Pathways and Practices for AI-Powered Digital Archives and Intelligent Services

Bing Geng[1], Weiyan Tan[1*], Juan Sun[2]

[1] Guangdong Provincial Veterans Service Center, Guangzhou, 510000, China

[2] Guangzhou Dongsheng Hospital, Guangzhou, 510000, China

* 36297294@qq.com

https://doi.org/10.70695/IAAI202504A7

Abstract

Given the practical challenges of digital archives, such as resource heterogeneity, insufficient semantic utilization, and limited service capabilities, this research constructs an integrated AI-enabled digital archive system encompassing a "data layer—model layer—service layer—operation and maintenance layer". This system unifies the modeling and processing of archival metadata, full-text content, and user behavior, building a knowledge graph and semantic vector framework. It develops semantic retrieval and multi-dimensional navigation algorithms combining deep text encoding, KG embedding, and learning-based ranking, and then utilizes retrieval enhancement to generate intelligent question-answering and personalized recommendation models. Starting from the system level, it constructs a closed loop of intelligent service and manual review through service orchestration and human-machine collaboration. Offline evaluations and simulated online trials conducted on pilot platforms demonstrate that the proposed solution outperforms traditional methods in terms of retrieval accuracy, question-answering quality, recommendation click-through rate, and processing efficiency, while ensuring that tail latency and the proportion of manual intervention remain within a controllable range. This system provides a scalable technical approach and practical reference example for the intelligent upgrading of digital archives and information services.

Keywords Digital Archives; Artificial Intelligence; Semantic Retrieval; Knowledge Graph; Retrieval Enhancement Generation; Personalized Recommendation; Human-computer Collaboration

1 Introduction

In recent years, digital archives and digital library systems have become the main development path in the field of information management amidst the surge of information, the differentiation of user needs, and the iteration of service scenarios. Trehan mentioned that AI-driven archive systems are revolutionizing information acquisition methods, but there are shortcomings in complex text processing and cross-media resource integration [1]. Meesad and Mingkhwan claimed that smart libraries should build a technical architecture oriented towards semantic understanding, but it still faces difficulties in model scalability and reliance on high-quality data [2]. Schellnack -Kelly and Modiba 's research revealed that after the introduction of AI, the description and indexing efficiency of audiovisual archive management in Africa was significantly optimized, but it was limited by incomplete data and basic support conditions [3]. Nduna pointed out, in combination with the evolution of automation and archive discourse, that AI technology urgently needs to be standardized and ethically transparent [4]. Adewojo et al. claimed that AI can improve the efficiency of knowledge services, but there are still differences in the consistency of user experience [5]. Arias Herná ndez and Rockembach reiterated that creating trustworthy AI requires incorporating AI literacy, interpretability, and risk governance into the planning of archive systems [6]. Kannaujia et al. ' s research confirmed that AI has key value for information organization and retrieval, but algorithm bias and model maintenance are still tricky issues [7]. Li et al. conducted research on digital twins of smart cities, created a big data analysis framework based on deep learning, integrated multi-source sensing data of the Internet of Things with the urban virtual-real mapping model, and confirmed the effectiveness of deep feature mining in optimizing urban operation status prediction and resource scheduling [8]. Tripathi et al. 's survey revealed the promoting effect of digital resources on education and scientific research, and reminded that intelligent services need to enhance accuracy and adaptability [9]. Okunlaya et al. finally showed that the AI-driven service framework has a supporting role in the digitalization of higher education, but cross-system collaboration and scenario-based applications need to be further deepened [10].

Existing research largely affirms the value of AI in archival and information services; however, challenges remain, including data heterogeneity, poor semantic understanding, low credibility and interpretability, and low system integration. To overcome these difficulties, an AI-enabled digital archival system capable of semantic modeling and intelligent services can be constructed, oriented towards real business processes . This paper aims to establish a technical framework involving data modeling, semantic indexing, knowledge graph reasoning, and intelligent services; evaluate its effectiveness in retrieval, question answering, and recommendation scenarios; and provide feasible engineering practices. The paper is structured as follows: Chapter 2 explores the characteristics and demands of digital archival data; Chapter 3 plans the AI technology system and key models; Chapter 4 creates intelligent service scenarios and implements performance verification; and Chapter 5 summarizes practical experience and looks ahead to future development.

