Construction of a Thematic Intelligence Data Management and Intelligent Analysis Platform Postprint

Authors: Yu Qianqian, Qian Li, Cheng Bing, Chang Zhijun, Wang Huili, Jin Xi

Date: 2023-04-01T16:16:05+00:00

Abstract

[Purpose/Significance] To address the demands for thematic intelligence services across multi-disciplinary domains and for diverse user types, a thematic intelligence data management and intelligent analysis platform is established. This platform realizes the proceduralization and intellectualization of thematic intelligence analysis, while simultaneously managing the process data of thematic intelligence analysis that incorporates expert wisdom, enriching service models, and enhancing the response speed to service demands. [Method/Process] Based on investigations of existing relevant research and practical analyses, the platform design concepts and construction framework are proposed, and the platform's main functions and key technologies are analyzed. [Results/Conclusion] The thematic intelligence data management and intelligent analysis platform has been completed. The platform integrates multi-source and multi-type data, establishing a seamless service chain from data to analysis. It embeds various intelligence analysis methods and deep learning algorithms, realizing multi-dimensional and multi-level analysis services. It can manage both the analysis process and historically accumulated data from intelligence analysts, achieving data sharing and reuse.

Full Text

Abstract

[**Purpose/Significance**] To meet the subject information service needs of multi-disciplinary fields and multi-type users, we constructed a subject information data management and intelligent analysis platform. The platform aims to achieve process-oriented and intelligent subject information analysis while

managing the process data of subject information analysis that incorporates expert wisdom, thereby enriching service models and improving service response speed. [**Method/Process**] Based on investigating existing relevant research and practical analysis, we proposed the platform's design concept and construction framework, and analyzed its main functions and key technologies. [**Result/Conclusion**] The subject information data management and intelligent analysis platform has been completed. It integrates multi-source and multi-type data, opens up the service chain from data to analysis, embeds various information analysis methods and deep learning algorithms, and realizes multi-dimensional and multi-level analysis services. The platform can manage both analysis process data and historically accumulated data from information analysts, enabling data sharing and reuse.

Keywords: subject information; data management; intelligent analysis; information analysis
Classification Number: G250.2
DOI: 10.13266/j.issn.0252-3116.2020.24.011

Introduction

In the era of big data information analysis and knowledge services, subject information services are undergoing disruptive transformation. The acquisition of multi-source heterogeneous data and the development of artificial intelligence technologies have brought new opportunities for subject information analysis [1]. The National Science and Technology Library (NSTL) has provided subject information services for national strategic needs for over a decade, accumulating vast amounts of subject information analysis process data. However, this data has remained in a decentralized, self-storage state. How to uniformly manage this data that incorporates the wisdom of scientific and technical information experts, and establish a foundation for rapid reproduction of subject information analysis processes, professional information sharing, and provision of new data services, is a question worth considering. NSTL and the National Science Library, Chinese Academy of Sciences (hereinafter referred to as "the Library") have acquired and integrated multi-source heterogeneous resources through various means [2-3]. How to fully exploit the value of these aggregated scientific and technological big data resources, compensate for the shortcomings of manual data source selection, data collection, data loading, and data analysis, and establish a rapid response mechanism for subject information analysis based on multi-source data computing, is another important consideration.

Related Research and Practice Analysis

2.1 Research on Subject Information Analysis Tools

Subject information analysis tools can be divided into platform-based tools and software-based tools based on whether they can provide online services. Platform-based tools include dedicated information analysis service platforms

(such as InCites [4], SciVal [5], Incopat [6], wisdomAI [7]) and scientific literature retrieval platforms with added analysis and evaluation functions (such as Dimensions [8], Web of Science [9], CNKI [10], and Wanfang Data Knowledge Service Platform [11]), with the latter transitioning from knowledge discovery to knowledge evaluation. Software-based tools include CiteSpace from Drexel University [12], VOSviewer from Leiden University [13], Sci2 from Indiana University [14], the open-source tool Gephi [15], Derwent Data Analyzer from Clarivate [16], and BibExcel from Umeå University [17].

