Artificial Intelligence Corpus Library: Connotation, Functional Requirements, and Construction Path

Authors: Liu Xiwen, financial power, Tu Zhifang

Date: 2026-01-14T10:08:54+00:00

Abstract

[Purpose/Significance] Against the backdrop of the accelerated development of AI4S and the national strategy of "Artificial Intelligence + Science and Technology," high-quality, computable corpora have become a core element supporting large-model pre-training and intelligent scientific discovery. As a form of social infrastructure, libraries are now facing a practical imperative to transform their functions toward the construction and service of artificial intelligence corpora. This paper aims to define the connotation and core functions of an "artificial intelligence corpus library," providing a reference for related theoretical research and construction practice. [Methods/Process] Through conceptual clarification and evolutionary analysis, the study identifies the essential characteristics of an artificial intelligence corpus library and interprets it as a deep integration and functional reconstruction of "database × corpus × digital library." From the three dimensions of technological evolution, social application, and regulatory governance, it examines the functional requirements, and, drawing on the "text-as-data" model of HathiTrust in the United States and the "digital scholarship" practices of the British Library, it summarizes the transformation logic from digital libraries to corpus libraries. [Results/Conclusions] The study argues that an artificial intelligence corpus library is a new type of knowledge infrastructure centered on multimodal, computable corpora, supporting the efficient and stable operation of intelligent facilities and constituting one of the effective approaches to realizing artificial intelligence governance. Its construction should follow an architectural system that takes data-driven logic as the top-level design, knowledge organization as the intermediate mechanism, and intelligent-agent applications as the driver of functional innovation. By upgrading existing library collections into corpus-based resources, the corpus library constructs a "non-consumptive use" value-added service model, embeds itself into digital scholarship workflows, and thereby achieves functional expansion and intelligent upgrading.

Full Text

Preamble

Artificial Intelligence Corpus Library: Connotations, Functional Requirements, and Construction Pathways

Liu Xiwen[1],[2], Qian Li[1],[2], Tu Zhifang[1]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190
[2]Department of Information Resource Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract

[**Purpose/Significance**] Under the accelerated development of AI for Science (AI4S) and the national "AI + Science and Technology" strategy, high-quality, computable corpus resources have become core elements supporting large model pretraining and intelligent scientific discovery. As social infrastructure, libraries face practical demands to transform toward AI corpus construction and services. This paper aims to define the connotation and core functions of the "AI corpus library" to provide references for related theoretical research and practical construction.

[**Method/Process**] Through conceptual analysis and evolutionary examination, this paper clarifies the essential characteristics of AI corpus libraries, interpreting them as a deep integration and functional reconstruction of "Database × Corpus × Digital Library." It analyzes functional requirements from three dimensions—technological evolution, social application, and regulatory governance—and synthesizes the "text-as-data" model of HathiTrust and the "digital scholarship" practice of the British Library to summarize the transformation logic from digital libraries to corpus libraries.

[**Result/Conclusion**] The study argues that the AI corpus library is a new type of knowledge infrastructure centered on multimodal, computable corpora, supporting the efficient and stable operation of intelligent facilities and representing an effective pathway for AI governance. Its construction should follow an architecture with data-driven logic at the top level, knowledge organization as the intermediate mechanism, and intelligent agent applications as functional innovation. Through the corpus-based upgrading of existing collections, corpus libraries build a "non-consumptive use" value-added service model embedded in digital scholarship workflows, achieving functional expansion and intelligent upgrading.

Keywords: Artificial Intelligence; AI Corpus Library; High-quality Data; Knowledge Infrastructure; AI Governance

1 Introduction

Artificial intelligence (AI), as the core driving force of a new round of scientific innovation and industrial transformation, is profoundly reshaping scientific research paradigms and knowledge production models. With the rapid development of generative AI, large models, and intelligent algorithms, the deep integration of AI and science and technology has become a crucial strategic direction for countries to enhance original innovation capabilities and seize the commanding heights of future technological competition. Consequently, major scientific and technological powers have systematically deployed national-level development pathways for "AI + Science and Technology," accelerating the process of AI-enabled scientific research and technological innovation through strengthening computing infrastructure, promoting data resource integration, and cultivating interdisciplinary talent systems.

On August 26, 2025, China' s State Council issued the "Opinions on Deepening the Implementation of the 'AI+' Action" [1], explicitly proposing the "AI + Science and Technology" initiative. The document emphasizes using AI to empower key stages of basic research, applied research, and technological innovation to accelerate scientific discovery, listing "strengthening data supply innovation," "continuously enhancing high-quality AI dataset construction," and "improving cross-modal complex scientific data processing capabilities" as important tasks. This highlights the fundamental role of high-quality data resources in the "AI + Science and Technology" strategy. On November 20, 2025, the UK government released the "AI for Science Strategy" [2], aiming to develop frontier capabilities in AI-driven science and consolidate the UK' s global scientific leadership. Focusing on five priority areas—engineering biology, fusion energy, materials science, medicine, and quantum technology—the strategy is built upon three pillars: data, computing, talent, and culture, with a targeted investment of £1.37 billion to accelerate AI-driven scientific breakthroughs. On November 24, 2025, the U.S. government officially launched the AI "Genesis Mission" [3], led by the Department of Energy (DOE) [4], integrating supercomputing resources from 17 national laboratories, massive scientific data accumulated by the federal government, cloud-based AI computing environments, and robotic laboratories to build a unified "American Science and Security Platform" to advance intelligent scientific discovery in key areas such as biotechnology, materials science, nuclear energy, semiconductors, and quantum information. These national strategies demonstrate that AI has evolved from a general-purpose technology to a foundational force leading scientific research paradigm transformation, with high-quality data and corpus resources becoming the key foundation supporting this transformation.

