction priorities [J]. China University Teaching, 2020(5): 4-8.

[48] Tang Yanjun, Jiang Cuizhen. Cross-border integration: A new path for new liberal arts talent cultivation in the new era [J]. Contemporary Education Science, 2020(2): 71-74.

[49] Jin Bo, Yang Peng. Analysis of archival data security governance strategies in the big data era [J]. Information Science, 2020, 38(9): 30-35.

[50] Zhao Yue, Sun Jingqiong, Duan Xiane. Archivalization: Rational thinking on the involvement of archival science in data resource management [J]. Archives Science Study, 2020(5): 83-91.

[51] CASWORKSHOP. Computational archival science [EB/OL]. [2021-10-21]. https://dcicblog.umd.edu/cas/4-cas-cyberinfrastructure/.

Author Contributions:
Zhao Yue: Conceived research questions and framework, wrote and revised the paper;
Ma Xiaoyue: Wrote and revised the paper;
Zhang Jiaxin: Analyzed literature and wrote the paper.

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