Liu et al. [18] noted that the InCites and SciVal platforms contain numerous evaluation indicators capable of handling most scientific research impact analysis and evaluation tasks. Xu et al. [19] identified intelligent semantic retrieval, integrated and flexible data processing, comprehensive analysis perspectives, and intelligent automatic report generation as the main development trends of patent information analysis tools like Incopat. C. Herzog et al. [20] pointed out that Dimensions integrates multi-type data (publications, patents, funding projects, policies, clinical trials) and different dimensions of analysis (trend analysis, researcher analysis, funding analysis, institutional analysis, comparative analysis) into a single platform, expecting integration to promote innovation. Taylor & Francis Group utilizes big data analysis and machine learning technologies to develop wisdomAI [7], covering multi-type data including publications, patents, funding projects, institutions, and authors, providing researchers and research institutions with in-depth analysis services across the entire value chain. Yu et al. [21] found that CiteSpace, VOSviewer, DDA, and BibExcel have been frequently applied in knowledge graph research. Yang et al. [22] discovered that Sci2 is suitable for deduplication of large datasets and has strong editable network output capabilities. Deng et al. [23] considered Gephi more appropriate for processing dynamic large data with powerful visualization functions.

Existing research and platform practices show that platform-based tools are developing toward multi-sourcing, intelligence, and fine-grained analysis. The aggregation and fusion of multi-source heterogeneous data have become a new data infrastructure for information analysis, and knowledge mining using artificial intelligence technologies has become a new growth point. Software-based tools have distinctive features, offering much to learn from in terms of data cleaning and visualization analysis. However, completing a report from data acquisition to data analysis often requires switching between multiple software tools [24], typically preventing one-stop operations and lacking analysis process data management functionality. Studies [25-26] have noted that while there are many foreign information analysis tools, some products have issues such as high prices, export restrictions, or intellectual property barriers. Domestic tool development is insufficient, playing a limited role in information research, and requiring increased R&D investment. Therefore, building information analysis tools with independent intellectual property rights is essential.

2.2 Research on Data Management Tools

In terms of data management tools, the most prominent are scientific data management platforms and practices. Domestic and international scientific data management platform construction has developed rapidly, including Harvard Dataverse [27], Dryad Data Repository [28], Australian National Data Service (ANDS) [29], Chinese Academy of Sciences Data Cloud [30], Peking University Open Research Data Platform [31], and Wuhan University Scientific Data Management Platform [32].

Cui et al. [33] believe that the core service functions of data management platforms include data management planning, data creation, data storage, data acquisition, data analysis, and data sharing, with Dataverse and ANDS possessing all these functions. Wei et al. [34] consider domestic scientific data management platforms to be data-dominated, primarily storing and managing scientific data already generated by users, such as the Wuhan University Scientific Data Management Platform and the Chinese Academy of Sciences Data Cloud. Zhu et al. [35] found through comparison that both Dataverse and Dryad are multidisciplinary, but the former focuses on social sciences with metadata schemes based on the DDI metadata standard, while the latter focuses on biological and ecological sciences with metadata schemes following the DC metadata standard.

Based on existing research and platform-stored data, different data management platforms have different functional characteristics, disciplinary focuses, and metadata schemes. Some platforms aim to manage and preserve institutional scientific data, such as some domestic platforms. Others aim to collect and manage scientific data from different social institutions, such as Dryad and ANDS. Currently, Wuhan University Scientific Data Management Platform's quantitative analysis research dataset stores 5 information analysis-related data items, with description fields including title, author, date, related description, URI, and dataset affiliation, with attachments being statistical analysis datasets and analysis reports. Peking University Open Research Data Platform also contains information analysis-related data, but storage is scattered across different data spaces and datasets. Data description fields include title, author, contact person, submitter, submission date, description, and discipline, primarily storing analysis datasets. Overall, information analysis process data is gradually gaining attention, but the level of attention is still far from sufficient.

2.3 Analysis of Subject Information Service Practice

We conducted interviews with eight information analysts from NSTL member units (including the National Science Library, Chinese Academy of Sciences, Institute of Scientific and Technical Information of China, Institute of Medical Information, Chinese Academy of Medical Sciences, and China National Chemical Information Center) to understand current pain points in subject information services and analyst needs. The investigation revealed shortcomings and pain points in subject knowledge organization system construction, data acquisition,

data cleaning, data analysis, and data management.