Meanwhile, scientific research is transitioning from traditional paradigms centered on experiments, theory, and computation to a new paradigm characterized by deep integration of data and intelligence. AI for Science (AI4S) or AI for Research (AI4R) [5], as the concentrated embodiment of "AI + Science and Technology," is systematically reconstructing the content, methods, and organi-

zation of scientific research. In key stages such as model construction, knowledge reasoning, scientific discovery, pattern mining, and research evaluation, AI has unprecedentedly high requirements for data scale, quality, semantic annotation, structuralization, cross-modal fusion capabilities, and accessibility [6]. Especially against the backdrop of widespread generative AI applications, the core resources supporting model pretraining and intelligent reasoning have evolved from general "data" to high-quality "computable corpus" with explicit semantic structures, disciplinary knowledge backgrounds, and computable features. Such corpora encompass various content types including papers, standards, patents, technical reports, datasets, and technical documents, as well as multiple knowledge modalities such as text, scientific data, experimental records, videos, images, tables, formulas, and code.

Across academia and industry, significant progress has been made in AI corpus construction. In the general domain, large-scale corpora for text, image, speech, and multimodal data have rapidly expanded, with typical examples including LAION-5B [7], Common Crawl [8], and ImageNet [9]. Meanwhile, key technologies for collection, cleaning, annotation, and semantic modeling have continuously matured, significantly improving the automation and intelligence level of corpus construction. In vertical domains, specialized corpora in medicine, agriculture, materials chemistry, finance, and linguistics have emerged rapidly, such as the multilingual medical corpus MMedC built by Shanghai Jiao Tong University [10], the AGROVOC linked open dataset developed by the Food and Agriculture Organization of the United Nations [11], and the Materials Project open-source database in materials science [12]. These corpora have continuously improved in content quality, structuralization, and knowledge representation capabilities, providing a solid foundation for model training, intelligent analysis, and industry applications.

However, AI corpus construction still faces numerous challenges. First, corpus supply is generally insufficient and of uneven quality, with high-value and multimodal resources lagging in updates. Second, the corpus circulation system is constrained by infrastructure conditions and governance capabilities, resulting in low sharing efficiency and inadequate full-lifecycle security guarantees. Additionally, in application scenarios, corpora suffer from shortcomings in quality credibility, compliance control, and scenario adaptability, leading to widespread problems of being "unusable, unusable with confidence, and not user-friendly"[6]. Meanwhile, corpus open sharing requires balancing intellectual property rights, data security, and ethical requirements [13], while interdisciplinary integration further increases the complexity of standardization, semantic consistency, and continuous updating. As model scales continue to expand and capabilities improve, higher requirements are also placed on the traceability of corpus sources, interpretability of training processes, and credibility of output results [14-15]. Overall, AI corpus construction is a systematic engineering endeavor spanning the entire chain of information resource collection, processing, organization, governance, circulation, and application.

As professional institutions that have long undertaken the functions of information resource collection, management, and development, libraries should further expand their systematic collection, standardized management, and specialized development of AI corpora based on existing digital library construction. This transformation is both an external requirement posed by profound changes in scientific research paradigms and technological environments, and an internal driving force for libraries to adapt to technological evolution and achieve functional expansion and service upgrading. In the AI era, library participation in corpus construction and services reflects the extension of their organizational objects, service targets, and service functions: First, the collection resource system extends beyond traditional literature information resources to include diverse types such as open academic resources, scientific datasets, technical documents, and web-based knowledge resources. Second, library service targets evolve from being "human-centered" to simultaneously serving "human-machine" interactions, with an increasing focus on services oriented toward "machines." Third, traditional professional services such as literature verification, citation indexing, and version preservation expand into areas such as standardized management, security assurance, source tracing, evidence preservation, credible verification, and long-term preservation of AI corpora [16]. In the AI era, new types or forms of libraries will emerge, which can be termed "AI corpus libraries."

Against the overall background of AI4S transformation and the national "AI + Science and Technology" strategy, this paper systematically explores the conceptual connotation, functional requirements, and construction pathways of AI corpus libraries. The paper aims to theoretically clarify the relationship between AI corpus libraries and related concepts such as corpora, data repositories, digital libraries, and smart libraries; to technically analyze the application-driven functional requirements and development-oriented architectural logic of AI corpus libraries; and to practically present representative cases and implementation pathways, thereby providing theoretical and practical references for high-quality scientific corpus construction, improvement of scientific infrastructure systems, and sustainable development of AI governance.

