*European Journal of AI,
Computing & Informatics*

Vol. 1 No. 2 2025

PINNACLE ACADEMIC PRESS
JOURNALS, BOOKS & PROCEEDINGS

Article Open Access

Innovative Exploration and Practice of Archives Collection Automation

Zhe Liu [1,*]

[1] Archives and University History Museum, China Agricultural University, Beijing, 100193, China
[*] Correspondence: Zhe Liu, Archives and University History Museum, China Agricultural University, Beijing, 100193, China

Abstract: As a fundamental part of archives management, the efficiency and quality of archives collection directly affect the utilization of archival resources. Currently, China's archives collection work is in a critical period of transition from traditional models to digitalization, facing multiple challenges such as technical bottlenecks, outdated concepts, and institutional obstacles. This paper systematically analyzes the current situation and existing problems of archives collection work from four dimensions: technical dilemmas in collection methods, backwardness in management concepts, innovative explorations in grassroots practice, and prospective development paths. It also proposes practical automation solutions for archives collection, providing references for promoting high-quality development of archival undertakings.

Keywords: archives collection; automation; practical solutions

Received: 15 July 2025
Revised: 21 July 2025
Accepted: 04 August 2025
Published: 13 August 2025

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As the Chinese Archives Association wraps up its “14th Five-Year Plan”, archival work has fully entered a new digital stage. Meanwhile, the single-set system for archives has become an inevitable requirement and a key development trend for project archive work in the context of the digital economy [1]. In this context, the exponential growth of digital archival resources has posed a systemic impact on traditional collection models. Although the exponential growth of heterogeneous data and cross-departmental collaboration demands a shift toward digitized workflows, some archival institutions still adhere to physical archiving models involving manual sorting and piece-by-piece verification, thereby failing to integrate into the digital reconstruction process. This has led to challenges such as high handover costs and a lack of internal motivation for participation. Essentially, the lag in digitalization of the collection is exacerbating the source obstruction in resource development. While technological empowerment has primarily focused on "management, preservation, and utilization," it has often neglected the process reengineering of "collection." Although the State Archives Administration has emphasized the development of basic services such as archives collection during the Chinese Archives Association’s "14th Five-Year Plan" period, the effectiveness of informatization reforms in archives collection remains minimal [2].

Currently, modern information technologies represented by big data, artificial intelligence, and cloud computing are advancing rapidly. The construction of digital government and smart cities is in full swing, and the digital and intelligent development trend in various social fields is evident [3]. In this context, traditional archives collection and passive reception models can no longer meet the needs of the times. There is an urgent need to construct a three-in-one research paradigm of "technological and empowerment-mechanism and innovation-service upgrading". Through automation reform and innovation, we can consolidate the foundation of the modern archival resource system, form theoretical achievements with both forward-looking vision and practical effectiveness, and provide systematic support for the high-quality development of archival undertakings.

2. Current Development Status of Archives Collection Work

Archives collection is essentially a systematic process that spans the entire life cycle of archival materials. Its responsible entities should not be limited to archives management departments but should also deeply connect with archives-generating departments, covering the complete process from the generation of archival documents, standardized sorting to transfer to the archives.

Currently, collection work in China's archives still generally suffers from passive limitations. They merely passively wait for archives-generating departments to transfer archives, lacking constructive guidance throughout the entire process of archives generation and transfer. This causes archivists to lack effective management of archive details. At the same time, they overly rely on campaign-style "full collection" propaganda during events like "International Archives Day" and annual archivist training meetings, or simply pressure transferring units through the "Archives Law" and management regulations, failing to establish a collaborative governance mechanism. This model leads to two major contradictions: 1) Due to the lack of active archiving awareness and standardized guidance, archives-generating departments tend to have problems such as incomplete document collection and non-standard sorting; 2) Passive reception makes it difficult to cover archive types that require active collection, such as major events, people's livelihood, and local characteristics.