In subject knowledge organization system construction, analysts primarily rely on manual collection and organization of materials, depending on expert guidance to form systems, lacking automated auxiliary tools. In data acquisition, obtaining multi-source data (funding projects, policy data, and other non-traditional literature) has become a trend, but non-traditional literature data sources are scattered, requiring manual searching across different websites. Bulk data acquisition is difficult; for example, Web of Science limits single downloads to 500 records and maximum queries to 100,000 records [36]. When data volume is large, download time and manpower consumption are significant. In data cleaning, although relevant tools provide some assistance, processing capabilities are limited. Cleaning methods mainly rely on controlled vocabularies and rules, while vocabularies accumulated by analysts remain in self-storage, self-management, and self-use states. In data analysis, current analysis tools struggle with large data volumes, typically operating slowly with over 50,000 records and easily freezing with over 100,000 records [37]. In data management, subject information analysis process files usually remain with project teams or individuals, lacking data management standards and platforms, making data sharing difficult.

Platform Design and Implementation

3.1 Platform Design Ideas

Subject information research is an information research work targeting specific users and specific needs [38]. Due to the strong dynamic and personalized nature of information problems and tasks, it is difficult to produce a universal information analysis system [39]. Drawing on existing research and practice, we propose the design ideas for the subject information data management and intelligent analysis platform: (1) Integrate multi-source and multi-type datasets such as journal articles, conference papers, patents, and funding projects into one platform, making full use of aggregated scientific and technological big data resources to support multi-source data acquisition; (2) Establish a human-machine combined data acquisition and cleaning approach, using relevant tools and algorithms to assist analysts in building knowledge organization systems, search strategies, and automated data cleaning; (3) Utilize big data technologies to improve analysis speed for large data volumes; (4) Design multi-dimensional and multi-level analysis modes, integrating various information analysis methods and deep learning algorithms to intelligently generate and export reports; (5) Uniformly store and manage high-value intermediate analysis results data from online or offline activities, achieving platform-based management, reusable utilization, and accumulable knowledge of subject information analysis data.

3.2 Platform Overall Architecture

Based on the platform design ideas, we determined the overall architecture of the subject information data management and intelligent analysis platform (see [Figure 1: see original paper]). This architecture includes five layers: big data infrastructure, big data resource system, subject data acquisition and cleaning specifications, subject information analysis computing models, and data management and analysis services.

3.2.1 Big Data Infrastructure

The National Science Library, Chinese Academy of Sciences has built a scientific and technological big data basic platform based on the open-source Apache Hadoop ecosystem to aggregate and fuse massive scientific and technological resources. We use the multi-source aggregated resources from this platform as the basic data source for the subject information data management and intelligent analysis platform. Elasticsearch is a distributed, scalable, high-real-time search and data analysis engine based on Lucene, used in the platform to store basic data and data retrieved or imported by subjects, and to support search result display and information analysis. Redis is a high-performance key-value database used to cache user access data and support platform performance optimization. MySQL is used to save subject information analysis process data.

3.2.2 Big Data Resource System

The platform possesses a multi-source and multi-type data resource system including "journal articles + conference papers + patent data + funding projects + policy data + enterprise data + capital support + market data." Among them, journal articles, conference papers, and patent data are resources obtained through negotiation, exchange, and purchase with domestic and foreign publishers and relevant information institutions by NSTL and the Library. The volume of journal articles and conference papers reaches over 110 million records, and patent data exceeds 80 million records. Funding project data includes self-collected funding projects from more than 10 countries, including the U.S. National Science Foundation (NSF) and China National Natural Science Foundation (NSFC), totaling over 5.2 million records. Policy data includes stored domestic relevant policy information, totaling over 260,000 records. Enterprise data, capital support, and market data provide storage support, with currently relatively small data volumes obtained.

3.2.3 Subject Data Acquisition and Cleaning Specifications

The platform establishes a human-machine combined data acquisition and cleaning specification approach, assisting users in sorting out and building authoritative and comprehensive data resources oriented toward subject research fields and directions. The platform supports unified description, representation, and storage management of multi-type data, parsing, integrating, deduplicating, fusing, cleaning, and standardizing subject data obtained through search, import, and knowledge topic screening. Based on platform automated data processing, users can manually edit and process retrieved datasets and research entities

such as countries/regions, institutions, personnel, and keywords, and set rules to optimize platform automated processing effects.

3.2.4 Subject Information Analysis Computing Models

The platform embeds various algorithm models, including research entity statistical analysis models, co-occurrence network analysis models, text mining visualization models, neural network semantic annotation models, large sample training models, and scientific and technological competitiveness evaluation models. Through data, algorithm, and computing-driven intelligent analysis, the platform achieves rapid production and delivery of information analysis reports combining intelligent computation of subject information data with expert wisdom.