2.1 Development and Evolution of Libraries in the Digital Intelligence Era

As a "growing organism" [17], libraries have always existed in a dynamic unity of "change" and "constancy." On one hand, libraries maintain relatively stable core values and institutional foundations in their long-term development. In terms of development philosophy, they adhere to a user-centered approach, safeguarding information equity and value neutrality. In social functions, they continuously serve as important actors in knowledge access guarantee, social and cultural heritage preservation, information infrastructure construction, and public decision-making support [18-19]. In core functions, they provide stable guarantees for knowledge source tracing, evidence preservation, and long-term credible storage through systematic collection, management, and development

[16]. In methodological systems, they rely on mature literature classification and knowledge organization systems [20-21] to systematically organize, standardize, and order dispersed and disordered information resources. On the other hand, as crucial infrastructure for human knowledge production and dissemination, library forms have always been closely related to information technology development and changes in knowledge production methods. Throughout this evolution, libraries have experienced development stages where physical libraries and digital libraries serve as the main threads, with electronic libraries, network libraries, virtual libraries, mobile libraries, and smart libraries emerging successively or coexisting over the long term [22]. For example, from physical libraries centered on "collection and use" of paper documents [23-24], to digital libraries characterized by networked access to digital resources [25-26], and then to smart libraries integrating AI, big data, and IoT technologies [22, 27]. Library forms exhibit distinct characteristics of their times.

Entering the digital intelligence era, the synergistic effects of data, algorithms, and computing power have profoundly changed knowledge production and utilization methods. Scientific knowledge increasingly presents a new form combining data intensity, model-driven approaches, and intelligent reasoning. The resource forms and service targets that libraries face have also undergone significant changes, posing new requirements for their resource organization logic, management mechanisms, and service models. Libraries are facing the demand for constructing and serving high-quality corpus resources and urgently need to explore new development forms for the intelligent era by building AI corpus libraries to better support intelligent learning, reasoning, and scientific research activities.

2.2 Conceptual Definition of AI Corpus Library

"Corpus" originally derived from linguistics research, typically referring to systematically collected, organized, and reusable collections of authentic language materials for describing, analyzing, and explaining linguistic phenomena. With the development of computational linguistics, machine learning, and AI technologies, its connotation has gradually expanded to refer to various structured or semi-structured data resources available for algorithm training, reasoning, and evaluation. Currently, academia and industry have widely used related concepts such as "corpus," "data corpus," "knowledge corpus," and "corpus datasets" [13, 28], but there remains a lack of clear definition for the emerging form of "AI corpus library." Overall, AI corpus libraries will inherit the demand-oriented service philosophy of digital libraries in terms of value orientation, but their service targets will no longer be limited to human users. Instead, they extend to AI models and intelligent application scenarios themselves, emphasizing the usability, understandability, and controllability of corpus resources in model training, reasoning, and practical application processes.

In terms of functional positioning, AI corpus libraries not only undertake the continuous supply function of high-quality corpora and credible knowledge but

also serve as important components of new knowledge infrastructure, fulfilling fundamental responsibilities such as long-term preservation, rights identification, evidence preservation, and source tracing of corpus resources. They provide verifiable and accountable knowledge bases for AI training and intelligent decision-making from institutional and technical perspectives. In terms of working methods, AI corpus libraries apply more semantic association, knowledge graphs, multimodal alignment, and automatic annotation methods based on traditional knowledge organization and resource management standards. Through resource ordering, association, and evidence-based reconstruction, they achieve credible organization and verifiable reconstruction of corpus resources.

From an overall perspective, AI corpus libraries represent neither a linear superposition of traditional corpora, databases, or libraries nor a simple extension of existing functions. Instead, they constitute a deep integration and functional reconstruction of three types of infrastructure under the background of the digital intelligence era. Their internal relationship can be summarized as "AI Corpus Library = Database × Corpus × Digital Library." Among them, "database" reflects the storage, computing, and governance capabilities for large-scale, multimodal, and heterogeneous resources; "corpus" reflects the resource organization logic oriented toward serving AI training, reasoning, and application; and "digital library" demonstrates the institutional advantages in knowledge source credibility, systematic organization, and standardized ordering. The coupling of these three elements driven by AI technology enables AI corpus libraries to break through the resource management paradigm of digital libraries centered on "human retrieval and reading," shifting instead to the core goal of "model usability, algorithm understandability, and process verifiability," and achieving an overall leap from "data resource management" to "intelligent knowledge supply."

This paper argues that an AI corpus library refers to a comprehensive knowledge infrastructure that, against the backdrop of deep integration among AI, science and technology, and economic and social development, takes high-quality, multimodal, and computable corpus resources as its core organizational objects. It integrates data governance, knowledge organization, and intelligent corpus service functions to provide corpus discovery, acquisition, understanding, training, and application support for diverse stakeholders including researchers, public institutions, and intelligent application systems.

3 Functional Requirements and Architectural Logic of AI Corpus Library

Social needs are the fundamental driving force for library functional evolution. The evolution of library functions from court collections to public services, and from universal services to specialized academic services, represents the inevitable result of demand orientation. The construction of AI corpus libraries not only responds to the development trend of widespread embedding of large intelligent models and agents in various application scenarios but also meets the model

Figure 1

Figure 1: Figure 1

training needs and data-driven demands of intelligent applications. It further enables centralized preservation, long-term preservation, and repeated utilization of corpora while providing basic guarantees for the certification and auditing of intelligent applications. The construction of AI corpus libraries must align with the development characteristics of the digital intelligence era, reflecting the functional requirements and architectural logic of new-era knowledge infrastructure, as shown in Figure 1

.

Figure 1. The construction approach of an AI corpus library

3.1 Functional Requirements of AI Corpus Library

The construction of AI corpus libraries represents not only a breakthrough in existing functions such as retrieval and storage of digital libraries but also the design of entirely new functions including participation in certification and auditing services. Its functional system is ultimately shaped by multidimensional demands including technological evolution, social application, and governance regulation, primarily concentrated in the following three aspects.

