Design and Implementation of a Research Data Management Platform for Humanities and Social Sciences from a Full Lifecycle Perspective: Post-print

Authors: Yao Zhanlei, Gu Jun, Xu Xin

Date: 2023-04-01T16:02:48+00:00

Abstract

[Purpose/Significance] Although China has issued policies in recent years that provide clear guidance and norms for the management, sharing, and utilization of scientific data, existing mainstream data management platform architectures focus on the scientific management of data, resulting in low efficiency in data sharing and utilization. This study systematically expands the data management functions of existing platforms and focuses on solving the challenges of scientific data sharing and utilization. [Method/Process] Based on extensive investigation and literature review, this study first clarifies the necessity and challenges in constructing research data management platforms for humanities and social sciences. Second, it systematically designs and elaborates the core functions and characteristics of such platform construction from a full lifecycle perspective. Furthermore, combined with platform examples, it details the key technical implementations of the core functions. [Results/Conclusions] With open interconnection as the foundation, development and utilization as the core, and self-service analysis as the distinctive feature, this study ultimately establishes a basic framework for a full lifecycle-oriented research data management platform. Based on this framework, it designs and implements a full lifecycle research data management platform for humanities and social sciences, which can provide a distinctive case and reference for relevant practice and research.

Full Text

Abstract

[Purpose/Significance] In recent years, although China has issued policies providing clear guidance and regulations on the management, sharing, and utilization of scientific data, existing mainstream data management platform archi-

tectures have focused primarily on the scientific management of data, resulting in low efficiency in data sharing and utilization. This study systematically expands the data management functions of existing platforms and focuses on solving the challenges of scientific data sharing and utilization. [**Method/Process**] Based on extensive investigation and literature review, this study first clarifies the necessity and challenges of constructing a humanities and social sciences research data management platform. Second, it systematically designs and explains the core functions and characteristics of such a platform from a full lifecycle perspective. Finally, combined with a platform instance, it details the key technical implementations of core functions. [**Result/Conclusion**] With open interconnection as the foundation, development and utilization as the core, and self-service analysis as the distinctive feature, this research establishes a basic framework for a full lifecycle-oriented research data management platform. Accordingly, a full lifecycle humanities and social sciences research data management platform was designed and implemented, which can provide a characteristic case and reference for related practices and research.

Keywords: data management; humanities and social sciences; development and utilization; platform construction
Classification Number: G252.5
DOI: 10.13266/j.issn.0252-3116.2021.07.003

Scientific data constitute a crucial foundational strategic resource for national scientific innovation and economic and social development. In today's big data era, scientific and technological innovation activities increasingly rely on the analysis, mining, and comprehensive utilization of scientific data. To this end, China has issued the "Administrative Measures for Scientific Data" (State Council Office Document No. 17 [2018]), emphasizing the need to "strengthen and standardize scientific data management, adapt to the development trend of big data, and actively promote the development, utilization, and open sharing of scientific data resources." However, although the value of research data sharing and reuse has become a consensus in academia, its implementation in practice remains unsatisfactory [1]. As a foundational resource, the construction of humanities and social sciences data resources in China is relatively lagging, and more ideas and inspiration are needed to enhance researchers' ability to conduct scientific research using interdisciplinary and cross-domain knowledge [3]. To promote data reuse by researchers, extensive studies and practices have been conducted from the perspectives of data managers and data reusers [1]. From the data manager perspective, this paper follows the "build well, manage well, use well" approach to further improve and refine existing humanities and social sciences research data management infrastructure, aiming to solve the challenges of research data sharing and utilization.

1 Related Research and Practice

1.1 Research Data Management

Research data broadly refers to the original and fundamental data generated in scientific research activities, which can help improve scientific reproducibility and credibility. Its management should address the entire lifecycle of research data [4] to efficiently carry out various data management activities. Currently, domestic and international researchers have conducted extensive studies and explorations on topics such as researcher needs surveys [5-7], data management lifecycles [8-10], data management services [11-13], data management policies [14-16], and data management education [17-19]. Research data management research and practice activities in the United States, United Kingdom, and Australia started earlier and have formed different development paths and solutions adapted to their respective national research cultures [20], with university libraries playing an increasingly prominent role and status [3]. Domestic research in China mostly introduces foreign research data management experiences [21] and conducts series of practical activities based on them. Particularly after 2011, the library and information science community and library industry began actively engaging in, tracking, and conducting research data management studies and practices, with typical cases including the Chinese Academy of Sciences' "Scientific Data Management and Sharing Cloud Service Platform" and Wuhan University Library's CALIS Phase III "University Scientific Data Management Mechanism and Platform Research." However, China's research data management research and practice remain in the exploratory stage, with a healthy open sharing culture and mechanism yet to be formed. There is insufficient comprehensive, holistic, and heuristic systematic theoretical research and process evaluation method innovation [18], and related research urgently needs strengthening.

