Construction of a Big Data System for Scientific Literature for Smart Knowledge Services: Post-print

Authors: Wu Zhenxin, Qian Li, Xie Jing, Chang Zhijun, Xu Liyuan, Zhao Yan

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] To explore the construction of a big data knowledge resource system for literature and intelligence, supporting intelligent knowledge services across multiple domains.

[Method/Process] Grounded in AI application requirements and drawing on industry best practices, we systematically identified limitations in existing resource systems and expanded them across multiple levels and dimensions; constructed reliable data processing pipelines and computing platforms to enable efficient data collection and processing; and developed intelligent data governance tools to achieve effective governance of knowledge resources and ensure the provision of high-quality data assets.

[Results/Conclusion] We have initially established a multi-type, multi-disciplinary big data knowledge resource system for scientific and technical literature, completed the construction of highly automated data collection and governance pipelines, implemented multi-layered data quality control mechanisms, accumulated hundreds of millions of high-quality data records, and provided data support for multiple knowledge services.

Full Text

Preamble

Construction of a Sci-Tech Literature Big Data System for Intelligent Knowledge Services

Wu Zhenxin[1],[2], Qian Li[1],[2], Xie Jing[1],[2], Chang Zhijun[1],[2], Xu Liyuan[1], Zhao Yan[1],[2]

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

Abstract: *[Purpose/Significance]* This paper explores the construction of a big data knowledge resource system for literature intelligence to support intelligent knowledge services across multiple domains. *[Method/Process]* Based on AI application requirements and drawing on industry experience, we identified problems in existing resource systems and expanded the resource framework across multiple levels and dimensions. We constructed reliable data processing workflows and computing platforms to support efficient data collection and processing, and developed intelligent data governance tools to achieve effective governance of knowledge resources and ensure the provision of high-quality data resources. *[Result/Conclusion]* We have initially formed a multi-type, multi-disciplinary big data knowledge resource system for sci-tech literature, established a highly automated data collection and governance process, implemented multiple data quality controls, accumulated hundreds of millions of high-quality data records, and provided data support for multiple knowledge services.

Keywords: sci-tech big data; knowledge resource architecture; data aggregation; intelligent knowledge services

Classification Number: G250.7
DOI: 10.13266/j.issn.0252-3116.2020.24.008

Artificial intelligence (AI) and big data have become general-purpose technologies influencing various social sectors, disrupting and transforming every industry they touch. Similarly, they have propelled scientific research breakthroughs through entirely new paradigms [1] and provided a novel paradigm for knowledge services, thereby stimulating strong demand for intelligent knowledge services.

Intelligent knowledge services involve building intelligent literature intelligence systems by fully leveraging AI and big data technologies, transforming sci-tech intelligence work into flexible "data cleaning plants," "information processing plants," "knowledge generation plants," and "decision-making plants" centered around intelligent literature systems. This enables sci-tech intelligence work to quickly perceive changes, refine problems, focus on targets, and formulate solutions, greatly compensating for human intelligence limitations and enhancing our ability to address complex problems and tasks. The construction of such systems will become crucial for building future core competitiveness.

2. Design of the Sci-Tech Literature Big Data System

The sci-tech literature big data system primarily comprises a data architecture, management platform, and the standards, specifications, and technical methods surrounding these two components. Based on a demand-driven design philosophy, we first analyzed the data requirements for supporting intelligent knowl-

edge services, then formulated design concepts and completed the framework design.

2.1 Data Requirements Analysis for Supporting Intelligent Knowledge Services

2.1.1 AI Application Requirements Intelligent knowledge services are characterized by AI applications, requiring us to analyze the impact of and demands placed by AI applications on big data systems.

Data scientist R. Monica proposed an AI Hierarchy of Needs [2], where AI application workflows progress from underlying data collection, storage, and cleaning to gradual AI application, with each stage corresponding to different data and processing requirements. The entire workflow increases in difficulty progressively (see Figure 1 [Figure 1: see original paper]). She argues that a solid data foundation is the primary element, while reliable data processes and convenient data tools are also critical for AI applications. The computer field generally recognizes that three factors drive AI applications: algorithms, computing power, and data, with data being the core competitive advantage. Benefiting from AI requires massive training datasets [3].