3.1.1 Technical Application Needs in the Context of AI Ubiquity

From the perspective of technological evolution, the rapid iteration of large models, generative AI, and multimodal intelligence has placed higher demands on corpora. Enhanced model capabilities depend on the stable supply of sufficient-scale, reliable-quality, multimodal corpora, requiring attributes such as rapid updating, dynamic expansion, and version control. As AI deepens from "general intelligence" to "domain intelligence," models' dependence on vertical domain knowledge, professional rules, and high-credibility training materials has significantly increased, making the rigor, professionalism, and explanatory frameworks of corpus resources core factors affecting model performance.

From the perspective of technical deployment and application diffusion, AI is comprehensively penetrating ubiquitous intelligent application scenarios in scientific research, industry, government, and social services [29], extending corpus needs from single-point training to comprehensive requirements covering multiple stages including reasoning, retrieval, generation, and evaluation. Corpora must not only possess structural and organizational characteristics but also adapt to different algorithmic processes in terms of expression methods, knowledge relationships, and data organization logic to ensure the stability and reliability of AI systems across multiple application scenarios.

Therefore, AI corpus libraries need to construct corpus systems that can simultaneously serve model training, knowledge representation, and cross-scenario

applications, enabling corpora to possess supporting capabilities for AI technology application evolution in terms of professional depth, structural integrity, updating mechanisms, and knowledge organization methods.

3.1.2 Social Operation and Service Needs in the AI Era As society enters the AI era, fields such as scientific and technological innovation, industrial development, public governance, and social services have widespread and continuously growing demands for various AI systems, presenting stronger scenario-based and personalized characteristics. Scientific research activities require intelligent research systems that can support interdisciplinary knowledge association and complex problem analysis; industrial sectors depend on professional models with industry knowledge, process understanding, and task execution capabilities; public governance requires intelligent governance systems that can support policy modeling, trend prediction, and emergency analysis; and social services require agents capable of public interaction, content generation, and multimodal processing.

The effective operation of these systems depends not only on basic corpus supply but also on corpus collections containing task attributes, contextual clues, industry logic, or decision-making bases, enabling them to make targeted responses to specific scenarios. Different from the "knowledge depth" emphasized in the technological evolution stage, social operation demands "task adaptability," "scenario fit," and "application embeddability" from corpora. For example, intelligent research systems require academic corpora that can support problem representation and knowledge deduction, intelligent industrial systems require professional corpora with industry regulations and business contexts, and intelligent governance systems require policy corpora with strict content, reliable sources, and clear explanatory paths.

Therefore, corpus libraries need to construct corpus systems that can support multi-domain, multi-task, and multi-system collaborative operation, providing continuous and stable knowledge supply for AI applications in key social links such as science and technology, industry, and governance.

3.1.3 Supervision, Governance, and Compliance Audit Needs As the scale of AI applications expands and risk structures become increasingly complex, regulation, auditing, and compliance have become important components of governance in the intelligent era. Insufficiencies in source annotation, authorization chains, and responsibility records of model training corpora have made supervision and review agencies urgently need infrastructure capable of corpus preservation, tracing, and verification.

The demands posed by the governance dimension on corpora are mainly reflected in the following aspects: (1) **Responsibility auditing needs**: requiring preservation of source records, version chains, and change histories of corpora used in model training to enable responsibility tracking and restoration analysis when

algorithms exhibit biases, content violations, or social risk events. (2) **Compliance management needs**: requiring clear authorization scopes, usage conditions, rights relationships, and legal basis of corpora to provide a reviewable evidence system for model training and application deployment. (3) **Trusted evidence preservation needs**: for scenarios such as judicial evidence collection, content verification, deep generation verification, and algorithm authentication, requiring authoritative, standardized, and tamper-proof corpus evidence preservation mechanisms.

Therefore, corpus libraries can play a critical role in AI governance systems, not only ensuring corpus preservation and management but also constructing institutionalized corpus infrastructure that supports regulatory review, compliance verification, and risk governance, enabling AI applications to operate safely within a verifiable, accountable, and regulatable framework.

3.2 Development-oriented Architectural Logic

The AI corpus library is a dynamic evolutionary system aimed at achieving the goals of "data/information–corpus/knowledge–intelligence/governance,"and its structural design should serve the long-term needs of AI development.

3.2.1 Top-level Logic Design with Data-driven Concept Despite extremely rapid technological updates, library development philosophy possesses long-term stability. Therefore, AI corpus libraries should adhere to the construction principle of being vision-led rather than technology-led [30], clarifying missions, goals, and public values as the foundation for absorbing new technologies and strengthening the systematic logic of data-driven approaches [31]:

First, configure resource elements with data flow as the core. Build complete data processes around data collection, processing, annotation, updating, invocation, feedback, and revision, rather than focusing solely on collection scale.

Second, orient performance toward data value transformation. Accumulate and amplify data value through training, reasoning, and generation, avoiding long-term data sedimentation that fails to realize utility.

Third, establish embedded data governance as the institutional foundation. Address weak governance links through traceability mechanisms, unified standards, and compliance verification to ensure full-process data controllability, traceability, and reliability.

3.2.2 Base Framework Construction with High-quality Data In the AI context, the consensus that "data flow determines the system capability ceiling" has been established. Both model training performance and intelligent system decision-making quality directly depend on the reliability of underlying data and corpora. Therefore, high-quality data constitutes an irreplaceable foundational base for corpus libraries [16].