1.2 Research Data Lifecycle Management

Research data differs from information resources in terms of "value aging" lifecycle patterns. Its lifecycle is closely linked to scientific research activities and is influenced by research methods, tools, and techniques [22]. It has a dual nature: managing the data's own lifecycle while reflecting the scientific activity lifecycle [9]. Although some scholars have explored data resource integration paths based on the scientific activity lifecycle—for example, Jing Runtian et al. [23] focused on research teams and analyzed characteristics of different team lifecycle stages, upon which Jia Yuwen et al. [24] established a resource integration model embedded in the research lifecycle—currently mainstream data management lifecycle models [25] are primarily data-oriented (see Table 1), and related research and practice activities are mostly based on these models. To effectively measure, evaluate, and continuously improve research data management practices and services, research data management capability maturity models [26] have been constructed for graded measurement.

Table 1 Several Typical Research Data Management Lifecycle Models

Data Lifecycle Model	Proposing Institution/Individual (Year)	Main Stages/Verbs
UK Data Archive	UK Data Archive (2004)	8 stages: conceptual research, data collection, data processing, data archiving, data publication, data discovery, data analysis, and data reuse
Research360	UK Data Archive Project Consortium (2014)	8 verbs: plan, collect, assure, describe, preserve, discover, integrate, analyze
DataONE	University of New Mexico Library, etc. (2009)	6 stages: data creation, data processing, data analysis, data preservation, data access, data reuse
UK Data Service	University of Essex (2007)	8 verbs: create, store, describe, identify, register, discover, obtain, develop
Australian National Data Service (2008)	2 stages: foundational stage (planning, peer review, experimentation, data processing/analysis/interpretation, final reporting) and idealized stage (evaluation/quality control, metadata/documentation, storage/archiving/preservation, intellectual property, embargo and access control)	

Data Lifecycle Model	Proposing Institution/Individual (Year)	Main Stages/Verbs
Structured Science Integration Infrastructure Project (2009)	6 functional entities: data collection, archival storage, data management, administration, preservation planning, and data access	
N. Beagrie et al. (2001)	6 stages: planning and design, collection and acquisition, interpretation and analysis, management and preservation, publication and publishing, mining and reuse	
University of Bath (2013)	6 stages: conceptualization, creation and receipt, evaluation and selection, long-term preservation and storage, access/use/reuse, transformation	

The lifecycle models shown in Table 1 primarily describe how to manage and control data effectively. Although they mention data reuse and development, they do not elaborate on these aspects in detail, and most are institution-oriented (for managers). Research data usage includes four aspects: verification, aggregation, mining, and reuse [27], which can further promote new academic discoveries and form new academic ecosystems. Therefore, optimizing data management lifecycle models is necessary.

1.3 Research Data Management Platforms and Tools

The provision of research data management services depends on platforms and tools to handle data management issues at various stages of the research data management lifecycle (see Figure 1 [Figure 1: see original paper]). Currently, research data management platforms and tools are flourishing and developing toward openness, integration, and standardization [28]. However, as shown in Figure 1, mainstream data management platforms and tools currently focus on data creation, processing, preservation, and access, with less emphasis on analysis and reuse.

Figure 1 Schematic diagram of data management platform tools and their distribution in typical data management lifecycles

(1) **Data Management Planning Tools**: These are formal documents providing summary descriptions of data management, including various stages during and after project completion [29]. Currently, the three most influential and widely used Data Management Plan (DMP) tools are DMPonline (https://dmponline.dcc.ac.uk), DMPTool (https://dmptool.org),

and DMPRoadmap (https://github.com/DMPRoadmap), all of which are open-source software.