A recent report by iResearch Consulting, "2020 China AI Basic Data Service Industry Development Report," states that AI commercialization has basically reached stage maturity in computing power, algorithms, and technology. To achieve greater implementation and address specific industry pain points, large amounts of relevant data processed through annotation are needed to support algorithm and model training [4]. Thus, high-quality data (especially annotated data), reliable data governance processes, and diverse governance tools are key to AI applications.

2.1.2 Analysis and Reference of International Publishers' Intelligent Knowledge Services The application of big data and AI technologies is also driving changes in sci-tech knowledge service models [5,6]. International publishers have leveraged their data advantages to pioneer new knowledge services using AI technology. We investigated their intelligent knowledge services as references for our architectural design.

Elsevier [7] has constructed a new research ecosystem covering data, evidence, tools, and intelligent services, releasing a series of digital and knowledge-based tools. Digital Science [8] has proposed a new research information service model for the entire research lifecycle, forming a data system covering four dimensions—researchers, research institutions, funding programs, and publications—and developing various intelligent tools. Taylor & Francis, in addition to its proprietary data, integrates data from multiple other sources. Its knowledge graph tool, Wizdom.ai [9], contains hundreds of millions of data records covering publications, patents, authors, institutions, concepts, and facts. Their work primarily focuses on these three aspects: data, platform, and tools.

2.1.3 Extension Requirements Analysis for Problem-Oriented Sci-Tech Big Data Systems Based on the above investigation and analysis, we carefully examined the existing architecture. To apply AI technology and support intelligent knowledge services, we need to provide richer high-quality data (including annotated datasets), reliable data governance processes, and diverse governance tools. We must also address challenges such as multi-source data integration, data indexing, deep knowledge fusion, and domain knowledge graph construction. These require the original data system to be expanded across multiple levels.

From the data perspective, we must enrich and extend the existing sci-tech literature data system to form a multi-level data cluster oriented toward different functions. Therefore, in addition to traditional sci-tech literature basic data clusters, we must establish a data governance support cluster to enable intelligent data processing and value addition, and a sci-tech knowledge association computing cluster to support knowledge computing and intelligent knowledge generation and decision-making.

From the process perspective, we must transform the original data processing workflow, centering on data governance and embedding AI and big data technologies to enhance overall reliability and efficiency. This ensures both efficient large-scale data collection and computation, enabling timely data governance and updates, and obtaining high-quality data resources to support intelligent knowledge services through intelligent data governance.

From the perspective of diverse intelligent tools, we must introduce new technologies and methods such as AI and knowledge mining, develop high-performance intelligent tools, conduct in-depth mining and reconstruction of sci-tech knowledge, expand entities and relationships, and promote the enrichment, fine-granularity, and semanticization of the knowledge system.

2.2 Basic Concepts

Based on the three requirements above, we formulated three construction concepts for the sci-tech big data system.

2.2.1 Expand the Big Data Resource System and Enrich Support Data Clusters Across Multiple Dimensions Building upon the original big data system, we comprehensively reviewed and expanded authoritative, accessible data sources, reorganizing basic data into the following five categories: (1) Research entities, including experts, research institutions, academic journals, research teams, publishing platforms, tech enterprises, and funding agencies; (2) Research activities, including research projects, academic conferences, training exchanges, tech competitions, data sharing, news, social activities, and sci-tech policies; (3) Research outputs, including papers, patents, reports, awards, monographs, standards, software, products, and data; (4) Research facilities, including large scientific installations, instruments, equipment, and consumables; and

(5) Scientific data, including research datasets. Ultimately, we aim to establish a complete sci-tech big data ecosystem covering multiple types, channels, and users, including literature, information, professional datasets, and research entities.

2.2.2 Center on Data Governance to Build an Efficient Data Governance Platform We transformed the original data processing workflow, centering on data governance and building an efficient data governance platform based on AI and big data technologies. This enables platform-based operations for data resource collection, storage, computation, and management within the sci-tech literature big data ecosystem, achieving multi-source data organization and integration. This forms a sci-tech literature big data system that effectively improves the collection, aggregation, computation, organization, fusion, and revelation capabilities of various resources.

2.2.3 Develop High-Quality Basic Data Resources Through Intelligent Data Governance Tools We developed diverse intelligent data governance tools to achieve effective governance of knowledge resources, ensuring the provision of high-quality data resources. These tools enable multi-source heterogeneous data aggregation, fusion, cleaning, standardization, knowledge extraction, and relationship extraction.