2 Data Modeling and Intelligentization Requirements Analysis for Digital Archives

2.1 Digital Archives Business Processes and Data Asset Characteristics

Digital archives services can be summarized as a closed loop of "acquisition–cataloging–storage–retrieval–utilization". As shown in Figure 1, the acquisition process aggregates multiple sources of objects, including digitized paper documents, original electronic files, and audiovisual materials; the cataloging stage generates metadata such as titles, authors, dates, and keywords according to standard guidelines; the storage stage inputs digital objects and metadata into a hierarchical storage and backup system; the retrieval stage uses field retrieval, full-text retrieval, and semantic retrieval to find archives; and the utilization period covers online viewing, downloading, access , and cross-system service calls.

From the perspective of data assets, archival resources contain structured, semi-structured, and unstructured data. The cataloging library and business database centrally store structured data; semi-structured data mostly exists in XML/JSON log and extended metadata formats; unstructured data consists of scanned images, PDFs, audio and video, and rich media content. Archival entries, metadata records, full-text content, and relationships between entries should be uniformly modeled at the "data layer" to meet the unified calling requirements of the subsequent model layer and application layer.

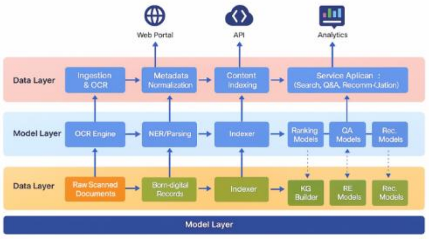

Fig. 1. Digital archive data lifecycle and processing pipeline

2.2 Archival Metadata and Knowledge Graph Modeling

To present archival resources and their semantic relationships in a systematic way, digital archives can be arranged into a heterogeneous knowledge graph. The entity set involves categories such as archival entries, institutions, people, events, place names, and keywords, and contains relational forms such as "creation", "preservation", "citation", "subject association", and "time association". This transforms traditional catalog data into a computationally computable and derivable graph structure.

The graph structure can be formalized as:

$$G = (V, E) \tag{1}$$

This G includes a comprehensive digital archive knowledge graph, V encompassing $E \subseteq V \times R \times V$ a set of entity nodes representing a set of directed or undirected edges between entities;

and defining a set of relation types R to distinguish different semantic relation categories. To facilitate downstream retrieval and reasoning, a numerical feature vector needs to be constructed for each entity.

$$\mathbf{x}_v = f_{\text{feat}}\left(m_v, c_v\right), \quad v \in V \tag{2}$$

Among them, $\mathbf{x}_v \in ò^{d}$ is the set of v dimensional d feature vectors of entities, m_v representing the structured features encoded by metadata fields such as title, author, date, and classification number . c_v It symbolizes the semantic features derived from the full text content, abstract, or multimodal content , $f_{feat}(\cdot)$ and serves as a feature fusion method. It can achieve the goal by relying on deep encoders or multilayer perceptrons. The differences between different collections in terms of document size, number of entities and relations, and completeness of record fields can be quantified using Table 1, providing evidence for subsequent model selection and parameter settings.

Table 1. Corpus and metadata statistics of the digital archive

Collection	#Documents	#Entities	#Relations	Avg. Fields/Record	Time Span (Year)	Missing Rate (%)
Administrative Records	1,250,000	380,000	2,150,000	22	1985–2024	4.3
Audio-visual Archives	210,000	145,000	860,000	18	1990–2024	7.8
Historical Manuscripts	95,000	120,000	540,000	26	1910–1989	12.5
Born-digital Records	630,000	260,000	1,320,000	20	2005–2024	3.1
Aggregated Total	2,185,000	905,000	4,870,000	21.4	1910–2024	6.4

2.3 User Behavior Data and Intelligent Service Requirements Modeling

Intelligent service design leverages a detailed profile of user behavior. User interactions with the archive system primarily involve operations such as retrieval, result clicking, full-text browsing, downloading and saving, and online consultation. The system combines sequential requests into sessions, then associates them with user type, access path, and task objectives. By performing structured modeling on access logs, features can be extracted from the user layer, session layer, and request layer, providing data support for retrieval re- ranking, recommendation, and question-and-answer quality assessment.