(2) **Electronic Laboratory Notebooks**: These primarily record and store experimental data electronically, providing collaboration, templates, data collection, and analysis functions to optimize research workflows and process documentation. In 2017, the University of Minnesota Libraries conducted a special survey on electronic laboratory notebook usage at top U.S. research universities, showing that most are expensive. Currently, many electronic laboratory notebook software options exist, but none meets all researchers' needs due to research differences. Common examples include LabArchives (triable, with professional and educational versions), RSpace (with community and enterprise versions, where the community version is free for trial), and sciNote (with free, premium professional, and premium enterprise versions, and is open-source).

(3) **Active Data Storage Platforms**: During scientific research, researchers continuously generate data called "active data," where security protection (hardware damage, virus intrusion, accidental deletion, etc.) is crucial. With mature cloud computing technology, besides traditional multiple and off-site backups, data storage increasingly "moves to the cloud," such as using general public cloud storage (Google Drive, Baidu Netdisk, etc.), purchasing commercial services to build campus cloud storage, or using open-source software for self-built cloud storage (for high-security requirements).

(4) **Archive Data Management Platforms**: These are traditional research data management platforms for managing highly stable, important, and long-term research data, such as self-built ICPSR, open-source Dataverse and DSpace, and commercial Figshare platforms. Currently, many archive data management platforms are built and in use across numerous disciplines, with specialized data publishing platforms emerging, such as Nature Publishing Group's Scientific Data (2014), the Global Change Data Journal (Chinese and English) editorial board's Global Change Scientific Research Data Publishing System (2014), and the Library Journal editorial board's data management platform (2017). According to re3data.org, as of January 13, 2021, 3,581 platforms have been registered, with the United States ranking first with 1,100 platforms. Excluding international associations/organizations (249), the top ten countries have 2,762 platforms, while China has only 47—a significant gap compared to leading countries like the United States.

(5) **Persistent Identifier Systems**: These assign globally unique and persistent identifiers to data to facilitate citation, identification, location, and long-term preservation. Currently, three persistent identifier schemes are widely used in data management platforms: Handle (http://www.handle.net), DOI (http://www.doi.org), and ARK (https://n2t.net/e/ark_{ids}.html).

(6) **Data Retrieval Systems**: These support researchers in finding required data resources, divided into dataset retrieval systems (directly searching dataset metadata) and data repository retrieval systems (focusing on repository metadata). Mainstream dataset retrieval systems include DataCitation Index (commercial), DataCite Search, and Google Dataset Search, while data repository retrieval systems include re3data and FAIRsharing.

1.4 Current Status of Humanities and Social Sciences Research Data Management Platforms

Unlike natural sciences, humanities and social sciences focus on systematic understanding of human and social phenomena and their regularities, with social consciousness characteristics. Although data scale is small, semantic content is rich and diverse, with high reusability—humanities and social sciences data have longer usage cycles, can be reused by multiple research teams in the same direction, and can generate sustained value. However, compared to natural sciences, most original and derived humanities and social sciences research data in China stop at academic publication, with a genuine sharing and deep development environment yet to form. Scholars' enthusiasm for data sharing and utilization remains to be improved, and data resource construction is relatively lagging, mostly driven by major national scientific research projects. Due to practical difficulties such as "single-function data service platforms, low retrieval efficiency, lack of support for machine reading and raw download, and poor overall system usability," these platforms struggle to meet users' needs beyond project requirements [2].

As the scientific research paradigm transforms toward data-driven approaches, humanities and social sciences research data management platform construction has received widespread attention, with numerous cases emerging (see Table 2). Although China's related platform construction has achieved considerable success, the humanities and social sciences field is still in the initial exploration stage, with non-standard and insufficient datasets. Compared to foreign platforms, common issues include lack of open-source philosophy in software development, incomplete platform service functions, and absence of collaborative construction concepts in some platforms [31]. In recent years, various data competition activities such as the "National University Data-Driven Innovation Research Competition," "Huiyuan Sharing University Open Data Innovation Research Competition," and "Master Cup Data League" have provided new paths and attempts for secondary development and utilization of research data resources. However, related platforms' support capabilities for data analysis (computing power, toolkits, etc.) remain in the initial stage.