In terms of resource structure, base data includes various types such as general corpora, professional corpora, scientific research corpora, synthetic data, and augmented data. In terms of quality requirements, they should possess authenticity, completeness, consistency, timeliness, verifiability, and explainability. In terms of technical attributes, they should also have structural characteristics convenient for algorithmic processing, supporting flexible invocation and automated processing. The maturity of the base framework will directly determine whether corpus libraries can become the core foundation of AI4S.

3.2.3 Intermediate Content Construction with Knowledge Organization Methods The goal of corpus libraries is not only to provide training data but also to construct a "computable knowledge space" for reasoning and knowledge generation through deep processing. Therefore, the construction focus of the intermediate layer lies in knowledge organization rather than simple data management.

The main mechanisms of corpus knowledge organization include: (1) **Knowledge extraction**: using natural language processing and information extraction technologies to identify concepts, entities, relationships, and rules from unstructured resources, achieving knowledge explicitation [32]. (2) **Associative organization**: through knowledge graphs, semantic networks, and association rule modeling, systematically associating dispersed corpus elements to form knowledge systems across documents, datasets, modalities, and disciplines [33]. Relying on these two mechanisms, corpus libraries evolve from "resource aggregation" to "knowledge computation," enabling AI to not only access information but also understand relationships, infer structures, and generate new knowledge.

3.2.4 Main Functional Innovation with Agent Applications Compared with digital libraries, the direct "users" of corpus libraries are increasingly turning to various intelligent agents. As comprehensive carriers for task execution, agents conduct reasoning, generation, and decision-making by invoking corpora [34]. Therefore, application support oriented toward agents has become a key innovation direction for corpus libraries.

On one hand, AI corpus libraries provide continuously updated training and reasoning resources for research agents, supporting AI4S capabilities such as automatic literature review, hypothesis generation, and experimental design optimization [35]. On the other hand, they also provide authoritative and reliably sourced corpora for government, industry, and public service agents, maintaining stability and credibility in knowledge sources, decision bases, and result explanations. Corpus libraries functionally transform from "resource providers" to "task participants," and from "serving human cognition" to "core nodes supporting human-machine collaborative intelligence."

4 Practical Cases and Construction Path of AI Corpus Library

Through the corpus-based transformation of existing digital library collections and external resources, some digital libraries are achieving a transition from "document centers" to "corpus centers." The following analysis of representative practices from HathiTrust and the British Library provides references for clarifying the construction pathways of AI corpus libraries.

4.1 Representative Cases

4.1.1 "Text as Data" : The HathiTrust Case HathiTrust is a digital library consortium initiated by multiple research university libraries in the United States, initially focusing on long-term preservation and academic access of large-scale digital collections [36]. With the rise of computational text analysis, digital humanities, and data-intensive research, traditional digital collection services oriented toward "reading" could no longer meet research needs. To address this, HathiTrust established the HathiTrust Research Center (HTRC) [37], proposing the "text as data" concept [38] to expand corpus service capabilities for computational analysis. Its key work includes: first, datafying and corporatizing texts in various forms; second, providing various data and corpus analysis tools.

(1) Text Datafication and Corporaization

HathiTrust has accumulated over 18 million digital resources, forming a massive collection of raw text. Regarding text datafication and corporaization, HTRC has primarily implemented the following measures:

First, constructing derived datasets [39]. HTRC provides derived datasets, particularly the flagship extracted features dataset, transforming full-text into downloadable, computable research data that complies with the "non-consumptive use policy" [40]. Such data does not directly provide copyright-protected full text but systematically extracts multidimensional features including volume and page metadata, part-of-speech-tagged tokens, and token statistics from full volumes, enabling researchers to conduct large-scale text analysis and model studies without accessing original content. HTRC continuously collaborates with researchers to develop various derived datasets, providing them openly to global researchers, expanding the usability and reuse value of text corpora in different research scenarios, and thereby transforming digital collections into high-value computational corpora.

Second, establishing a workset customization mechanism [41]. Based on text datafication, HTRC enables on-demand organization and fine-grained invocation of corpora through the worksets mechanism. Worksets are formed by researchers selecting and combining volumes from collections according to specific questions, serving directly as analysis objects and enabling text analysis through HTRC' s computational tools. Through this mechanism, corpus organization shifts from platform-unified supply to researcher-led contextual construction,

enabling precise matching of subsets to different research tasks and methods while supporting sharing and citation, enhancing research reproducibility and transparency, and demonstrating the important function of corpus libraries in standardized research.

(2) Providing Data and Corpus Analysis Tools

After completing corpora datafication and customization, HTRC enables effective corpus invocation and direct service to research practice through a supporting analysis tool system. This approach reduces the technical threshold for computational text analysis while ensuring analysis process security and compliance through controlled computing environments.

First, launching the "Bookworm" word frequency trend visualization tool [42]. Based on HathiTrust' s massive digital collections, Bookworm provides visualization of individual word frequency occurrence and temporal variation trends in large-scale text corpora. The tool uses pre-calculated word frequency statistics, does not provide full-text reading, but supports "distant reading" style macro-text research, applicable to conceptual history, intellectual history, and academic discourse evolution analysis. While ensuring copyright and system performance, Bookworm supports researchers in exploratory analysis and hypothesis generation through low-threshold, immediate feedback, embodying the tool-based service philosophy of corpus libraries for computational research.