Let the user set U be and D be the identifier of the document set . For any user $u \in U$ and document $d \in D$, the overall interaction strength is defined $r_{u,d}$ as:

$$r_{u,d} = \alpha c_{u,d} + \beta d_{u,d} + \gamma t_{u,d} \tag{3}$$

Among them, $c_{u,d}$ the number of clicks or click indications $d_{u,d}$ represent the number of downloads or download guidance, $t_{u,d}$ and are normalized values of document page dwell time ; $\alpha, \beta, \gamma \geq 0$ the weight is used to adjust the effect of different behavioral signals on the intensity of interest. Using the session set as S statistical material, task completion rate indicators can be set.

$$CR = \frac{\sum_{s \in S} z_s}{|S|} \tag{4}$$

The probability of task completion is represented $|S|$ by CR the number of sessions, $z_s \in \{0,1\}$ and the indicator for session completion is s. When a user achieves their goal of finding the required file, completing the download, or closing the consultation ticket, based on indicators such as interaction intensity and task completion rate, the different needs and pain points of different user types in the search, browsing, and consultation stages can be identified. Relevant statistics can be displayed in Table 2, involving User Type, #Sessions, Av. Queries/Session, Click-through Rate (%), Download Rate (%), Av. Dwell Time (s), etc., providing quantitative support for the training and evaluation of intelligent retrieval and recommendation models .

Table 2. User interaction log summary on the archive platform

User Type	#Sessions	Avg. Queries/Session	Click-through Rate (%)	Download Rate (%)	Avg. Dwell Time (s)
Staff	38,500	2.3	68.4	41.7	185
Researchers	52,300	3.1	74.9	55.3	242
Students	79,200	2.7	62.5	29.4	161
Public	44,800	1.9	48.2	18.6	123
Overall	214,800	2.5	63.7	34.2	178

3 System Architecture and Key Models for AI-Powered Intelligent Digital Archives

3.1 Overall Technical Architecture Design for AI-Powered Digital Archives

Focusing on large-scale digital archive scenarios, the overall technical architecture consists of a data layer, a model layer, a service layer, and an operation and maintenance monitoring component. The data layer performs collection, purification, and annotation operations to acquire scanned images, native electronic archives, and log data. It uses an ETL pipeline to perform format standardization, field standardization, and desensitization. Then, it imports structured metadata and unstructured content into a unified storage and indexing cluster.

The model layer builds upon this foundation to construct recognition, representation, and reasoning models, including an OCR engine, text and multimodal encoders, knowledge graph embedding models, and retrieval and recommendation models. Each model exposes gRPC/REST interfaces in a containerized service model, relying on message queues or feature services to achieve feature caching and online inference scheduling. The service layer aggregates business APIs such as semantic retrieval, intelligent question answering, and personalized recommendation , providing access to user portals, external systems, and management consoles through a unified gateway. The operation and monitoring layer covers the entire chain, collecting logs, metrics, and alarms to achieve model version coordination, A/B testing, and automatic resource scaling. The overall data flow and call relationships are presented in Figure 2.

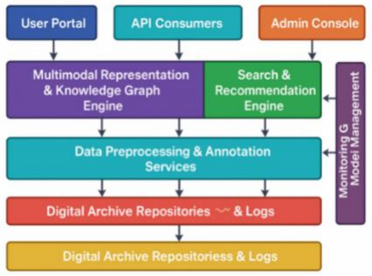
User Portal
API Consumers
Admin Console
Multimodal Representation & Knowledge Graph Engine
Search & Recommendation Engine
Monitoring G Modei Management
Data Preprocessing & Annotation Services
Digital Archive Repositories ~ & Logs
Digital Archive Repositoriess & Logs

Fig. 2. Overall architecture of the AI-enabled digital archive and intelligent services

3.2 Intelligent Indexing and Semantic Representation Learning Model

Automatic document indexing uses a deep text encoder to jointly model OCR outputs and native electronic text. First, we concatenate the title, body text, and structured metadata fields (author, date, classification number) of each document into one token sequence after OCR normalization. Add segment embeddings and positional embeddings to show the boundary of the field and the position of the token. The corresponding OCR text fragments are aligned with these segments, and a Transformer/BERT-like encoder is then used to get a dense semantic representation for the entire document.