Table 2 Several Typical Humanities and Social Sciences Research Data Management Platforms

Platform	Dataset Count	Main Functions and Official Website
UK Data Archive (UKDA)	-	Dataset creation, submission, search, download; consulting services; discussion community. https://www.data-archive.ac.uk
Inter-university Consor-tium for Political and Social Research (ICPSR), University of Michigan	15,600+	Dataset creation, submission, search, download; data analysis; consulting services; news and events; discussion community. https://www.icpsr.umich.edu/icpsrweb
Harvard-MIT Data Center (HMDC), Dataverse	106,870+	Dataset creation, submission, search, download; online statistical analysis; research computing support; desktop services and hosting services. https://dataverse.harvard.edu
Peking University Open Research Data Platform, Dataverse	-	Dataset creation, submission, search, download; online statistical analysis. https://opendata.pku.edu.cn
Fudan University Social Science Data Sharing Platform, Dataverse	-	Dataset creation, submission, search, download; data classification statistics. http://dvn.fudan.edu.cn
Renmin University China Academic Survey Data Archive	-	Dataset creation, submission, search, download; data analysis report sharing. http://www.cnsda.org

Note: Dataset counts are as of January 13, 2021, without disciplinary distinction.

It should be noted that existing research data management platforms mostly treat data as a type of information resource. Although data analysis, mining, and utilization are considered in top-level design, the application focus is not on these aspects. Currently, as the value of data resources becomes increasingly important and secondary development and utilization needs grow stronger, existing research data management platforms require upgrading and transformation.

2 Design of Humanities and Social Sciences Research Data Management Platform

Current humanities and social sciences research data management platform construction in China is mostly led by universities, with unique but non-standard data content and insufficient secondary development functions. This paper focuses on secondary development and utilization of data resources, following the "build well, manage well, use well" approach. While accommodating mainstream data management platform basic functions, it centers on multi-channel data resource collection and standardized management, self-service analysis, and development and utilization to form a distinctive construction scheme, as shown in Figure 2 [Figure 2: see original paper].

Figure 2 Basic framework of a full lifecycle-oriented humanities and social sciences research data management platform

Under this framework, platform construction should emphasize three directions:

(1) **Open and Interconnected Data Sharing Mechanism**: Emphasize data sharing capabilities and security protection of the platform. Under legal compliance and data privacy protection terms, achieve scientific and smooth flow and effective utilization of data resources. Therefore, platform construction should accommodate data access (collection, exchange, etc.), standardization (metadata management, backup, etc.), and utilization (aggregation, analysis, download, etc.) across different scenarios to ensure scientific flow of data resources inside and outside the platform. Consider introducing blockchain technology for data rights confirmation and copyright protection, and sandbox technology to ensure "non-landing" development and utilization.

(2) **Self-Service Analysis Software Access Standards**: Emphasize platform analytical capabilities to empower the platform, enhance usability, and transform it from "emphasizing collection" to "collection and utilization." Therefore, beyond common resource statistics and preview functions, the platform must solve the problem of utilizing complex data resources, thereby promoting researchers' data sharing. To ensure analytical flexibility and scalability, focus on solving the free access of third-party analysis software rather than building proprietary analysis environments,

emphasizing the development of platform data analysis interface specifications to form an open interconnection mechanism for software tools and meet diverse analytical needs.

(3) **Data Resource Development and Utilization Models**: Emphasize secondary development and utilization activities of platform data resources, conducting value-added services under legal compliance. The platform should support dataset evaluation and academic promotion, data publishing, and other activities, with capabilities for aggregating thematically related multi-source data, mining potential research teams, identifying research data tracking and academic integrity, and identifying high-quality thematic datasets (based on data quality, utilization rate, etc.).

2.1 Data Collection and Metadata Management

Addressing the diversification of data collection and metadata description standardization, the platform should meet collection needs for researchers' self-stored data, thematic databases, and public open data to adapt to the widespread distribution of research data and support unified data description.

(1) **Diversified Data Collection Strategies**: Data sources in research data management platforms are divided into researcher submission (thematic data) and platform active collection (public open data/commissioned processing data). Regardless of method, data must be collected and integrated into the platform efficiently, stably, and accurately. Beyond traditional file attachment uploads, data often exists in large quantities and variable forms such as databases and web pages. Therefore, the platform should innovate collection strategies by designing JDBC/ODBC interfaces, crawler/protocol collection, and API calls to adapt to this reality, while considering data collection volume and frequency, data presentation methods (statistical charts, data lists, etc.), and providing good user experience.