3.2.5 Data Management and Analysis Services

The platform provides management functions for subject information analysis process data and local data, offering various analysis services including quantitative analysis, content analysis, and competitiveness evaluation analysis. Different analysis services correspond to different data resource types and employ different analysis computing models. The platform supports filtering of analysis dimensions for quantitative analysis and content analysis, intelligent generation and export of analysis reports, and front-end page publishing and display of analysis dimensions for quantitative analysis, content analysis, and competitiveness evaluation analysis.

3.3 Main Functions and Key Technologies

The subject information data management and intelligent analysis platform includes four main functional modules: professional information analysis, rapid analysis, competitiveness evaluation analysis, and data management. Rapid analysis draws on Web of Science, CNKI, and other platforms that add quantitative analysis functions based on literature retrieval, providing services to platform-approved users for quick understanding of field overviews. Professional information analysis and data management serve information analysts. Competitiveness evaluation analysis serves specific information analysts.

3.3.1 Professional Information Analysis

Based on investigation of analysts' workflows, we divided platform professional information analysis functions into five steps: creating subjects, subject knowledge organization, subject data aggregation, subject data cleaning and standardization, and subject information analysis (see [Figure 2: see original paper]).

(1) Creating Subjects

In this step, users can browse the list of created subjects, viewing subject names, data volume, data time range, data type, subject creation time, and status, and create new subjects as needed (see [Figure 3: see original paper]).

(2) Subject Knowledge Organization

In this step, the platform provides functions for creating, importing, editing,

and saving subject knowledge organization systems. Users can build hierarchical subject knowledge organization systems or import existing systems. The platform provides functions for adding, deleting, and modifying nodes, and supports building search strategies based on the knowledge organization system.

(3) Subject Data Aggregation
The platform supports data acquisition through basic search discovery, professional search strategies, or local import, and supports obtaining data through selecting knowledge topics (nodes in the subject knowledge organization system, defaulting to selecting nodes, synonyms, and hyponyms) to automatically generate search strategies (see [Figure 5: see original paper]). Search results are sorted by relevance and time, with faceted display of obtained data from multiple perspectives.

(4) Subject Data Cleaning and Standardization
Data cleaning and standardization is a crucial step in information analysis work and a prerequisite for ensuring accurate and reliable analysis results. In this step, the platform cleans and standardizes datasets through deduplication, searching, sorting, and deletion. The platform allows users to export data following fair use principles, with 5,000 records per export. For automated cleaning and standardization of research entities, the platform applies full-data automatic cleaning and standardization results to retrieved datasets. Compared with cleaning only the dataset itself, this better mines associations between research entities and improves standardization effects.

The platform automatically cleans and standardizes countries/regions based on world country and region name standard code tables. It uses a hierarchical hybrid structured deep learning framework model, employing single-layer bidirectional LSTM network vector semantic matching combined with character edit distance, supplemented by country, city, postal code, and institution name sorting features to calculate institution name similarity and automatically clean and standardize institutions. It uses author name disambiguation rule sets to automatically clean and standardize authors [41]. Using STKOS thesaurus standard concepts and synonyms, it automatically cleans and standardizes keywords. The platform supports manual editing and standardization of non-standard names and merging multiple non-standard names, defaulting to sorting standardized names by publication volume (see [Figure 6: see original paper]).

(5) Subject Information Analysis
In this step, considering that single-dimensional information analysis is insufficient to meet needs in the big data era, requiring multi-dimensional innovation from both data and methodological perspectives [42], the platform sets up quantitative analysis and content analysis modules. Different data types correspond to different analysis dimensions, embedding open-source tools such as ECharts and Gephi to visualize analysis results in line charts, bar charts, bubble charts, stacked charts, network diagrams, word clouds, maps, etc., with downloadable analysis result charts.

In the quantitative analysis module, the platform mainly conducts statistical analysis, cooperation network analysis, and co-word analysis on structured content such as year, country/region, institution, author, and technical composition (see [Figure 7: see original paper]). Papers, patents, and funding projects have 15, 21, and 13 analysis dimensions respectively. In the content analysis module, the platform extracts research problems and key technologies from unstructured scientific and technological literature content using an active learning-guided deep learning extraction framework [43], enriching the subject information intelligent analysis system at the semantic level. It conducts quantitative statistical analysis and association analysis on extracted research problems and key technologies (see [Figure 8: see original paper]).