Let $D = \{d_1, d_2, \ldots, d_N\}$ represent the set of training documents. For each document $d \in D$, let $\mathbf{x}_d$ be the input token sequence (OCR text and metadata included). A deep encoder $f_\theta(\cdot)$ with parameters θ maps $\mathbf{x}_d$ to a k-dimensional semantic vector:

$$\mathbf{o}_d = f_\theta\left(\mathbf{x}_d\right) \in \text{ò}^{\,k} \tag{5}$$

Here, $\mathbf{o}_d$ is the semantic representation vector of document d, θ represents all the learnable parameters of the encoder, and k is the dimension of the representation space.

In addition to $\mathbf{o}_d$, we construct a multi-label classifier to forecast semantic labels such as topics, entity types, and time/place categories. Assume that there are K different labels. For each label index $k \in \{1,\ldots, K\}$, the model predicts the probability that document d has label k as:

$$p_{d,k} = \sigma\left(\mathbf{w}_k \mathbf{o}_d + b_k\right) \tag{6}$$

Where $\mathbf{w}_k \in \text{ò}^{\,k}$ and $b_k \in \text{ò}$ are the parameters of the classification layer, and $\sigma(\cdot)$ is the sigmoid function. Collect all the label probabilities into $\mathbf{p}_d = \left[p_{d,1},\ldots, p_{d,K}\right]$, which is the predicted multi-label distribution for document d.

Let $y_{d,k} \in \{0,1\}$ be the ground-truth indicator that document d is associated with label k. Training objective is a combination of the standard multi-label cross-entropy loss and L_2 regularization:

$$\mathrm{L} = -\frac{1}{N}\sum_{d=1}^{N}\sum_{k=1}^{K}\left[y_{d,k}\log p_{d,k} + \left(1-y_{d,k}\right)\log\left(1-p_{d,k}\right)\right] + \lambda\|\theta\|_2^2 \tag{7}$$

In (7), N is the number of training documents, K is the number of labels, $y_{d,k}$ and $p_{d,k}$ are the true label indicator and predicted probability for label k on document d, respectively, and $\lambda > 0$ is the regularization coefficient that controls the strength of the L_2 penalty on the encoder parameters θ. The hyper-parameters and training resource settings of the indexing and representation models are listed in Table 3.

Table 3. Hyper-parameters and training resources of indexing and representation models

Model	#Params (M)	Hidden Dim	Max Seq Len	Batch Size	#Epochs	Training Time (h)	GPU Type
BERT-base Encoder	110	768	512	32	8	14.5	V100 × 2
Longformer Encoder	148	768	2048	16	6	19.2	A100 × 2
Multimodal Encoder	120	1024	1024	twenty four	10	22.7	A100 × 4

3.3 Knowledge Graph Embedding and Intelligent Association Discovery

By leveraging metadata and relational schemas, knowledge graph embedding models can be used to create vector representations of archival entities and relationships. This facilitates entity alignment, similar archive discovery, and multi-hop path inference, creating vector representations for each entity and relation type $\mathbf{e}_v, \mathbf{e}_r \in \text{ò}^{\,d}$. Taking TransE as an example , a triple (h,r,t) scoring function can be written:

$$s(h,r,t) = -\|\mathbf{e}_h + \mathbf{e}_r - \mathbf{e}_t\|_2 \tag{8}$$

Here, $\mathbf{e}_h$ and $\mathbf{e}_t$ represent the head and tail entity vectors , respectively, $\|\cdot\|_2$ and are L2 norms. Using positive and negative sample pairs, training can be performed using the margin ranking loss.

$$\mathrm{L}_{KG} = \sum_{(h,r,t)\in\mathrm{T}}\ \sum_{(h',r,t')\in\mathrm{T}^-} \max\left(0, \gamma + s\left(h',r,t'\right) - s(h,r,t)\right) \tag{9}$$

Among them, T the set of real triples T^{-} is the set of negative samples formed by replacing the head entity or the tail entity, which is used $\gamma > 0$ as the margin hyperparameter . After training, graph embedding $\mathbf{e}_v$ and semantic vector $\mathbf{h}_v$ concatenation or attention fusion operations can be implemented to enhance retrieval recall efficiency and recommendation diversity. The loss reduction during the training stage and the fluctuation of indicators such as MRR and Hits@K with epoch can be presented in Figure 3 to evaluate the performance difference before and after the introduction of knowledge graph.