Second, providing web-based "click-and-run"algorithmic tools [43]. HTRC offers a suite of online algorithmic services enabling researchers to conduct computational analysis on public or self-built worksets through web interfaces without complex programming. The algorithms cover text exploration, statistical analysis, and visualization tasks, transforming corpus analysis from a "data acquisition—local processing" model to a "platform-based computation—result output" model, reflecting the demand-oriented design philosophy of corpus libraries.

Third, constructing secure, controlled computational analysis environments through data capsules [44]. To address the issue that copyright-protected texts cannot be directly downloaded, HTRC has built secure, controlled computational environments enabling researchers to conduct research-driven text analysis within the platform. Public domain resources are available for analysis by all users, while computational access to copyright-protected content is strictly limited to researchers from member institutions. By "sending computation to data" rather than "delivering data to users," data capsules enable intensive computational analysis while ensuring copyright security and compliance, demonstrating the balancing capability of corpus libraries between technical implementation and institutional constraints.

4.2.2 "Digital Scholarship" : The British Library Case Based on its existing digital library infrastructure, the British Library has systematically introduced machine learning and AI methods to continuously advance collection corporaization transformation and computational utilization, gradually forming

an AI corpus practice path with distinctive features of public cultural institutions. Its corpus capabilities do not rely on a single technical project but are built upon long-accumulated national-level digital collection systems, legal deposit systems, and mature governance frameworks, reflecting the endogenous logic of evolution from digital libraries to corpus libraries.

(1) Enriching Collections with AI: Text Transcription as the Core

Text transcription is the foundational link in corpus construction. The British Library has widely adopted non-generative machine learning technologies such as Optical Character Recognition (OCR), Handwritten Text Recognition (HTR), and Automatic Speech Recognition (ASR) to conduct large-scale processing of printed documents, manuscripts, multilingual resources, and oral history materials [45], systematically incorporating transcription results into digital collection management and transforming them from "readable texts" to "analyzable data."

(2) Utilizing Collections with AI: Computational Use Governance Mechanisms

At the corpus utilization level, the library has proposed the concepts of "collections as data" and "computational use" [46], implementing structural processing and batch computational support for digital collections to meet text mining and machine learning needs. On one hand, it enhances collection usability in algorithmic processing through standardized metadata, language identification, and text structure optimization. On the other hand, it supports large-scale computational analysis of protected resources within copyright and legal frameworks through non-consumptive text and data mining mechanisms and tiered access control.

(3) Creating Analysis Tools and Demonstration Projects: Embedding in Digital Scholarship Workflows

Relying on the British Library Labs [47] and digital scholarship & data science support systems [48], the library continuously develops analysis tools and demonstration projects for collection corpora through internal pilots and cross-institutional collaborations, embedding AI methods into specific research workflows. For example, in the "Living with Machines" project [49], research teams comprehensively applied text mining, word embedding, and computer vision methods to conduct computational analysis of large-scale historical collections, demonstrating the potential of digital collections as general AI corpora.

(4) Institutional Construction Oriented toward Responsible AI

As a national-level public cultural institution, the British Library emphasizes publicness, responsibility, and long-term sustainability in AI corpus construction, ensuring a balance between openness, usability, and regulatability of corpora through clear ethical principles, usage norms, and responsibility boundaries. Its construction path based on collections, premised on governance, and guided by digital scholarship demonstrates that AI corpus libraries are not independent

technical platforms but structural upgrades of digital libraries in the direction of intelligence and computability, providing a replicable practical paradigm for public sector and academic institutions.

4.2 Construction Path of Corpus Library

Comprehensive analysis of the above cases reveals that AI corpus libraries are not platforms constructed independently from existing digital library systems. Instead, they represent systematic upgrades of resource forms, service targets, and functional structures through the introduction of "computational use" concepts and corpora capabilities based on digital libraries. Their construction pathways exhibit obvious gradualness and transformative characteristics, mainly reflected in the following aspects.

(1) Conducting Corporaization Transformation Based on Digital Collections

Whether through HathiTrust' s construction of derived datasets from large-scale digital texts or the British Library' s use of OCR, HTR, and other technologies to promote text transcription and structural processing, high-quality, large-scale digital collections are important resources for generating general AI corpora. By repositioning "readable literature" as "computable objects," libraries lay a solid foundation for subsequent computational analysis and model training.

(2) Constructing Computational Use Mechanisms Premised on Corpora Governance

The key distinction between corpus libraries and general data platforms lies in institutional guarantees. Both cases attach great importance to copyright compliance, non-consumptive use, and access control, embedding computational processes within platform environments through institutional design. This approach supports large-scale analysis of protected resources while ensuring resource security and public trust. This governance-first corpora model demonstrates the unique advantages of libraries in public sector corpus supply.

(3) Traction by Research Needs for Corpora Organization and Analysis Services

HathiTrust supports researchers in organizing corpora on demand through worksets and online analysis tools, while the British Library embeds corpora into specific research workflows through digital scholarship projects and experimental platforms. Both cases demonstrate that corpus libraries not only provide static datasets but also reduce computational research thresholds and enhance corpus usability in scientific practice through tool, environment, and workflow design.

(4) Transformation Construction Rather Than New Construction

In the context of public research and cultural institutions, AI corpus libraries typically rely on mature digital library systems, introducing computational cor-

pora, intelligent processing, and digital scholarship services to gradually expand service targets from literature services for human readers to corpus services for both researchers and intelligent systems. This transformation construction model features controllable investment, stable institutions, and strong sustainability, possessing promotion and replication value.