Users can filter analysis dimensions for different data types and automatically generate and export information analysis reports. Reports feature NSTL icons and other proprietary characteristics, including analysis search strategies, data volume, literature type, time range, and analysis chart results with relevant textual descriptions. Users can also select analysis dimensions and subjects to publish on the platform front-end for other users to browse subject quantitative and content analysis results and gain deeper understanding of subject development trends.

3.3.2 Data Management

Based on interviews with information analysts and analysis of existing information analysis data storage status, we included analysis search strategies, datasets, standardized data, and analysis reports in the platform's data management scope. Analysis search strategies are the strategies used by analysts when retrieving data. Datasets are retrieved result datasets or datasets processed with human participation. Standardized data includes country/region standardized data, institution standardized data, author standardized data, and keyword standardized data. Analysis reports are automatically generated or manually processed and written by analysts.

We designed an interactive approach between online subject information analysis and subject information data management to meet real-time saving and management requirements for high-value intermediate analysis result data during subject information analysis. Users clicking the "save search strategy" button automatically saves the search strategy from the subject data aggregation step to the search strategy management list. Clicking the "confirm import to subject database" button automatically saves the retrieval result dataset to the dataset management list. Clicking the "save to my standard library" button automatically saves research entity standardized data from the subject data cleaning and standardization step to the standardized data management list. Clicking the "save report" button automatically saves the report generated in the subject information analysis step to the analysis report management list. The platform also supports user upload and import of local data, solving the problem of expert wisdom-incorporated data being in decentralized self-storage (see [Figure 9: see original paper]).

Standardized data saved in real-time by users during subject information analysis or uploaded from local sources is automatically applied to subsequent research entity data cleaning and standardization for newly created subjects, assisting in improving subsequent subject research entity data cleaning and standardization effects. Different data types have different description methods: analysis search strategy description fields include search strategy, search terms, and subject; standardized data description fields include standardized name, other names, and creation time; dataset description fields include dataset volume, acquisition method, subject affiliation, data year range, data type, and creation time; analysis report description fields include report name, report type, data year range, subject affiliation, generation method, and creation time (see [Figure 10: see original paper]).

3.3.3 Competitiveness Evaluation Analysis

Since subject country or regional competitiveness evaluation analysis often involves industry data, which is not a strength of the scientific and technological big data basic platform, the subject information data management and intelligent analysis platform focuses on visualizing and revealing competitiveness evaluation analysis dimensions and comprehensive evaluation results. Based on the designed competitiveness evaluation indicator system, we set up data organization templates for competitiveness evaluation analysis for platform users to download and use. Users can organize data according to the template and upload it to the platform for corresponding data visualization display.

Competitiveness evaluation analysis results can be published on the platform front-end for other users to understand the development levels of different countries or regions in subject fields. Visualization charts can be downloaded to assist information analysts in improving the speed and efficiency of writing or producing subject competitiveness analysis reports.

3.4 Implementation Effects

3.4.1 Fully Utilizing Aggregated Scientific and Technological Big Data Resources to Obtain Multi-Type Analysis Reports Through One Set of Operations

The platform breaks through the limitations of single data sources or single data types, fully utilizing aggregated scientific and technological big data resources to integrate multi-source and multi-type data such as journal articles, conference papers, patents, and funding projects. With a user-friendly interface and simple operation, the platform designs a wizard-style professional information analysis process, achieving process-oriented management of professional information analysis. Through one set of streamlined operations, users can obtain analysis reports based on different data types. Compared with scientific literature retrieval platforms, it more deeply embeds data cleaning and standardization functions and management functions for different data types. Compared with information analysis software, it realizes the entire process from data retrieval to data analysis, supporting both local retrieval and data import, compensat-

ing for the limitation of information analysis software that only supports data import.

During the COVID-19 outbreak, we used the platform's Demo version's professional information analysis function to quickly analyze and generate MERS-CoV and SARS-CoV data analysis reports based on papers, patents, and projects, attracting industry attention and resonance. Analysts responsible for the "Advanced Rail Transit" subject used the platform's full-process automated information research report production mechanism to complete relevant analysis reports, providing faster, more accurate, and comprehensive information support services for rail transit industry users [45]. Subjects published on the platform front-end (MERS-CoV, SARS-CoV) are shown in [Figure 13: see original paper], with data types used for analysis displayed on the right side of subject names and partial analysis results shown below. Clicking the "more" button on the right allows logged-in users to view all quantitative and content analysis results for the subject.