(2) **Metadata Description Standards**: Establish complete metadata description standards for research data characteristics to achieve standardized description of research data (including three major attributes: basic attributes, characteristic attributes, and value attributes, as shown in Figure 3 [Figure 3: see original paper]), including a field description standardization system and metadata field mapping. The field description standardization system specifies platform data publishing field description rules, guiding data publishers to improve other users' understanding of data and enhance data sharing capabilities.

Figure 3 Schematic diagram of standardized description of research data

2.2 Data Sharing and Copyright Protection

Addressing scientific management and sharing of research data, focusing on data catalogs and data files to promote organic flow of data inside and outside the platform.

(1) **Data Catalog Management and Sharing**: Data catalogs are extracted from research data metadata. To maximize data sharing and dissemination, the platform should support sharing through RSS subscription, one-click sharing, and APIs to actively push and increase data resource exposure. Follow mainstream data interoperability protocols (such as OAI-PMH, SRW/U, SDARTS protocols) to meet cross-platform data resource integration and sharing needs.

(2) **Data File Management and Sharing**: Focus on research data itself. For multi-source heterogeneous data types (.txt, .xlsx, .csv, .sql, etc.), do not stop at data submission. Design and implement a multi-type file fusion storage mechanism (see Figure 4 [Figure 4: see original paper]) to achieve deep content-level integration of data files, laying the foundation for secondary development and utilization and thematic multi-source data aggregation. To fully protect data privacy, data file development, utilization, and sharing activities must obtain data owners' consent. Before formal data publication, the platform should prompt owners to set corresponding permissions. When data file permissions are set as "restricted," the platform must notify owners via email or SMS to obtain authorization for subsequent development, utilization, and sharing activities involving the data file.

Figure 4 A multi-type file fusion storage mechanism for research data management

Blockchain's characteristics of decentralization, openness, autonomy, information immutability, and anonymity offer natural advantages in autonomy, tracking, and traceability during humanities and social sciences data sharing [32]. Considering that public blockchains are open to all public, with data management 不受 any individual or organizational control, and that humanities and social sciences data have sustainable use and slower update rates, this platform recommends using private or consortium chains for data rights confirmation and copyright protection. Compared to public chains, transaction costs between nodes are lower, as shown in Figure 5 [Figure 5: see original paper].

Figure 5 Schematic diagram of inter-node transaction steps in Hyperledger Fabric, a typical consortium chain framework

2.3 Online Self-Service Analysis for Research Data

Addressing platform research data usage for researchers, assisting in-depth internal data observation, mining, and analysis, primarily reflected in compatibility with mainstream analysis software tools.

(1) **General Data Exploration**: Should meet researchers' needs for shallow data overview of interesting data files to intuitively grasp data morphology, quality, and content, including field descriptions, data instances, statistical reports, etc. To further assist research and observation, statistical report design should adopt the framework shown in Figure 6 [Figure 6: see original paper].

Figure 6 Key points of statistical report design

(2) **Professional Mining and Analysis**: The richness of platform analysis tools can both support researchers' professional mining analysis activities around interesting data files and stimulate researchers to share research data and enhance platform activity. Figure 6 solves multi-source data expression issues at the data level. Here, platform internal data migration and tool access issues need resolution. Platform internal data migration concerns whether platform data should flow to tools during mining analysis, focusing on execution efficiency—data flow to tools itself is time-consuming, which is crucial with massive data. Tool access methods are divided into hard access (deeply integrated with the platform) and soft access (through API calls). Hard access doesn't involve data migration, while soft access involves data migration but supports free third-party tool access, as shown in Figure 7 [Figure 7: see original paper].

Figure 7 Mechanism for third-party tool access to the platform

2.4 Secondary Development and Utilization of Data Resources

For management, promoting secondary development and utilization activities of platform research data and conducting targeted value-added services, the platform must provide technical support for these services.

(1) **Thematic Data Aggregation**: Primarily relies on metadata descriptions and data item characteristics in data files to extract tags, measure thematic similarity, and conduct multi-dimensional aggregation of data from different projects or researchers, aiming to overcome potential biases from single-source data. Although the platform has unified specifications at the data level for single data files, it must also provide mechanisms supporting researchers' multi-source data splicing and association, while recording manual data splicing behaviors and using machine learning to continuously enhance intelligent data aggregation capabilities.