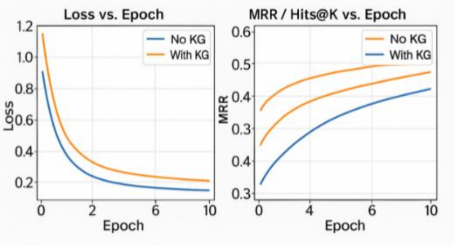

Fig. 3. Training dynamics of semantic indexing and KG embedding models

3.4 Intelligent Retrieval and Recommendation Based on Learning Ranking

Relying on semantic representation and graph embedding, retrieval and recommendation can be uniformly categorized as a learning ranking problem and modeled accordingly. For queries q and candidate documents d , representation vectors are first obtained separately $\mathbf{h}_q, \mathbf{h}_d$, and then feature combination vectors are constructed and relevance values are calculated.

$$s(q,d) = g\left(\mathbf{h}_q, \mathbf{h}_d\right) = \mathbf{w} \quad \left[\mathbf{h}_q; \mathbf{h}_d; \mathbf{h}_q \odot \mathbf{h}_d\right] \qquad (10)$$

Here, $[\cdot;\cdot]$ represents the action of vector concatenation, $\odot$ represents element-wise multiplication, w and serves as the parameter vector q to be learned . For a given query , softmax is used to normalize the scores of the candidate set $\{d_j\}$, thereby constructing a list-style loss for the labeled relevant documents.

$$\mathrm{L}_{\text{rank}} = -\sum_{q} \sum_{d_i \in \mathrm{D}_q^{+}} \log \frac{\exp\left(s\left(q, d_i\right)\right)}{\sum_{d_j \in \mathrm{D}_q} \exp\left(s\left(q, d_j\right)\right)} \mathrm{D}_q^{+} \mathrm{D}_q \qquad (11)$$

Here, D_q the candidate document combination representing query q D_q^{+} is a subset of the relevant documents. By minimizing the Lrank using the training set , the relevance of the search results and the accuracy of the ranking can be improved simultaneously. During the engineering implementation, different configurations are combined into several model variants based on whether or not graph features and user behavior features are used and the real-time calculation method is used. The features used, latency, memory usage and other indicators can be reflected in Table 4, which can serve as the basis for subsequent experiments and system selection.

Table 4. Retrieval and recommendation model variants

Variant	Features Used	KG Enabled?	User Behavior Used?	Latency (ms)	Memory (MB)
Baseline-BM25	Keyword + Field Weights	No	No	12	480
Semantic-only LTR	Text Embeddings	No	Yes (CTR features)	35	920
KG-enhanced LTR	Text Emb. + KG Emb.	Yes	Yes	48	1,150
Session-aware RecSys	Text Emb. + KG Emb. + Session Seq.	Yes	Yes (Seq features)	62	1,380

4 Design and System Implementation of Intelligent Service Scenarios for Digital Archives

4.1 Semantic Retrieval and Multidimensional Navigation Services

Using the ranking model described above, the semantic retrieval subsystem begins by parsing natural language queries, distinguishing between keywords, entity names, and time constraints. It standardizes the expression by relying on synonym expansion and domain terms, and selects candidate archives from the full-text index and knowledge graph neighborhood using a hybrid recall method of vector index and inverted index. Then, the learning ranking model adjusts the ranking by integrating semantic similarity, entity matching degree, and user behavior characteristics. The search results are presented on the front end through multi-dimensional navigation, allowing interactive filtering based on keywords, archive entities, timelines, and collection sources. Under high concurrency, caching and sharded indexes are used to keep the average response time stable at around 100 milliseconds.

Offline evaluation uses manually compiled query-document relevance labeled materials to analyze the differences in accuracy, ranking quality and efficiency of different retrieval strategies. Table 5 presents the performance of each method on metrics such as P@5, P@10, nDCG@10 and MAP. In addition, query latency and index size are given to facilitate the trade-off between performance and resource consumption.