2.3 Framework for a Data Governance-Centered Sci-Tech Literature Big Data System

Based on the above concepts, we completed the framework design for the sci-tech literature big data system, as shown in Figure 2 [Figure 2: see original paper]. The new framework consists of three components: underlying basic literature data, an intermediate data governance platform, and a post-governance application data layer comprising high-quality authoritative data and domain knowledge data.

From the perspective of supporting intelligent knowledge services, we divided the data architecture into three layers based on different data functions, as shown in Figure 3 [Figure 3: see original paper].

2.3.1 Sci-Tech Literature Basic Data: The Raw Data Layer The underlying data primarily consists of literature and web resources, representing the most fundamental and original data resources of the sci-tech literature big data system and serving as the basis for intelligent mining and analysis. This includes: (1) Commercial publishing resources: journal articles, conference papers, dissertations, patents, sci-tech intelligence, reports, standards, books, periodicals, reference works, product samples, and numerical datasets; (2) Open access resources: journal articles obtained through official OA interfaces, mainly from Cornell University Library's arXiv [10] and the U.S. National Center for Biotechnology Information' s PMC [11]; (3) Chinese Academy of Sciences (CAS) system

resources: information accumulated over many years from nearly 100 institute institutional repositories, dozens of featured data centers and professional centers under construction during the 13th Five-Year Plan, and self-processed data from NSLC research teams covering broad resource types with strong domain specialization; (4) Web-crawled resources: content automatically collected, selected, described, organized, and revealed from domestic and international institutional websites in NSTL priority areas, including major news, research reports, budgets, funding information, and research activities; and (5) Resources exchanged with relevant institutions: including journal articles, conference papers, patents, sci-tech intelligence, and standards.

2.3.2 Data Governance Foundation Data: The Data Governance Layer This layer primarily includes knowledge base data for data quality control. (1) Normative databases: As a traditional method for data quality control, normative databases remain an important foundation in the sci-tech literature big data system. The big data center aggregates scattered normative databases from various centers and teams through unified management and centralized services to promote collaborative sharing and value realization. These normative data are applied during data cleaning, processing, and organization to improve data quality, including institutional normative databases, name databases, journal databases, and funding project databases. (2) Domain thesauri: Using the STKOS project from the National Science and Technology Library [12] and NSLC' s accumulated resources, we have aggregated massive domain thesauri covering science, engineering, agriculture, and medicine. (3) Knowledge bases: These include three types of information. Multi-source institutional information aggregates data from over 600 institutions and research institutes, including CAS and major domestic research institutions and universities, containing basic institutional information, institutional IP information, and commercial publisher subscription information. Multi-source user information utilizes user data from WOS, iAuthor [13], IR [14], Baidu Scholar, and the CAS unified authentication system, combined with user information accumulated from relevant information service systems, to compile a user base information repository. Multi-platform log information includes years of accumulated log data from institute users on commercial publisher platforms and real-time user log data from various NSLC service platforms, as well as post-processed data extracted and analyzed from log information. (4) Rule bases: Specific cleaning rule sets are established for particular resources and resource attributes to support intelligent data governance tools.

2.3.3 Sci-Tech Knowledge Association Computing Data Cluster: The Knowledge Graph Layer This layer utilizes basic data from the data governance layer, processing raw data through cleaning, standardization, fusion, and extraction to form different types of data collections that provide data services to the application layer. Entity data includes research entities such as scholars, institutions, journals, research topics, conferences, and funding projects

extracted from papers, projects, journals, patents, and other resources. Relationship data includes various relationship data extracted simultaneously with entities, forming an entity relationship database that provides data services for relationship mining and knowledge computing.

3. Key Problem Solutions

Based on the above design, the sci-tech big data system must also provide annotated datasets, reliable data governance processes, and diverse governance tools, while addressing challenges such as data fusion, indexing, and domain knowledge graph construction.

3.1 Refined Data Governance Process Covering the Full Data Ecosystem Lifecycle

3.1.1 Governance Process Covering the Full Data Ecosystem Lifecycle

According to the management requirements of the full data ecosystem lifecycle, we reshaped a refined data governance process covering the entire lifecycle, forming a standardized process with six main stages: data source registration, data harvesting, data warehousing, integration and fusion, knowledge graph construction, and microservices (see Figure 4 [Figure 4: see original paper]). This process embeds multiple intelligent governance function modules developed based on big data, machine learning, and knowledge mining technologies to achieve refined data management.