In summary, the construction of AI corpus libraries is not only a technological upgrade but also a systematic transformation of digital libraries based on collections, guaranteed by governance, and oriented toward computational research. This practice provides feasible pathways and experiences for library and information institutions to reshape knowledge infrastructure in the AI context.

5 Conclusion

This paper systematically analyzes the connotation, functional requirements, and construction pathways of AI corpus libraries. At the theoretical level, it clarifies their relationship with traditional corpora, databases, and digital libraries, defining a functional positioning of "model usability, process verifiability, and resource traceability." At the technical level, it sorts out functional requirements for technological development, social services, and compliance auditing, as well as an architectural logic based on high-quality datasets, knowledge organization methods, and agent application outputs. At the practical level, it selects cases from HathiTrust and the British Library to demonstrate the gradual transformation pathways from digital collections to corpus libraries.

The study shows that AI corpus libraries not only achieve the scaling and structuralization of corpus resources but also, through institutionalized governance, knowledge organization, and application-oriented service mechanisms, transform corpora from "accessible" to "computable, verifiable, accountable, and sustainably usable" core social infrastructure. This transformation enables libraries to gradually upgrade from traditional knowledge preservation and provision roles to intelligent knowledge supply and governance pivot points for scientific research, industry, and social governance in the intelligent era. Practical experience demonstrates that the gradual construction pathway based on existing digital collections, oriented toward computational use, and premised on responsible governance possesses practical operability and promotion value.

In the future, as large model capabilities improve and intelligent application scenarios expand, AI corpus libraries must further achieve dynamic balance between open sharing and security compliance, maintain institutional tension between technological innovation and public value, and continuously explore new methods for constructing multimodal, high-credibility, traceable corpora, cross-domain knowledge integration, and intelligent service support, providing solid support for trustworthy AI and public knowledge infrastructure development.

References

[1] China State Council. Opinions on Deepening the Implementation of the "AI+"Action[EB/OL]. (2025-08-26)[2025-10-20]. https://www.gov.cn/zhengce/content/202508/content_{7037
[2] UK Government. AI for Science Strategy[EB/OL].[2025-12-08]. https://www.gov.uk/government/publicatio for-science-strategy/ai-for-science-strategy
[3] The White House. Launching the Genesis Mission[EB/OL].[2025-12-08]. https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/.
[4] US Department of Energy. Genesis Mission: A National Mission to Accelerate Science Through Artificial Intelligence [EB/OL].[2025-12-08].https://genesis.energy.gov/.
[5] LI Guojie. AI4R: The fifth scientific research paradigm[J]. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 1-9.
[6] QIU Baohua. AI Large Models Must Grasp the Key of Training Data[N]. Study Times, November 28, 2025, Page 3.
[7] LAION-5B[EB/OL].[2025-12-09].https://laion.ai/blog/laion-5b/.
[8] Common Crawl[EB/OL].[2025-12-09].https://commoncrawl.org/.
[9] ImageNet[EB/OL].[2025-12-09].https://www.image-net.org/.
[10] Qiu Pengcheng, Wu Chaoyi, Zhang Xiaoman. et al. Towards building multilingual language model for medicine[J]. Nature Communication, 2024, 15: 8384. https://doi.org/10.1038/s41467-024-52417-z.
[11] Food and Agriculture Organization of the United Nations. AGROVOC[EB/OL].[2025-12-09].https://www.fao.org/agrovoc/.
[12] The Materials Project[EB/OL].[2025-12-09].https://next-gen.materialsproject.org/.
[13] CHEN Qiang, XING Qieqie. Corpus Development for Generative AI: Current Landscape, Challenges, and Strategic Pathways[J]. Journal of Social Sciences2025(6): 177-192.
[14] CAI Lin. The Legal Framework for the Obligation of Traceability of Corpus Sources in Generative Artificial Intelligence[J]. Jiangsu Social Sciences, 2025(2): 161-169, 243-244.
[15] WANG Limei, WANG Qian, TAO Qian, et al. Legal Issues in the Development of Corpus for Artificial Intelligence[J]. Digital Law, 2024(5): 28-46.
[16] ZHANG Xiaolin. High-quality Dataset Construction in Al + Environment: Library' s Opportunities and Challenges[J]. Journal of Library Science in China, 2025, 51(280): 4-17.
[17] RANGANATHAN S R. The Five Laws of Library Science[M]. Translated by XIA Yun, WANG Xianlin, ZHENG Ting, et al. Proofread by HOU Hanqing. Beijing: Bibliography and Document Publishing House, 1988: 308-337.
[18] JIN Shengyong, ZHANG Wenqiu. Recognition of Library Social Function[J]. Library, 2014 (6): 38-41.
[19] WANG Shiwei. Recognition of the Three Functions of Public Library in Inheriting Civilization and Serving Society[J]. Library Journal, 2019, 38 (10): 24-28.
[20] JIANG Yongfu. Library and Knowledge Organisation: Understanding Library Science from the Point of View of Knowledge Organisation[J]. Journal