Table 5. Offline evaluation of semantic search and navigation models

Method	P@5	P@10	nDCG@10	MAP	Latency (ms)	Index Size (GB)
BM25	0.612	0.578	0.643	0.421	18	120
BM25 + Field Boost	0.647	0.601	0.671	0.446	22	122
BERT Semantic Only	0.702	0.664	0.718	0.489	39	138
BERT + KG Features	0.734	0.691	0.752	0.517	46	145
BERT + KG + Behavior Features	0.761	0.714	0.781	0.538	53	152
Distilled Online Model	0.749	0.705	0.769	0.529	31	133

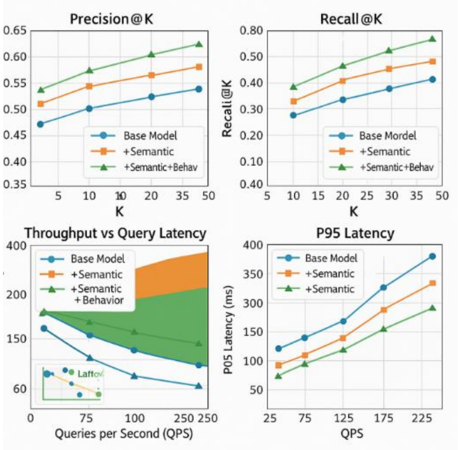

Fig. 4. Semantic search performance and efficiency comparison

Figure 4 also shows the Precision@K / Recall@K curves and the comparison chart of QPS and P95 Latency. The top row shows the benefits of semantic enhancement and the addition of behavioral features at high K values, while the bottom row shows the differences in throughput and tail latency of various solutions under different concurrency pressures, which helps in online solution decision-making.

4.2 Intelligent Question Answering and Document Interpretation Services

The intelligent question-answering module employs a retrieval-enhanced generative framework, combining full-text archival indexing with a large-scale language model. Upon receiving a user's question, the system first retrieves several highly relevant fragments and their metadata from the archive using the same semantic retrieval component as Chapter 3 , and then expands the relevant entities and timeline using a knowledge graph. Subsequently, the RAG pipeline organizes this evidence into structured prompts, injecting citation sources, summary instructions, and style constraints into the generative model input to generate an answer containing natural language explanations, key archive numbers, and time nodes. To control for illusions and compliance risks , sensitive field masking rules are introduced during the answer generation stage, automatically desensitizing or replacing confidential information and personal privacy data, and mandating the inclusion of citation annotations in the text.

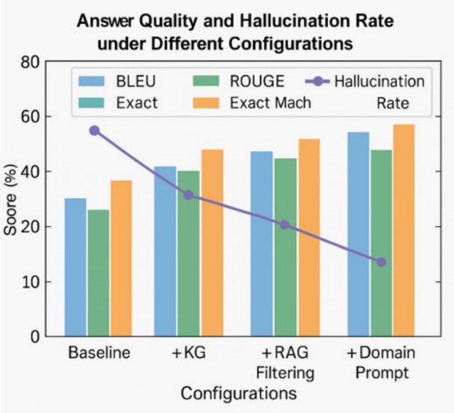

Fig. 5. Answer quality and hallucination rate under different configurations

The question-and-answer quality assessment employs both automated metrics and human evaluation. The evaluation dataset consists of questions from real users and standard answers provided by archival professionals. Text overlap metrics such as BLEU, ROUGE, and Exact Match are calculated for each. Furthermore, experts evaluate the illusion rate and usability. Figure 5 presents a performance comparison of four configurations: Baseline generation, knowledge graph integration, added search filtering, and domain prompts. The bars show the upward trend of BLEU, ROUGE, and Exact Match, while the line shows that the illusion rate decreases significantly with increasing configuration complexity. The comparison results demonstrate that the combined use of RAG Filtering and Domain Prompt significantly suppresses unfounded generation and improves answer traceability, providing quantitative evidence for the subsequent application and promotion in key business scenarios.

4.3 Personalized Profile Recommendation and Knowledge Path Guidance

The personalized recommendation module, based on user behavior logs in Section 2.3 and knowledge graph embedding in Section 3.3, conducts comprehensive modeling of users' long-term interests and conversational intent. The system begins by building a long-term profile for each user based on file entities and keywords, and reflects click, download, and consultation behaviors onto the knowledge graph to create interest subgraphs. It uses RNN or Transformer encoders to construct short-term state vectors from single-session access sequences to reflect the local preferences of the current retrieval task. After entering the candidate generation stage, it uses collaborative filtering, graph walking, and content-based recall to co-create candidate files. Then, it uses a ranking system that integrates long and short-term vectors and file embedding to output recommended items. Finally, it formulates "extended reading routes starting from this file" for users according to the knowledge path.