(1) Data source registration involves implementing basic information such as data source selection, access methods, and business cooperation forms. (2) Data harvesting performs corresponding acquisition processing based on data source release methods, including OAI interface access, database direct connection, FTP file services, and manual acquisition from storage media. Matching configuration templates were developed for each data type, enabling new data source extraction by simply configuring target field paths across sources, greatly improving reception efficiency. (3) Data warehousing uses Hive' s external table storage, with structured data processed through distributed MapReduce ETL parallel computing stored in the data warehouse as the foundation for subsequent computation. (4) Integration and fusion performs business-level deduplication and character complementation on parsed structured data, along with necessary information conversion and supplementation. Current paper fusion rules use title + journal + year as the unique identifier; patents use patent number or application number; and crawled data uses URL md5 codes. The resource aggregation module employs multi-level rule patterns for processing, demonstrating high efficiency for large-scale data fusion. (5) Knowledge graph construction includes three sub-processes: data enrichment, entity extraction, and relationship establishment. Data enrichment improves entity recognition accuracy through web crawling, targeted processing, and

multi-source optimization. Entity extraction processes source data according to entity definitions through extraction and splitting to build entity objects, currently yielding six major entity types: scholars, institutions, journals, research topics, conferences, and funding projects. Relationship building, a crucial part of knowledge graphs, preserves relationships between entities after separation and completes weight calculation through statistical analysis of relationship data to solidify relationship weights. (6) Microservices: The sci-tech literature big data system provides data access services through Restful API interfaces using distributed technology with advantages including elastic scalability, hot registration, high performance, and anti-crawling capabilities.

3.1.2 Modular Process with Personalized Configuration We designed each processing stage modularly, enabling personalized configuration for diverse data sources and formats to build different harvesting, parsing, and cleaning function modules. Each data resource' s processing can achieve personalized configuration. Different resource types are provided with different package identifiers, using different classes to implement different source harvesting. For each new data source, configuring harvesting frequency, fields, type (incremental or full), and start time according to type-specific rules enables new data source harvesting.

3.1.3 Visualized Fully Automated Processing Workflow Implementation uses MapReduce and Spark frameworks for distributed computing. Each complete processing workflow can be configured as a job project, with data processing jobs distributed across multiple servers for synchronous processing through job queue threshold settings and real-time dynamic loading. The platform visually displays each step in the processing workflow, such as "unprocessed," "processing," and "processed." Full and incremental reprocessing can be performed for data from a specific source. This automated workflow reduces complexity, ensures a degree of personalization, and improves processing efficiency.

3.2 Multi-Level Data Quality Control

The main characteristics of intelligent knowledge services are personalization and precision, both requiring high-quality data support. We implemented multi-level data quality control by analyzing data characteristics, standardizing data, and monitoring and verifying data quality to effectively improve data quality and usability.

3.2.1 Unified Metadata Standards and Models The sci-tech literature big data system formulated unified metadata standards based on NSTL's Unified Literature Metadata Standard 3.0 (official version) [15], using XML language and DTD for formal description. The metadata includes 13 element sets: source,

single literature, topic/classification/keywords, contributor/institution, conference, funding, operation information, access management, full-text files, figures, tables, supplementary materials, and references. Excluding duplicate elements and attributes, the standard contains 97 descriptive elements, 53 auxiliary elements, 49 attributes, and 4 special character elements, flexibly combined to describe diverse, multi-level resources.

3.2.2 Normative Database Construction By integrating and aggregating normative databases from different institutions and projects, we have formed four entity normative databases: institutional, name, journal, and funding project databases. These are used to standardize relevant metadata content during data cleaning.

3.2.3 Multi-Level Data Cleaning Rules As a supplement to normative databases, we added quality control for specific resources and attributes. Each resource type has specific cleaning rule sets, and element-specific cleaning rules were established based on relevant standards to normalize fields such as country/region, city, institution name, journal name, publication year, data type, discipline classification, scholar name, and keywords.

3.2.4 Repeatable Cleaning Process Based on the above standards and normative databases, the sci-tech literature big data system monitors and manages the complete data quality lifecycle through standard API interfaces for data collection, aggregation, cleaning, processing, organization, computation, and services. The platform provides reusable cleaning workflows to ensure repeatable operations at each stage for cyclic quality improvement.