The most prominent problems currently facing China's archives collection work are concentrated in two aspects:

2.1. The Volume of Archives Collection Work Continues to Grow, and the Pressure of Archives Transfer Is Increasing

Current archives collection work presents a prominent contradiction between the popularization of digital production methods and the lag of traditional collection models. With the in-depth advancement of office automation and e-government, electronic documents generated daily by government departments and enterprises and institutions at all levels have exceeded 90% in proportion. However, The progress of archives collection has failed to keep pace with digital transformation, resulting in a serious "digital disconnect". The core crux of this phenomenon lies in the "technical bottleneck of cross-system data connectivity". Most archiving departments' systems lack effective connection with archives management systems, and data formats and interface standards are not unified, making it impossible to achieve automatic circulation of electronic documents. According to a survey conducted by a municipal archival institution, more than 75% of directly affiliated units still need to print electronic documents into paper format before transferring them to archives departments, which are then scanned and digitized again by the archives departments, creating a concerning phenomenon of "digital atavism". This process not only results in significant waste of human and material resources but also leads to the loss of essential electronic document metadata, which in turn compromises the traceability, authenticity, and long-term preservation value of archives. It is estimated that this inefficient process requires archiving department staff to invest an additional 4-6 hours per week in handling archival affairs, directly triggering widespread resistance.

2.2. Archives Management Departments Adhere to Traditional Mindsets in Collection Concepts, Resulting in Insufficient Innovative Momentum

The predicament of archives collection work stems not only from technical constraints but also reflects deep-seated backwardness in management concepts. Most archives management departments still adhere to the role positioning of "passive recipients" and lack a proactive service awareness, leading to a serious disconnect between collection work and the front end of business. This outdated concept has become an invisible obstacle to the transformation and modernization of archival work, as evidenced by the following three aspects:

2.2.1. Single and Inefficient Publicity Methods

The limitations of annual campaign-style publicity are very obvious. Surveys show that more than 82% of city and county archives mainly carry out legal publicity through nodes such as the annual "International Archives Day", mostly in traditional forms such as distributing manuals and holding exhibitions, lacking continuity and targeting.

2.2.2. Insufficient Understanding of the Front End of Business

Archives department staff rarely conduct in-depth frontline research on document generation and circulation processes, resulting in collection standards that are disconnected from actual business. At the same time, due to the lack of on-the-spot research, archives departments find it difficult to truly understand the difficulties and pressures faced by archiving departments in the process of collecting archives and cannot provide practical and effective help.

2.2.3. Obvious Lack of Service Awareness

Most archives departments focus their main energy on organizing and preserving archived archives, with seriously insufficient guidance and support for the archives generation stage [4]. Data shows that only 31% of archives departments have established dedicated archiving consultation positions, with only 15% offering on-site guidance services. Archiving departments often have nowhere to turn when encountering problems and can only handle them based on experience, directly leading to uneven archiving quality.

3. Challenges Facing Reform and Innovation in Archives Collection Methods

3.1. Challenges in Transforming Traditional Work Concepts

The reform and innovation of archives collection work first face deep-seated obstacles of lagging concepts. For a long time, most archives management departments have adhered to the role positioning of "passive recipients" and lacked proactive service awareness, leading to a serious disconnect between collection work and the front end of business. This outdated work concept continues to hinder the transformation and upgrading of archival work and is reflected in the following key challenges:

3.1.1. Insufficient Service Awareness

Archives departments often focus their main energy on backend organization and preservation, with insufficient guidance and support for the archiving generation stage. Data shows that only 31% of archives have established special archiving consultation positions, and the proportion providing on-site guidance services is even lower at 15%.

3.1.2. Lack of Research Mechanisms

Archivists rarely conduct in-depth research on document generation and circulation processes at the business frontline, resulting in collection standards that are disconnected from actual business. For example, a case from a municipal Housing and Urban-Rural Development Bureau shows that the engineering archives format required by the archives is inconsistent with the export format of the actual approval system, forcing staff to manually reformat and convert data, increasing workload and reducing efficiency.

3.1.3. Inefficient Publicity Methods

82% of city and county archives mainly carry out campaign-style legal publicity through nodes such as "International Archives Day", mostly in traditional forms such as distributing manuals and holding exhibitions, lacking both continuity and targeting.

3.2. Challenges of Lacking Low-Cost Solutions

Another significant challenge lies in the lack of affordable and applicable automation solutions for archives collection. This shortage is primarily reflected in the following aspects:

3.2.1. High Technology Docking Costs

Currently, most archives system solutions require archiving departments to call their archiving interfaces, which requires changing the business system architecture, affecting system stability and increasing transformation costs. Data from a provincial archives indicates that while introducing an AI-driven archiving system can enhance efficiency by up to 76%, the initial technical investment remains relatively high, posing challenges for implementation at the grassroots level.