of Library Science in China, 1999 (5): 19-23.
[21] JIA Junzhi. From Cataloging to Metadata Management: The Development of Library Knowledge Organization[J]. Journal of Library Science in China, 2023, 49 (2): 121-131.
[22] CHENG Huanwen. Three Dimensional Analysis of Smart Library[J]. Library Tribune, 2021, 41 (6): 43-55.
[23] HUANG Zongzhong. "On the Collection and Utilization of Libraries" [J]. Journal of Wuhan University (Humanities Science), 1962 (2): 91-98.
[24] PENG Junting. Reflections on the "Collection" and "Utilization" of Libraries[J]. Journal of Beijing Normal College (Social Sciences Edition), 1991(3): 105-109, 57.
[25] FOX E A, AKSCYN R M, FURUTA R K, et al. Digital libraries[J]. Communications of the ACM, 1995, 38(4): 22-28.
[26] ZHAO Xichen. Research and Construction of the Digital Library[J]. Modern Library and Information Technology, 1999(1): 28-31, 43.
[27] WANG Shiwei. On Three Main Features of the Smart Library[J]. Journal of Library Science in China, 2012, 38 (6): 22-28.
[28] LV Tingyu, LI Xiaoying, ZHANG Ying, et al. Study on the Construction of a Question-Answer Corpus Dataset for Chinese Medical Knowledge Large Language Models[J]. Journal of Medical Informatics, 2024, 45 (5): 20-25.
[29] ZHANG Xiaolin, LIANG Na Knowledge Is towards Being Wisdom, Wisdom Needs to Be Scenario-based, and Intelligence Can Be Ubiquitously Embedded Exploration of the Logical Framework of Intelligent Knowledge Services[J]. Journal of Library Science in China, 2023, 49 (3): 4-18.
[30] Johnson R, Refsum C. Governing in the Age of AI: Building Britain' s National Data Library[R/OL].(2025-02-25).[2025-10-30].https://institute.global/insights/tech-and-digitalisat ion/governing-in-the-age-of-ai-building-britains-national-data-library.
[31] LIU Xiwen, FU Yun. The Research Advancements in Information Science Amidst Data & Al Empowerment: Data-Driven, Model-Driven and Knowledge Discovery[J]. Advances in Information Science, 2024, 15 (00): 131-171.
[32] CHEN Yucheng, HAN Tao, HU Zhengyin. Research on the Hint Framework for Knowledge Extraction From Scientific and Technological Literature Text[J/OL][2025-12-18]. Journal of Modern Information, https://link.cnki.net/urlid/22.1182.G3.20250923.1636.006.
[33] ZHAO Chenyang, WANG Desheng, ZHANG Jun, et al. Application of Heterogeneous Resource Organization and Association Discovery Technology[J]. Computer and Network, 2019, 45 (11): 69-71.
[34] LIU Xiwen, SUN Mengge, FU Yun. From MIS to DIS Agent: Reshaping the Paradigm of S&T Documentation and Information Service[J]. Library and Information Service, 2025, 69 (17): 3-15.
[35] GUO Limin, LIU Yueru, FU Yaming. Design and Transformation Pathways of Library Systems Driven by Generative Agents[J]. Journal of Library and Information Science in Agriculture[J/OL][2025-12-18], https://doi.org/10.13998/j.cnki.issn1002-1248.25-0562.
[36] HathiTrust[EB/OL].[2025-12-19].https://www.hathitrust.org/.

[37] HathiTrust Research Center[EB/OL].[2025-12-19]. https://www.hathitrust.org/about/research-center/.
[38] HathiTrust. Research Center Analytics[EB/OL].[2026-01-08]. https://analytics.hathitrust.org/.
[39] HathiTrust. Derived datasets[EB/OL].[2026-01-08]. https://analytics.hathitrust.org/deriveddatasets.
[40] HathiTrust. Non-Consumptive Use Policy[EB/OL].[2025-12-19]. https://www.hathitrust.org/the-collection/terms-conditions/non-consumptive-use-policy/.
[41] HathiTrust. Worksets[EB/OL].[2026-01-08]. https://analytics.hathitrust.org/staticworksets.
[42] HathiTrust. Bookworm[EB/OL].[2026-01-08]. https://bookworm.htrc.illinois.edu/develop/.
[43] HathiTrust. Algorithms[EB/OL].[2026-01-08]. https://analytics.hathitrust.org/statisticalalgorithms
[44] HathiTrust. Data capsules[EB/OL].[2026-01-08]. https://analytics.hathitrust.org/staticcapsules.
[45] British Library. AI and machine learning with British Library collections[EB/OL].[2025-12-19]. https://www.bl.uk/stories/blogs/posts/ai-and-machine-learning-with-british-library-collections
[46] Candela G. Collections as Data: Getting Started[EB/OL].[2026-01-08]. https://bl.iro.bl.uk/concern/generic_{works}/dbc38f7b-664c-42ec-8c7d-c6db095ddefd.
[47] British Library. British Library Labs: Projects[EB/OL].[2025-12-19]. https://bl.iro.bl.uk/collections/36116aa1-7037-40f3-9b91-ecb1be15e226?locale=it&view=gallery.
[48] British Library. Digital Scholarship & Data Science Topic Guides[EB/OL].[2025-12-19]. https://bl.iro.bl.uk/collections/a4e0122e-c4c9-47e4-b5c9-7921baedea5b?locale=en.
[49] British Library.. Living with machines[EB/OL].[2026-01-08]. https://livingwithmachines.ac.uk/.

Author Contributions:

Liu Xiwen: Conceptualized the research problem and framework, collected materials, wrote, revised, and finalized the manuscript;
Qian Li: Discussed research ideas, revised and reviewed the manuscript;
Tu Zhifang: Discussed the manuscript framework, wrote and revised the manuscript.