3.2.5 Data Processing Tools and Workflows Integrating Expert Knowledge The sci-tech literature big data system uses normative databases and third-party resources as the foundation for computation to perform automated processing of raw data, while adopting data processing tools that integrate expert knowledge. We established a workflow where professionals participate in deep processing of institutions, scholars, journals, and topic terms through data management and processing systems to ensure effective quality control.

3.3 Data Fusion and Indexing for Multi-Source Sci-Tech Big Data

3.3.1 Data Fusion Data fusion is a data governance process combining machine-automated fusion and manual governance, integrating ETL workflow, entity standardization, deduplication, and enrichment into a standardized process. It is both a rule-algorithm-based customized data fusion workflow and a crowdsourced data fusion process.

Each entity type has its own deduplication elements and rules. For journal articles, DOI is the primary deduplication element, followed by a combination

of title + author count + author name + publication year as the secondary deduplication element to complete data deduplication and lay the foundation for data fusion. Based on high-performance computing technologies such as MapReduce and Spark on the big data platform, with Hive data warehouse as the data source and high-speed service indexing through Elasticsearch, the system completes data identification and fusion, as shown in Figure 5 [Figure 5: see original paper].

3.3.2 Data Indexing Dataset classification and indexing represent intellectualization in scientific research and are concerns for data public services. The sci-tech literature big data system achieves multi-dimensional indexing from topic keywords, discipline classification, importance, and publication timeliness. For discipline classification indexing, the system studied multiple classification systems from sources such as NSTL, Chinese Library Classification, Chinese Academy of Sciences Library Classification, ESI, and publisher classifications. Through comprehensive analysis and computational fusion of multi-level disciplines, a discipline classification library suitable for the sci-tech literature big data system was formed and applied to various resources within the system, with each data resource "stamped" with its own discipline classification set. Centered on discipline classification, external services and usage are provided.

The sci-tech literature big data system has developed diverse data governance tools based on AI and semantic technologies, establishing personalized indexing rules for various entities to achieve computer-automated indexing. It has also developed a data processing platform integrating expert knowledge and established a coordinated participation mechanism for data quality control, enabling different user roles such as scholars, institution managers, discipline service teams, and data management teams to participate in data governance and achieve intelligent, precise domain profiling for accurate push and retrieval services.

4. Construction Results and Application Introduction

After more than three years of construction, we have formed a multi-domain, multi-level sci-tech literature big data knowledge resource system [16]. Compared with traditional literature big data systems, the data content is more abundant, including not only traditional literature data resources but also richer data governance data and sci-tech knowledge association computing data. The system features a refined data governance process covering the full data ecosystem lifecycle, embedded with multiple intelligent governance modules and tools capable of providing high-quality knowledge data for intelligent knowledge services.

4.1 Initial Formation of Multiple Data Clusters with Considerable Volume

4.1.1 Sci-Tech Literature Basic Data Cluster This data cluster currently covers eight well-known domestic and international databases including Web of Science, Elsevier, Wiley, Taylor, VIP, CSCD, PMC, and arXiv. It also collects data from over 2,200 important research institutions worldwide such as NIH, NSF, and NSFC, with total data volume exceeding 300 million records. Beyond traditional data types such as journal articles, books, patents, dissertations, sci-tech reports, standards, and ancient texts, the system also covers global funding projects, full-network sci-tech data from major global research institutions and societies, open sci-tech network data from major global sci-tech think tanks, socio-economic information data, policy and regulation data, and numerical data on sci-tech competitiveness from the World Bank and Lausanne reports. Overall, this data cluster features large annual spans and high freshness, with the earliest data dating back to 1799 (patents) and 1900 (literature), spanning 221 years. Data is updated regularly at frequencies of 1 day (literature) and 3 days (patents) to ensure freshness.

4.1.2 Data Governance Foundation Data We have accumulated two major categories: normative databases and domain thesauri, as shown in Table 1 .