(2) **Data Mining and Pattern Discovery**: Includes data usage and content mining. Data usage focuses on research data's value and utilization efficiency, establishing measurement and evaluation models (data integrity, data citation, etc.) similar to papers to support researchers in finding high-value data resources. Content mining focuses on analysis, comparison, and knowledge discovery from multi-domain data and multiple datasets, such as identifying implicit research teams based on similar datasets and iden-

tifying academic integrity based on data quality (or consistency). The platform should support these activities.

3 Key Technologies for Humanities and Social Sciences Research Data Management Platform

Currently, many platform prototypes such as Dataverse, DSpace, EPrints, Fedora, and Nesstar have emerged. Most are based on digital asset systems, focusing on digital asset preservation and management, with weak support for data analysis and visualization, but they provide relatively complete solutions for institutional data management platform construction, helping avoid cumbersome basic function development and focus on data resource development and utilization function development. Notably, building data management platforms based on open-source software through secondary development is also common in domestic and foreign universities [33].

The authors used DSpace as the platform prototype for personalized design and secondary development to complete basic function development of the humanities and social sciences research data management platform. DSpace was initially jointly developed by MIT Libraries and HP Labs and launched in October 2002. As a digital resource storage system targeting content management and publication, it can collect, store, index, and publish digital resources in various formats, with a complete user interface, strong customizability, good scalability, and secondary development support. It is currently used online by over 300 institutions worldwide with numerous users and successful cases.

The platform developed in this paper (with Yao Zhanlei responsible for specific design, Gu Jun for specific development and implementation, accessible at http://222.204.246.126/rdmp/) is based on DSpace 4.9 secondary development, using the springboot2 development framework, openjdk1.8, postgresql9.6, and solr7.3. During development, basic data items and parameters were restructured to meet front-end page content customization, data requirement management (see Figure 8a [Figure 8: see original paper]), data association with research outcomes (see Figure 8b [Figure 8: see original paper]), and data application authorization management (see Figure 9 [Figure 9: see original paper]).

Figure 8 Operation diagrams of the data management platform after DSpace secondary development (partial)

Figure 9 Schematic diagram of data management platform backend application authorization management

3.1 Open and Interconnected Data Sharing

Data sharing includes platform-to-platform data sharing and data file sharing among different users on each platform.

Platform-to-platform data sharing mostly concerns data catalog information

exchange, a platform-level application for maximizing data sharing and dissemination. The platform uses the OAI-PMH metadata harvesting protocol for data sharing between two independent platforms, provided they have established trust relationships. The platform supports the OAI-PMH protocol, allowing other data platforms to harvest metadata, as shown in Figure 10 [Figure 10: see original paper].

Figure 10 Platform metadata harvesting parameter configuration

After data is harvested to other platforms, user-to-user data sharing can be achieved on that platform. To protect data providers' legitimate rights and prevent data abuse, the platform requires data providers to authorize data use when sharing among users. Currently, the platform uses email and system methods for authorization. When users request data use and download, the system automatically sends an authorization request email to the provider's mailbox and synchronizes it in the provider's system management backend. Only after the provider authorizes the data does the user obtain usage permission, as shown in Figure 9 [Figure 9: see original paper].

Simultaneously, the platform improved the database supported by Hyperledger Fabric, replacing traditional super ledger key-value data storage with relational database storage (see Figure 11 [Figure 11: see original paper]) to enhance on-chain data query and processing capabilities. Accordingly, the platform reserved blockchain infrastructure interfaces, enabling seamless integration with blockchain infrastructure while ensuring stable platform operation.

Figure 11 Core code for on-chain data structure transformation

3.2 Self-Service Data Exploration

To enhance data analysis and visualization functions in existing data management platforms and support researchers' data exploration activities, the platform draws on commercial reporting concepts to provide multi-dimensional, dynamic, interactive reporting and visualization functions under a unified analysis window (see Figure 12 [Figure 12: see original paper]), supporting flexible and personalized data exploration for structured data files.