Offline evaluation utilizes historical logs to divide the training set and the holdout set, and conducts simulated online experiments simultaneously. Click-through rate and coverage are calculated by replaying real conversation sequences. Table 6 summarizes the Recall@20, NDCG@20, CTR, Coverage, and list diversity of various recommendation methods in different usage scenarios (such as research users, student users, public access, etc.). The conversation-aware and knowledge graph-integrated model surpasses traditional collaborative filtering in terms of recall and click-through rate, while significantly improving coverage and diversity, supporting the transition of archival services from "single-point search" to "knowledge path guidance".

Table 6. Offline and simulated online metrics of recommendation

Scenario	Method	Recall@20	NDCG@20	CTR (%)	Coverage (%)	Intra-list Diversity
Researcher Desk	Baseline CF	0.612	0.433	9.8	21.4	0.41
Researcher Desk	KG-Enhanced Seq	0.732	0.521	13.7	34.9	0.56
Student Mobile	Baseline CF	0.547	0.396	7.3	18.2	0.37
Student Mobile	Session-aware Rec	0.689	0.478	11.2	31.7	0.52
Public Portal	Popularity Rank	0.421	0.315	5.6	12.5	0.29
Public Portal	Hybrid Rec (All)	0.603	0.432	8.9	28.4	0.51

4.4 Intelligent Service Process Orchestration and Human-Machine Collaboration

When multiple intelligent capabilities emerge simultaneously, service process orchestration should be used to achieve collaboration between retrieval, question answering, recommendation, and human support tickets. All front-end requests should first be connected to the API gateway. The policy center should guide the requests to the corresponding microservice links according to the request type, user identity, and system load. If a retrieval request enters the semantic retrieval and navigation module, complex questions should be included in the RAG question answering path, and long conversations should drive personalized recommendation operations. Before each link returns results, compliance review and confidence assessment should be carried out. If the confidence level does not reach the threshold or triggers sensitive rules, the system will automatically generate a contextualized support ticket and hand it over to a human archivist, who can add explanations or directly adjust the answer. At the same time, the feedback content will be included in the training data for subsequent model upgrades and updates.

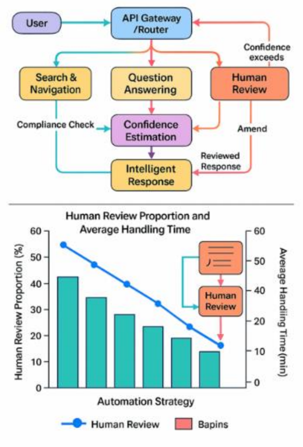

Fig. 6. Intelligent service orchestration and human-in-the-loop workflow analysis

To determine the effect of intelligent orchestration strategies on business efficiency and human workload, the system devised various levels of automation and conducted a comparative experiment in pilot institutions. The experimental results in Figure 6 show that with the introduction of confidence level grading and human-machine collaboration mechanisms, the proportion of human intervention remained stable within a controllable range, the overall problem-solving time decreased sharply, and the complaint rate did not increase significantly. This indicates that reasonable process planning can effectively save the time of archival professionals while maintaining service quality and compliance , allowing them to undertake more complex business tasks.

5 Conclusions and Outlook

This research focuses on the intelligent needs of digital archives. It constructs a technical system integrating data modeling, semantic representation, knowledge reasoning, and intelligent services, which is then validated in real-world business scenarios. Utilizing a unified data and knowledge graph model, and employing deep semantic indexing, KG embedding, and learning ranking for semantic retrieval, it significantly improves retrieval quality metrics such as P@K and nDCG while maintaining controllable response latency . The intelligent question-answering system using RAG, relying on knowledge filtering and domain guidance, significantly reduces illusions. A personalized recommendation model integrating behavioral logs and graph structures effectively balances recall, click-through rate, and diversity. Employing service orchestration and human-machine collaboration mechanisms, a closed-loop system of intelligent results and human review is established, improving overall processing efficiency while adhering to security and compliance principles in archival services. AI-enabled digital archives enhance the retrieval and utilization experience and provide applicable technical methods for optimizing archival business processes and innovating service models.