Table 1 Supporting Data Resources in the Data Governance Foundation Data Cluster

Resource Type	Source	Scale
Institutional Normative Database	Approximately 90,000 institutional normative names and 900,000 institutional aliases from sources including WOS, iSwitch, patent databases, etc. Institutional websites and institutional repositories of Chinese Academy of Sciences	

Resource Type	Source	Scale
	List of Chinese higher education institutions from the Ministry of Education of the People' s Republic of China [17]	
	Nearly 80,000 institutions in 221 countries from the Global Research Identifier Database (GRID)	
	Global university data from Wikipedia DBpedia [18]	
Scholar Normative Database	Approximately 1.86 million scholars from WOS, iAuthor, IR, Baidu Scholar, etc., associated with 4.2 million literature resources	
	Official websites of CAS institutes	
	CAS institutional repositories	
	China Scientists Online [19]	
Journal Normative Database	Approximately 45,000 records from the National Union Catalog of Serials and automatically collected journal data	

Resource Type	Source	Scale
Funding Project Normative Database	Approximately 5 million funding projects from 11 countries including Australia, Germany, Russia, Canada, USA, EU, Japan, Switzerland, India, UK, and China	
Domain Thesauri	344,735 terms, 104,063 concepts 605,604 terms, 157,570 concepts 241,530 terms, 92,869 concepts 1,128,835 terms, 260,329 concepts	

4.1.3 Sci-Tech Knowledge Association Computing Data Cluster This cluster has accumulated over 300 million research entities including scholars, institutions, papers, patents, journals, and funding projects. It has also accumulated 3.4 billion pairs of relationship data across six entity types and 21 relationship categories, forming a relatively comprehensive entity relationship database that provides intelligent data services for relationship mining and knowledge computing. It can build authoritative knowledge graphs for academic research circles and provide a strong knowledge foundation for intelligent knowledge services, as shown in Table 2 .

Table 2 21 Categories of Research Entity Relationship Data

Constraint-Relationship Type	Entity Types
publish	Literature set/Single literature → Journal
address_{is}	Literature set/Single literature/Institution/Researcher → Country/State/Province/City
source_{is}	Literature set/Single literature → Journal

Constraint-Relationship Type	Entity Types
subject_{is}	Literature set/Single literature/Institution/Researcher/Project → Topic/Classification/Keyword
contributor	Literature set/Single literature → Researcher
affiliation	Researcher → Institution
proceeding_{include}	Conference → Literature set/Single literature
hold_{conference}	Institution → Conference
fund_{by}	Project → Literature set/Single literature
reference	Literature set/Single literature → Literature set/Single literature
fundapply	Literature set/Single literature → Project
manageby	Institution → Project
related_{org}	Institution → Institution
attach_{with}	Full-text/Figure/Table/Supplementary material → Literature set/Single literature
hold_{collection}	Institution → Literature set/Single literature
contribute_{institution}	Literature set/Single literature → Institution
undertake_{conference}	Institution → Conference
associatemedia_{conference}	Institution → Conference
cooperate_{conference}	Institution → Conference
support_{conference}	Institution → Conference
guid_{conference}	Institution → Conference

4.2 Initial Formation of Intelligent Knowledge Service Support Capabilities

The sci-tech literature big data system construction has achieved initial success, providing data support for multiple services including the NSLC portal [20] and the "Hui" series products [21-23], as well as offering various forms of data services [24].

Taking institutional knowledge management and data analysis services [25] as an example, the sci-tech literature big data system provides basic sci-tech literature data clusters and knowledge association computing data clusters. These support automatic aggregation of institutional sci-tech achievement data by institutional dimension, intelligent calculation and depiction of institutional academic profiles, and analysis of current layout and development directions. The system also supports real-time provision of researcher data, research funding project data, and published journal article data for the institution, as shown in Figure 6 [Figure 6: see original paper].

As the sci-tech literature big data system provides data support for an increasing number of application services, certain emerging issues require careful consideration and resolution in future construction.

First is the issue of sustainable development. We must carefully analyze all resource channel guarantees, identify possible sources and acquisition methods for various resource types, and find appropriate operational models. With limited funding, we must use mechanisms such as collaborative construction and sharing, and data services to incentivize participation and contribution from multi-source data providers. Simultaneously, we must continuously enhance the quality control capabilities of the sci-tech literature big data system. Due to multiple data sources with varying quality and standards, many hidden problems exist during data cleaning and fusion, affecting fusion accuracy, organizational standardization, and the effectiveness of various top-level intelligent knowledge service applications. Moving forward, we must strengthen the enrichment of normative database dimensions and levels, combined with new technologies and models, to effectively improve data completeness and quality.