3.4.2 Realizing Multi-Dimensional and Multi-Level Analysis, Integrating Multiple Information Analysis Methods and Deep Learning Algorithms

The platform supports multi-dimensional and multi-level analysis modes, including quantitative analysis, content analysis, and competitiveness evaluation analysis, meeting different types of subject information analysis needs. Quantitative analysis employs information analysis methods such as statistical analysis, cooperation network analysis, and co-word analysis to reveal publication trends, scientific cooperation networks, and research hotspots. Subject data cleaning and standardization and content analysis respectively use different deep learning algorithms for research entity disambiguation and normalization and for identifying and extracting research problems and key technologies from unstructured literature content. Competitiveness evaluation analysis uses statistical analysis, comparative analysis, and comprehensive evaluation methods to display indicator sub-item results and comprehensive evaluation results.

Taking the "New Generation Artificial Intelligence" competitiveness evaluation analysis application demonstration as an example, the platform built a competitiveness evaluation model focusing on the new generation artificial intelligence and its subdivided industries, using strategic policies, industrial layout, scientific and technological development, capital support, and industrial prospects as first-level indicators, with corresponding 13 second-level indicators and 14 third-level indicators, to comprehensively compare and evaluate the industrial development levels of countries worldwide and domestic innovation centers. The comprehensive analysis and evaluation results of global new generation artificial intelligence industry development are shown in [Figure 14: see original paper]. Strategic policy analysis evaluates policy support strength and trends using policy data. Industrial layout analysis evaluates enterprise distribution in basic, technical, and application layers using enterprise data. Scientific and technological development includes scientific and technological input, output levels,

and cooperation levels using paper, patent, and project data. Capital support evaluates social capital investment using investment and financing data. Industrial prospect analysis evaluates industrial market development potential using market data. Clicking first-level indicators such as strategic policy, industrial layout, and scientific and technological development reveals analysis dimensions formed by second-level or third-level indicators.

Furthermore, the platform clarifies data management objects, enabling the preservation and management of data generated during subject information analysis processes and users' local data. Currently in the promotion and trial operation stage, the platform has over 40 users from 18 units (NSTL member units or service stations). The platform still has considerable room for improvement in data accuracy, data computing speed, and user experience. The project team will optimize and improve platform functions through a "service while building while improving" approach to provide new support and development for NSTL subject information services.

Data services and platform tools are essential stages for future intelligent information model transformation and upgrading, and are new paths and methods to effectively address the urgent information service needs faced by analysts with large volumes, multiple demand types, and tight deadlines. The subject information data management and intelligent analysis platform is an independently developed data management and information analysis tool that integrates multi-source and multi-type data, providing a means to deeply explore and release the value of multi-source aggregated scientific and technological data resources. It provides multi-dimensional and multi-level analysis services, process-oriented and intelligent information analysis, opens up the service chain from data to analysis, explores diversified analysis service implementation methods, seamlessly embeds multiple information analysis methods and deep learning algorithms, manages analysis process data and analysts' historical accumulation data, enriches service models, enhances analysts' service capabilities, and improves service response speed. In the future, we will explore more solutions for complex information analysis needs from perspectives such as text analysis and semantic analysis, continuously optimizing, improving, and iteratively upgrading the platform to build it into a commonly used tool for information analysts, helping them complete information analysis work better and faster.

References

[1] Liu Xiwen, Wang Li. Ten Years of Exploration in Subject Information Services for National Key Science and Technology Projects [C]//Proceedings of the 20th Anniversary of the National Science and Technology Library. Beijing: Science and Technology Literature Press, 2020: 346-349.
[2] Xian Guojian, Luo Tingting, Zhao Ruixue, et al. From Labor-Intensive to Computation-Intensive: The Transformation Path of NSTL Database Construction Model [J]. Digital Library Forum, 2020(7): 52-59.
[3] Qian Li, Xie Jing, Chang Zhijun, et al. Research and Design of Intelligent