3.2.2. Poor Standard Compatibility

Archives management systems in different regions and departments operate independently, with inconsistent data formats and interface standards, making it difficult to directly replicate advanced experiences. A survey by a provincial archives bureau shows that only 23% of departments have established standardized electronic document archiving processes.

3.2.3. Imbalanced Academic Research

Innovative research on archives collection is obviously insufficient, with most academic achievements concentrated in the archives management and utilization, forming a distorted research pattern of "emphasizing management over collection", leaving grassroots units lacking theoretical guidance and case references.

3.3. Challenges of Shortage of Compound Talents

The third major challenge facing the reform of archives collection work is the structural shortage of talents, specifically manifested as:

3.3.1. Significant Skill Gaps

Training data from a provincial archives bureau in 2018 showed that only 30% of archives management personnel possess basic informatization operation capabilities. This skill gap is particularly prominent against the background of the implementation of the single-set system for electronic documents. Archives education in colleges and universities focuses on traditional theories and involves little in cutting-edge fields such as information technology and data science, resulting in graduates whose skill sets do not align with the practical requirements of modern archival positions.

3.3.2. Imbalance between Talent Supply and Demand

Professional talents with information technology backgrounds are unwilling to join archives departments due to low salaries and limited development space. The recruitment ratio for informatization positions in a provincial archives reached 1:12, much higher than the 1:3 ratio for traditional positions, and less than 10% of the recruited personnel have dual backgrounds.

3.3.3. Inefficient Training Mechanisms

Existing training programs primarily adopt a one-size-fits-all approach, lacking customization and practical orientation. 85% of archivists believe that they are of limited help in solving actual technical problems, and 70% of people prefer problem-oriented practical training.

4. Low-Cost Technological Innovation Solutions for Automating Collection and Automatic Review Work

As a compound talent with both informatization background and archival professional knowledge, I deeply recognize three major pain points in traditional archives collection models: low efficiency, high error rates, and high costs of implementing automatic collection solutions. Faced with the rapid growth of massive digital archives, the traditional manual operation method of archives collection has become unsustainable, bringing a heavy burden to staff in archiving units and archives. However, current mainstream market solutions require archiving departments to carry out business system transformations, and the high transformation costs make automatic archiving interfaces difficult to implement.

Based on actual research, I propose a novel solution using message middleware, constructing a three-tier data processing model of "reading–reorganizing–output". Its principle is illustrated through a three-tiered framework (Figure 1), and its core advantages are summarized as follows:

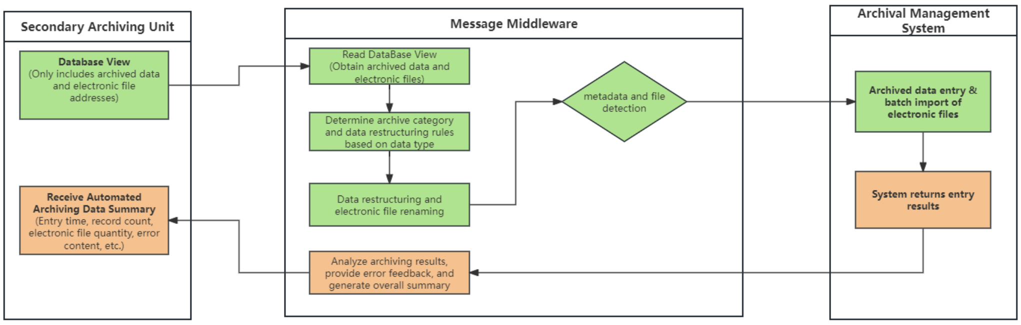
Secondary Archiving Unit
Database View (Only includes archived data and electronic file addresses)
Receive Automated Archiving Data Summary (Entry time, record count, electronic file quantity, error content, etc.)
Message Middleware
Read DataBase View (Obtain archived data and electronic files)
Determine archive category and data restructuring rules based on data type
Data restructuring and electronic file renaming
metadata and file detection
Analyze archiving results, provide error feedback, and generate overall summary
Archival Management System
Archived data entry & batch import of electronic files
System returns entry results

Figure 1. Three-Tier Data Processing Model Based on Message Middleware.

4.1. Data Security Guarantee

The system is strictly limited to reading read-only data views provided by archiving units. This view only contains structured data sets that need to be archived, and implements physical access isolation for other data in the business system to ensure the security of core data in the source system.