Figure 12 Schematic diagram of general report analysis

To give report analysis universality, the platform currently focuses on general mining and analysis of data file items and values, including: (1) generic analysis of database table structures such as data integrity, missing values and outliers, data item grouping, and time series; and (2) basic text analysis in research activities such as keyword clouds and entity extraction. Such functions enhance platform visualization capabilities while providing researchers with simple, one-stop data analysis services, thereby increasing researchers' willingness to use the platform and revitalizing and expanding data resources.

The general report analysis is an intelligent analysis module that automatically

scans and identifies relevant fields and content in data files, generating corresponding visualization charts based on backend-defined data analysis modules. Data volume describes the total records in the data file. If incomplete records are found, the field is extracted as a missing field. The system also compares values of all field data types to identify maximum and minimum fields, marking them as outlier fields. During field scanning, the system uses regular expressions to determine field types. Fields not matching numeric or time field criteria are identified as text fields. For text fields, the system uses the "jieba" word segmentation tool for text processing and combines the TextRank algorithm to extract tags from field content, ultimately drawing thematic cloud maps for the data file to facilitate users' deep mining of dataset themes.

Meanwhile, the platform has been deeply integrated with the Transwarp big data platform (core code shown in Figure 13 [Figure 13: see original paper]), leveraging third-party platforms' rich analysis tools to meet scholars' needs for more professional analysis and mining activities, as shown in Figure 14 [Figure 14: see original paper].

Figure 13 Core code for integrating with the Transwarp big data platform

Figure 14 Schematic diagram of accessing third-party analysis platforms

3.3 Value-Added Data Development and Utilization

Humanities and social sciences data have high reusability and continuously generate value. Data in the same research direction can be reused by multiple R&D teams, making the value of multi-source data splicing and aggregation prominent. The platform has established a multi-source data aggregation platform, focusing on internal structured data resources, reading characteristics of data resources to be aggregated, and providing scholars with an online tool for free extraction, combination, and flexible content editing (see Figure 15 [Figure 15: see original paper]). It also supports accessing scholars' own data (excel, sql, and other file formats).

Figure 15 Schematic diagram of multi-source data aggregation platform

For high-value data resources, with data owners' authorization, the platform can also conduct special value-added activities, such as processing and compiling into featured thematic datasets for publication and distribution by theme, discipline, or event. Utilizing data usage and evaluation on the platform provides richer support for data publishing. Combined with data resources' research attributes and online tools, the platform can build communication platforms for new researchers in disciplines to help them reproduce classic studies, understand research paradigms, and quickly grasp relevant research paths. These new attempts will provide new paradigms and paths for further promoting research data resource development and utilization and become the focus of subsequent platform optimization and upgrading.

References

[1] Sun Yuwei, Cheng Ying, Xie Juan. Researchers' data reuse behavior research: Systematic review and meta-synthesis[J]. Journal of Library Science in China, 2019(5): 110-130.

[2] Wang Xiaoguang. Strengthen humanities and social sciences data resource construction and management[N]. Guangming Daily, 2018-07-05(11).

[3] Zhang Yanan, Huang Jingli, Wang Gang. Construction method of three-dimensional and precise portrait of researchers' scientific research behavior considering global and local information[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(10): 1012-1021.

[4] Qian Jinlin, Liu Guifeng. Review of foreign scientific research data management research[J]. Information Studies: Theory & Application, 2017, 40(10): 130-134.

[5] Xing Wenming, Yang Ling. Investigation on current status of scientific research data management in China's scientific research institutions[J]. Digital Library Forum, 2018(12): 27-33.

[6] Hu Yongsheng, Liu Ying. Analysis of university scientific data management needs based on user surveys[J]. Library and Information Service, 2013, 57(6): 18-22.

[7] Wang Dandan. Research on scientific data management service needs identification methods[J]. Journal of Academic Libraries, 2013(6): 10-17.

[8] BALL A. Review of Data Management Lifecycle Models[EB/OL]. [2019-09-26]. http://opus.bath.ac.uk/28587/.

[9] Qian Peng. Analysis of the duality of information lifecycle management: Taking scientific data management as an example[J]. Information Studies: Theory & Application, 2013, 36(3): 11-14.

[10] DARLINGTON M, BALL A. A Research Data Management Plan for Engineering Research[EB/OL]. [2019-09-30]. http://opus.bath.ac.uk/30104/.