Knowledge Service System Based on Scientific and Technological Big Data [J]. Data Analysis and Knowledge Discovery, 2019, 3(1): 4-14.
[4] InCites [EB/OL]. [2020-09-22]. https://incites.clarivate.com.
[5] SciVal [EB/OL]. [2020-09-22]. https://www.scival.com/.
[6] Incopat [EB/OL]. [2020-09-22]. https://www.incopat.com/.
[7] wisdom.ai [EB/OL]. [2020-09-22]. https://www.wisdom.ai/.
[8] From idea to impact-The next evolution in linked scholarly information [EB/OL]. [2020-09-22]. https://www.dimensions.ai/.
[9] Web of Science [EB/OL]. [2020-09-22]. http://apps.webofknowledge.com.
[10] CNKI [EB/OL]. [2020-09-22]. https://www.cnki.net/.
[11] Wanfang Data Knowledge Service Platform [EB/OL]. [2020-08-03]. http://www.wanfangdata.com.cn/index.html.
[12] CiteSpace: visualizing patterns and trends in scientific literature [EB/OL]. [2020-10-20]. http://cluster.ischool.drexel.edu/~cchen/citespace/download/.
[13] VOSviewer: visualizing scientific landscapes [EB/OL]. [2020-10-20]. https://www.vosviewer.com/.
[14] Sci2 tool [EB/OL]. [2020-10-20]. https://sci2.cns.iu.edu/user/index.php.
[15] The open graph viz platform [EB/OL]. [2020-10-20]. https://gephi.org/.
[16] Derwent data analyzer [EB/OL]. [2020-10-20]. https://clarivate.com/derwent/solutions/derwent-data-analyzer-automated-ip-intelligence/.
[17] PERSSON O, DANELL R, SCHNEIDER J. How to use bibexcel for various types of bibliometric analysis [C]//Celebrating scholarly communication studies: a festschrift for Olle Persson at his 60th birthday. Leuven: International Society for Scientometrics and Informetrics, 2009: 19-24.
[18] Liu Fei, Zhang Meiqi. Comparative Study of InCites Platform and SciVal in Application of Scientific Research Impact Evaluation [J]. Library Journal, 2019, 38(7): 60-68.
[19] Xu Jinglong, Lu Lucheng, Zhao Yajuan. Comparative Study of Patent Intelligence Analysis Tools for Patent Analysis Process [J]. Information Theory and Practice, 2020, 43(8): 178-185, 151.
[20] HERZOG C, HOOK D, KONKIEL S. Dimensions: bringing down barriers between scientometricians and data [J]. Quantitative science studies, 2020, 1(1): 387-395.
[21] Yu Xiaotong, Pan Xuelian, Hua Weina. Software Citation and Diffusion Analysis in Knowledge Graph Research [J]. Information and Documentation Services, 2019, 40(2): 19-29.
[22] Yang Jing, Cheng Changxiu. Comparative Analysis of Literature "Big Data" Analysis Software CiteSpace and Sci2 [J]. Computer Science and Application, 2017, 7(6): 580-589.
[23] Deng Jun, Ma Xiaojun, Bi Qiang. Comparative Study of Social Network Analysis Tools Ucinet and Gephi [J]. Information Theory and Practice, 2014, 37(8): 133-138.
[24] Wang Li. Comparison of Open Source/Free Tools and Research on Full-Process Solutions for Patent Analysis [J]. Information Theory and Practice, 2016, 39(1): 118-122.
[25] Liu Yuqin, Wang Xuefeng, Lei Xiaoping. Design and Implementation of