4.2. Whole-Process Management of Archival Data

Based on the real-time reading mechanism of business system data views, it realizes a complete traceability chain of archival data from the source of generation to transfer to the archives. Metadata verification rules and transfer progress dashboards are embedded to ensure full-cycle controllability.

4.3. Zero-Transformation Deployment

Data reorganization and conversion are realized through message middleware. Neither the business system nor the archives management system requires structural transformation. Archiving units only need to offer the necessary data views required for archiving. For archives management departments, they only need to import the batch archiving data organized by the middleware into the system. The transformation cost is almost zero.

4.4. Intelligent Review Preposition

A built-in basic archive review module can automatically verify file integrity, format compliance, and metadata completeness. The system increases the problem detection rate by 80% and provides real-time feedback to archiving units, significantly reducing later rework.

4.5. Significant Cost-Effectiveness

Compared with traditional solutions, the implementation cost is reduced by more than 90%, and no professional operation and maintenance team is needed, making it particularly suitable for small and medium-sized institutions with limited resources.

5. Application Practice

Taking the actual case of transferring undergraduate student status information and electronic transcripts from a university to the archives, the university's undergraduate school transfers more than 8,000 paper transcripts to the archives every year. Due to the lack of an automatic data transmission channel between the two systems, the archives can only scan and digitize the received paper transcripts to achieve digital management of transcripts. Student status information needs to be manually entered by part-time archivists in the undergraduate school. To link electronic transcripts with student status data, it is also necessary to perform image recognition on electronic transcripts to extract student IDs, and then match them with student status information entered into the system. The entire workflow is cumbersome, time-consuming, and resource-intensive.

By deploying message middleware, it is possible to automatically read the database view provided by the undergraduate school that contains student status information and electronically authenticated electronic transcript addresses, to obtain graduates' student status information and automatically capture electronic transcripts according to the electronic transcript addresses. After the data is obtained on the archives server, it automatically completes tasks such as metadata reorganization, review, electronic transcript availability verification, and linking of student status information with electronic transcripts, achieving automatic archiving of graduates' student status information and electronic transcripts, significantly saving manpower and scanning expenses. Against the background of nationwide funding constraints, this successful solution for automatic archiving of student status information and electronic transcripts has important reference significance for other universities to realize automatic collection, transfer, and review of undergraduate graduation student status information and archives.

6. Summary

In the context where data transmission and sharing between archives-generating departments and management departments still face multiple barriers, especially for relatively weak archiving links in the collaborative chain, cross-unit data sharing is particularly difficult. The author proposes an archive collection and transfer solution based on message middleware technology, and successfully interlinks the data chain between generating departments and archives. At the same time, it has low cost, solves the systemic impact of the exponential growth of digital archival resources on traditional collection models, liberates archives collection and transfer staff from low-level repetitive mechanical work, greatly improves archiving willingness and enthusiasm, and significantly reduces implementation resistance while avoiding system transformation of archives-generating departments to the greatest extent. It provides a reusable technical path and practical paradigm for the informatization construction of archives collection.

References

1. G. Colavizza, T. Blanke, C. Jeurgens, M. Ridge, M. Esteva, and L. Borbinha et al., "Archives and AI: An overview of current debates and future perspectives," *ACM J. Comput. Cult. Herit.*, vol. 15, no. 1, pp. 1–15, 2021, doi: 10.1145/3479010.
2. L. Jaillant and A. Caputo, "Unlocking digital archives: Cross-disciplinary perspectives on AI and born-digital data," *AI Soc.*, vol. 37, no. 3, pp. 823–835, 2022, doi: 10.1007/s00146-021-01367-x.
3. L. Jaillant, "How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions," *Arch. Sci.*, vol. 22, no. 3, pp. 417–436, 2022, doi: 10.1007/s10502-022-09390-7.
4. A. Hawkins, "Archives, linked data and the digital humanities: Increasing access to digitised and born-digital archives via the semantic web," *Arch. Sci.*, vol. 22, no. 3, pp. 319–344, 2022, doi: 10.1007/s10502-021-09381-0.

Disclaimer/Publisher's Note: The views, opinions, and data expressed in all publications are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of PAP and/or the editor(s). PAP and/or the editor(s) disclaim any responsibility for any injury to individuals or damage to property arising from the ideas, methods, instructions, or products mentioned in the content.