References

[1] What are the application scenarios of AI now that it is so popular? [EB/OL]. [2020-11-16]. https://www.zhihu.com/question/282715644.

[2] The AI Hierarchy of Needs [EB/OL]. [2020-11-16]. https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007.

[3] AI Bigbull2050. The three driving forces of AI: algorithms, computing power, and data [EB/OL]. [2020-11-16]. http://blog.itpub.net/69946223/viewspace-2734390.

[4] Liu Lin. What changes will occur in AI basic data services under the new infrastructure boom? [EB/OL]. [2020-06-03]. https://www.leiphone.com/news/202006/WNW3OH7baaG0RBi5.

[5] Qian Li, Zhang Xiaolin, Wang Qian. Research on a research design fingerprint description framework based on sci-tech literature [J]. Journal of Academic Libraries, 2015, 33(1): 14-20.

[6] Ke Ping, Zou Jinhui. Library transformation in the post-knowledge service era [J]. Journal of Library Science in China, 2019, 45(1): 4-17.

[7] Research Intelligence [EB/OL]. [2020-10-06]. https://www.elsevier.com/research-intelligence.

[8] Digital Science [EB/OL]. [2020-10-06]. https://www.digital-science.com.

[9] Wizdom.ai [EB/OL]. [2020-10-06]. https://www.wizdom.ai/#about.

[10] arXiv [EB/OL]. [2020-10-06]. https://arxiv.org.

[11] PubMed [EB/OL]. [2020-10-06]. https://www.ncbi.nlm.nih.gov/pubmed.

[12] STKOS Sci-Tech Knowledge Organization System Sharing Service System [EB/OL]. [2020-10-06]. http://stkos.las.ac.cn/stkosservice/user/welcome.htm.

[13] iAuthor China Scientists Online [EB/OL]. [2020-10-08]. http://iauthor.cn/welcome/index.

[14] Chinese Academy of Sciences Institutional Repository Grid [EB/OL]. [2020-10-08]. http://www.irgrid.ac.cn.

[15] NSTL Unified Literature Metadata Standard 3.0 [EB/OL]. [2020-10-08]. http://spec.nstl.gov.cn/embed/home.htm.

[16] SciFire Knowledge Service Platform Based on Collective Intelligence [EB/OL]. [2020-10-08]. http://159.226.100.96/bi/bi.html.

[17] List of Chinese higher education institutions provided by the Ministry of Education of the People' s Republic of China [EB/OL]. [2020-10-08]. http://www.moe.gov.cn/srcsite/A03/moe_{634}/201706/t20170614_{306900}.html.

[18] DBpedia global university data [EB/OL]. [2020-10-08]. https://wiki.dbpedia.org/develop/datasets.

[19] China Scientists Online [EB/OL]. [2020-10-08]. https://iauthor.cn.

[20] Chinese Academy of Sciences Knowledge Service Platform, National Science Library, Chinese Academy of Sciences [EB/OL]. [2020-10-08]. https://www.las.ac.cn.

[21] Sci-Tech Big Data Knowledge Discovery Platform [EB/OL]. [2020-10-10]. https://scholareye.cn.

[22] Hui Research Personal Edition: Intelligent Portable Research Assistant [EB/OL]. [2020-10-10]. https://scholarin.cn.

[23] Chinese Academy of Sciences Literature and Information Center Data Observation Platform [EB/OL]. [2020-10-10]. http://kgview.las.ac.cn.

[24] Chinese Academy of Sciences Knowledge Service Platform Data Service Center [EB/OL]. [2020-10-10]. https://www.las.ac.cn/front/dataCenter/dataResources.

[25] Hui Research Institutional Edition: Institutional Knowledge Management and Analysis Service Platform [EB/OL]. [2020-10-10]. https://inst.scholarin.cn.

Author Contributions:

Wu Zhenxin: Responsible for the design and construction of the sci-tech literature big data system, paper writing, and final revision.
Qian Li: Participated in the system design, responsible for big data infrastructure platform construction.
Xie Jing: Participated in the system design, responsible for sci-tech big data management platform construction.
Chang Zhijun: Participated in the system design, responsible for big data infrastructure platform implementation.
Xu Liyuan: Aggregated and cleaned data, participated in establishing data cleaning rules, wrote and revised the initial draft.
Zhao Yan: Participated in the sci-tech literature big data system design.

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