Scientific Research Relationship Construction and Visualization System [J]. Library and Information Service, 2015, 59(8): 103-110.
[26] Cui Ming, Pan Xuelian, Hua Weina. Research on Software Use and Citation in China's Library and Information Science Field [J]. Journal of Library Science in China, 2018, 44(3): 66-78.
[27] Harvard Dataverse [EB/OL]. [2020-12-10]. https://dataverse.harvard.edu/.
[28] Dryad [EB/OL]. [2020-12-10]. https://datadryad.org/stash.
[29] Australian national data service [EB/OL]. [2020-12-10]. https://www.ands.org.au/working-with-data.
[30] Chinese Academy of Sciences Data Cloud [EB/OL]. [2020-12-10]. http://www.csdb.cn/.
[31] Peking University Open Research Data Platform [EB/OL]. [2020-12-10]. https://opendata.pku.edu.cn/.
[32] Wuhan University Scientific Data Management Platform [EB/OL]. [2020-12-10]. http://sdm.lib.whu.edu.cn/jspui/.
[33] Cui Xu, Zhao Ximei, Wang Zheng, et al. Analysis of Achievements, Shortcomings, Countermeasures, and Trends of Scientific Data Management Platform Construction in China: Based on Domestic and International Comparison [J]. Library and Information Service, 2019, 63(9): 21-30.
[34] Wei Junchao, Zhang Chunfang. Comparative Study of Domestic and Foreign Scientific Data Management Platforms [J]. Library and Information Knowledge, 2017(5): 97-107.
[35] Zhu Ling, Nie Hua, Cui Haiyuan, et al. Construction of Peking University Open Research Data Platform: Exploration and Practice [J]. Library and Information Service, 2016, 60(4): 44-51.
[36] MORAL-MUNOZ J A, HERRERA-VIEDMA E, SANTISTEBAN-ESPINOSA A, et al. Software tools for conducting bibliometric analysis in science: an up-to-date review [J]. El profesional de la información, 2020, 29(1): 1-20.
[37] Zhou Chaofeng. Comparative Study of Common Bibliometric Software [D]. Wuhan: Central China Normal University, 2017.
[38] Xu Junlin, Liang Guangde, Zhong Hongying, et al. Research on Quality Control of Subject Intelligence Product Production in University Libraries [J]. Information Theory and Practice, 2013, 36(6): 68-72.
[39] Hua Bolin, Li Guangjian. Research on Architecture Design and Key Technologies of Intelligent Information Analysis System [J]. Library and Information, 2017(6): 74-83.
[40] Wang Ying, Zhang Zhixiong, Li Chuanxi, et al. Design and Implementation of Scientific and Technological Knowledge Organization System Open Engine System [J]. New Technology of Library and Information Service, 2015(10): 95-101.
[41] Zhang Jianyong, Qian Li, Yu Qianqian, et al. Research and Practice of Scientific Research Entity Name Standardization [J]. Data Analysis and Knowledge Discovery, 2019, 3(1): 27-37.
[42] Teng Guangqing, Ye Xin, Guo Siyue, et al. Evolution of Scientific and Technological Information Analysis from Single Dimension to Multi-Dimensional

Composite [J]. Digital Library Forum, 2019, 12(12): 2-8.
[43] Tao Yue, Yu Li, Zhang Runjie. Research on Active Learning Method for Phrase-Level Topic Extraction in Scientific and Technical Literature [J]. Data Analysis and Knowledge Discovery, 2020, 4(10): 134-143.
[44] Li Feng. How Libraries Conduct Discipline Competitiveness Evaluation: Enlightenment from the "International Comparison of UK Research Performance" Report [J]. Journal of Academic Libraries, 2015, 33(2): 72-76.
[45] Sun Yuling, Qin Aning, Peng Hao, et al. Experience in Advanced Rail Transit Technology Information Monitoring and Research Services Oriented to National Economic Main Battlefield [C]//Proceedings of the 20th Anniversary of the National Science and Technology Library. Beijing: Science and Technology Literature Press, 2020: 449-451.

Author Contributions

Yu Qianqian: Platform requirement investigation, research plan and functional system design, drafting, writing, and revising the paper;
Qian Li: Proposed platform construction framework and ideas, proposed paper revision suggestions;
Cheng Bing: Investigated subject information service needs, participated in analysis scenario design;
Chang Zhijun: Platform multi-source and multi-type data maintenance and import, platform technical support;
Wang Huili: Participated in new generation artificial intelligence competitiveness evaluation analysis application demonstration;
Jin Qian: Proposed platform construction goals, overall requirements, and platform improvement ideas.

Research on the Construction of Data Management and Intelligence Analysis Platform for Subject Information

Yu Qianqian[1,2] Qian Li[1,2] Cheng Bing[1] Chang Zhijun[1,2] Wang Huili[3] Jin Qian[4]
[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
[3]China National Chemical Information Centre, Beijing 100029
[4]Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081

Abstract: [Purpose/significance] In order to meet the subject information service needs of multi-disciplinary and multi-type users, we constructed a data management and intelligent analysis platform for subject information. The platform aims to achieve process-oriented and intelligent subject information analysis while managing the process data of subject information analysis that

reflects experts' wisdom, thereby enriching the service model and improving service response speed. [Method/process] Based on investigating existing relevant research and practice, we proposed the platform's design concept and construction framework, and analyzed its main functions and key technologies. [Result/conclusion] The platform has been completed. It integrates multiple sources and types of data, opens up the service chain from data to analysis, embeds a variety of information analysis methods and deep learning algorithms, and realizes multi-dimensional analysis services. It can manage analysis process data and historical data from information analysts, enabling data sharing and reuse.

Keywords: subject information; data management; intelligent analysis; information analysis

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