Acknowledgement

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Trehan, V. (2023). AI-powered archives: Revolutionizing information access for the future. In 2023 IEEE International Conference on Big Data (BigData) (pp. 6298-6300). IEEE.
2. Meesad, P., & Mingkhwan, A. (2024). AI-powered smart digital libraries. In Libraries in transformation: Navigating to AI-powered libraries (pp. 391-428). Cham: Springer Nature Switzerland.
3. Schellnack-Kelly, I., & Modiba, M. (2025). Developing smart archives in society 5.0: Leveraging artificial intelligence for managing audiovisual archives in Africa. Information Development, 41(3), 626-641.
4. Nduna, V.(2025). The new automation: Artificial intelligence and the archives discourse. In Artificial intelligence in records and information management (pp. 273-304). IGI Global Scientific Publishing.
5. Adewojo, A. A., Amzat, O. B., & Abiola, H. S. (2025). AI-powered libraries: Enhancing user experience and efficiency in Nigerian knowledge repositories. Library Hi Tech News, 42(2), 12-16.
6. Arias Hernández, R., & Rockembach, M. (2025). Building trustworthy AI solutions: Integrating artificial intelligence literacy into records management and archival systems. AI & Society, 1-18.
7. Kannaujia, S. K., Verma, P. K., Verma, S. K., et al. (2024). AI-powered revolution: Automating information management in libraries. Academic Libraries, 291.
8. Li, X., Liu, H., Wang, W., et al. (2022). Big data analysis of the internet of things in the digital twins of smart city based on deep learning. Future Generation Computer Systems, 128, 167-177.
9. Tripathi, S., Bhushan, C., & Upreti, N. C. (2025). Empowering education and research: Unveiling the dynamic influence of digital libraries and information centres. In AIP Conference Proceedings (Vol. 3224, No.1, p. 020015). AIP Publishing LLC.
10. Okunlaya, R.O., Syed Abdullah, N., & Alias, R. A. (2022). Artificial intelligence (AI) library services innovative conceptual framework for the digital transformation of university education. Library Hi Tech, 40(6), 1869-1892

Biographies

1. **Bing Geng** Bachelor, he works at Guangdong Provincial Veterans Service Center, have punished 3 publicly academic papers;
2. **Weiyan Tan** PHD, Senior Engineer, Social Worker, she works at Guangdong Provincial Veterans Service Center, . Led the project "Key Technologies and Applications of Big Data Integration and Intelligent Services for Veterans" which was awarded the Third Prize of the Golden Bridge Award by China Technology Market Association;
3. **Juan Sun** Bachelor, he works at Guangzhou Dongsheng Hospital with 2 publicly published academic papers.

AI賦能數字檔案和服務智能化的路徑與實踐

耿兵[1]，譚蔚妍[1]，孫鐫[2]

[1]廣東省退役軍人服務中心，廣州，中國，510900

[2]廣州市東昇醫院，襄陽，中國，441000

摘要：鑑於數字檔案在資源異構、語義利用不足、服務能力有限等方面的實際困境，本項研究打造了「數據層—模型層—服務層—運維層」一體化的AI賦能數字檔案體系，把檔案元數據、全文內容與用戶行爲做統一建模處理，組建知識圖譜與語義向量格局；打造結合深度文本編碼、KG嵌入和學習排序的語義檢索與多維導航算法，而後運用檢索增強生成的智能問答及個性化推薦模型；以系統層面爲出發點，憑藉服務編排與人機協同構建智能服務與人工審覈閉環。試點平臺開展的離線評估以及模擬在線試驗證明，提出的方案在檢索精度、問答質量、推薦點擊率以及處理效率等指標方面勝過傳統方法，保證尾延遲及人工介入比例在可控範圍之內，該體系可爲數字檔案和信息服務實現智能化升級給予可推廣的技術途徑和實踐參考範例。

關鍵詞：數字檔案；人工智能；語義檢索；知識圖譜；檢索增強；個性化推薦；人機協同

1. 耿兵，学士，公开发表学术论文3篇；
2. 谭蔚妍，博士，高级工程师，社会工作师，主持的退役军人大数据整合与智能服务关键技术及应用项目获得中国技术市场协会金桥奖三等奖；
3. 孙镌，学士，公开发表学术论文2篇。