[11] RICE R, HAYWOOD J. Research data management initiatives at University of Edinburgh[J]. International journal of digital curation, 2011, 6(2): 232-244.

[12] Chen Daqing. Research on implementation framework of foreign university data management services[J]. Journal of Academic Libraries, 2013(6): 10-17.

[13] Zhang Peifeng, Zhang Lianfen. Innovation of library scientific data management services under global scientific research paradigm transformation: From the perspective of data management lifecycle[J]. Library Theory and Practice, 2019(5): 39-48.

[14] Zhou Yuqin, Xing Wenming. Research on China's scientific research data

management and sharing policy system[J]. Chinese Journal of Medical Library and Information Science, 2018, 27(8): 1-7.

[15] He Qingfang. Investigation and analysis of foreign scientific data management policies[J]. Shanghai University Library and Information Work Research, 2016, 26(2): 9-13.

[16] Jiang Xin. Analysis of current status and future research trends of scientific data open policy research[J]. Journal of Modern Information, 2016(2): 167-171.

[17] Xing Wenming, Tang Yajing, Qin Shun. Interpretation and enlightenment of foreign educational institution scientific research data management policy outlines[J]. Digital Library Forum, 2019(5): 9-16.

[18] E Lijun. Scientific research data management education in foreign university libraries[J]. Information and Documentation Services, 2014(1): 101-105.

[19] Zhang Yanmei. Library scientific data management research from the perspective of user data literacy education[J]. Library and Information, 2015(4): 139-141, 109.

[20] Cui Yuhong, Li Weimian. Review of research data management progress[J]. Library Journal, 2017(1): 12-19.

[21] Chai Huiming, Zhang Libin, Zhao Yajie. Review of domestic library scientific data research[J]. Library and Information Service, 2019, 63(7): 116-126.

[22] Ding Ning, Ma Haoqin. Comparative study and reference of foreign university scientific data lifecycle management models[J]. Library and Information Service, 2013, 57(6): 18-22.

[23] Jing Runtian, Wang Rui, Zhou Jiagui. Research on lifecycle stage characteristics of scientific research teams: Multi-case comparative study[J]. Science of Science and Management of S.& T., 2011(4): 173-180.

[24] Jia Yuwen, Li Chaoqun. Research on data resource integration model embedded in research lifecycle[J]. Journal of Library Science, 2019(2): 51-55.

[25] Li Weimian. Research on research data management service evaluation based on lifecycle theory[D]. Beijing: Beijing Institute of Technology, 2016.

[26] CROWSTON K, QIN J. A capability maturity model for scientific data management: Evidence from the literature[J]. Proceedings of the American Society for Information Science and Technology, 2011, 48(1): 1-9.

[27] MARCHIONINI G, Yang Guancan, Lu Kun. Research data management: Ensuring data quality, promoting new scientific research in iSchools[J]. Library and Information Knowledge, 2013(4): 4-9.

[28] Cui Haiyuan. Research Data Management and Service Guide[M]. Beijing: Ocean Press, 2019.

[29] Data Management General Guidance[EB/OL]. [2020-06-25]. https://dmptool.org/general_{guidance}/.

[30] SAYRE F D, BAKKER C J, JOHNSTON L R, et al. Where in academia are ELNs? Support for electronic lab notebooks at top American research universities[C]// Poster presented at the Association of College & Research Libraries Conference. Baltimore: ACRL, 2017.

[31] Cui Xu, Zhao Ximei, Wang Zheng, et al. Analysis of achievements, deficiencies, countermeasures, and trends of China's scientific data management platform construction: From a domestic and foreign comparative perspective[J]. Library and Information Service, 2019, 63(9): 21-30.

[32] Gu Jun, Xu Xin. Design and implementation of humanities and social sciences data sharing model: Taking consortium chain technology as an example[J]. Journal of the China Society for Scientific and Technical Information, 2019(4): 354-367.

[33] Hong Zhengguo, Xiang Ying. Building university scientific data management platform based on DSpace: Taking scorpion species and toxin database as an example[J]. Library and Information Service, 2013, 57(6): 39-42, 84.

Author Contributions: Yao Zhanlei: Platform design and manuscript writing; Gu Jun: Platform development and manuscript revision; Xu Xin: Overall design and manuscript finalization.

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