Large-Scale Knowledge Graph Construction for Extant Chinese Classical Texts Across Dynasties: A Postprint

Authors: Ouyang Jian, Liang Zhufang, Ren Shuhuai

Date: 2023-04-01T16:02:47+00:00

Abstract

[Purpose/Significance] This study explores the construction of a knowledge graph for extant Chinese historical books and records, aiming to provide researchers with a one-stop platform for mining the hidden knowledge behind massive ancient bibliographic data, expanding the connotation of ancient book knowledge services. Simultaneously, large-scale knowledge graphs of classical texts constitute an important foundation for machine intelligence. [Method/Process] Through knowledge graph technology, this research organizes knowledge of extant Chinese historical books and records across dynasties, constructs a framework model for classical text knowledge graphs from three components: requirement layer, model layer, and application layer, performs classical text data extraction and multi-source data fusion through human-machine collaboration, completes data collation, and analyzes and defines entity types and attributes, as well as entity relationships and types for the classical text knowledge graph. [Results/Conclusion] The constructed classical text knowledge graph contains 649,549 ancient book entities, 221,783 responsible parties for classical texts, 1,498,383 ancient book versions, and 13,960 geographical nodes, forming a three-dimensional, multi-dimensional, multi-purpose ancient book knowledge association network that provides a relatively comprehensive description of the main extant Chinese historical bibliographic information worldwide.

Full Text

Preamble

Research on the Construction of a Large-Scale Knowledge Graph of Extant Chinese Classical Books Through the Ages

Ouyang Jian[1] · [2], *Liang Zhufang*[3], *Ren Shuhuai*[1]

[1]Shanghai International Studies University Library, Shanghai 201620
[2]School of Journalism and Communication, Shanghai International Studies University, Shanghai 201620
[3]School of Management, Guangxi University for Nationalities, Nanning 530006

Abstract: [Purpose/Significance] This study explores the construction of a knowledge graph for extant Chinese classical books through the ages, providing researchers with a one-stop platform to mine the hidden knowledge behind massive bibliographic data of ancient books, thereby expanding the connotation of ancient book knowledge services. Meanwhile, a large-scale classical books knowledge graph also serves as an important foundation for machine intelligence. [Method/Process] Through knowledge graph technology, we organize the knowledge of classical books and construct a framework model from three components: the demand layer, model layer, and application layer. Using human-machine collaboration, we perform classical books data extraction and multi-source data fusion, complete data organization, and analyze and define entity types and attributes as well as entity relationships and their types for the classical books knowledge graph. [Result/Conclusion] The constructed knowledge graph contains 649,549 ancient book entities, 221,783 responsible persons, 1,498,383 ancient book versions, and 13,960 place name nodes, forming a three-dimensional, multi-dimensional, and multi-purpose ancient books knowledge association network that provides a relatively comprehensive description of the main extant Chinese classical books bibliographic information worldwide.

Keywords: ancient books; knowledge organization; knowledge graph; humanities research; digital humanities
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2021.05.013

Classical bibliographies represent a valuable treasure and serve as a key to unlocking China's ancient cultural heritage [1]. As an important component of philology, classical bibliographies have begun experimenting with new information technologies in their development and utilization amidst the wave of ancient books digitization, where the medium of ancient documents has shifted from physical forms to electronic, searchable digital formats. China's bibliography has a long history, and traditional bibliography and philology face unprecedented challenges and impacts. The digitization of ancient books has renewed the concepts and connotations of traditional philology [2], giving rise to the concept of digital philology in the modern digital environment [3] and creating demands for establishing digital bibliographies. Building digital bibliographies is necessary for protecting and promoting Chinese civilization, enabling users to comprehensively and systematically understand the 脉络 of Chinese civilization's development, while also meeting the needs of new documentation and researchers. Classical books knowledge graphs constitute an important component of ancient books digitization—a semantic knowledge organization and service model designed for the new information environment. By describing, revealing, and expressing the knowledge structure of classical books, they achieve the goals

of ancient books knowledge management and knowledge discovery, satisfying different users' needs for knowledge expression and presentation, thereby more intelligently feeding back to users the basic and extended information about classical books they require. Classical books knowledge graphs have become an important component of Chinese ancient books research platforms, providing foundational data for digital philology research. Establishing a knowledge graph of extant Chinese classical books is fundamental to ancient books knowledge services and can provide effective assistance for researchers in fields such as Chinese ancient books, history, philosophy, and linguistics.

2. Current Status of Knowledge-based Development and Utilization of Classical Book Catalogs

The compilation and indexing of classical book catalogs have been the primary forms of early classical books development. *Zhongguo Guji Zongmu* (General Catalog of Chinese Ancient Books) and *Zhongguo Guji Shanben Shumu* (Catalog of Chinese Rare Ancient Books) are typical representatives. With the development of computer technology, digital classical book catalogs and indexes have become the main research direction, with many scholars conducting beneficial explorations [4-7]. Additionally, there are numerous international studies on Chinese classical books. In the development and utilization of digital classical book catalogs, knowledge formalization is an important component. He Lin et al. [8] and Luo Chenguang et al. [9] proposed ontology-based classical books knowledge construction, beginning attempts at semantic and knowledge-based approaches. In 2009, the "Chinese Classical Books Through the Ages Catalog Analysis System" developed by Peking University and the National Library made meaningful attempts, creating an epoch-making classical literature catalog knowledge service system [10]. Compared with traditional classical book catalog applications, this system achieved significant progress, though it still had many limitations. The constructed classical knowledge units were somewhat deficient. Although it featured multi-dimensional correlation analysis functions for responsible persons, responsible behaviors, version characteristics, and binding features, the lack of supporting linked data for time, space, and various elements—such as critical compiler information—prevented analysis from more dimensions. The data granularity was also too coarse for more refined analysis, limiting the application's functionality. Therefore, it is necessary to expand knowledge related to classical books, including person information, time, and place names, linking, integrating, and aggregating different types and granularities of classical literature content to establish an ancient books knowledge association network. This would achieve functions such as ancient books knowledge storage, editing, indexing, knowledge mining, and knowledge discovery, meeting the demand for deep knowledge mining and recreation of classical book content value [11] to further discover the implicit knowledge within ancient books and greatly enhance the value of traditional classical book content.

Many scholars have also attempted to develop the "distinguishing academic

schools and examining origins" function of classical books. Song Denghan et al. utilized the RDA framework to design an overall description of ancient book version resources from three levels: norms, bibliography, and collection, aiming to achieve knowledge clustering functions for textual criticism in the description of ancient book version resources [12-13]. Deng Zhonghua et al. used ontology library construction technology to design classes, attributes, and instances for ancient book version knowledge data [14]. Xia Cuijuan et al. proposed the concept of "evidence-based ancient books" [15]. The "distinguishing academic schools and examining origins" function represents only a partial application of classical book knowledge value; more valuable applications of classical book knowledge await in-depth exploration.

Overall, current research and development applications for classical book knowledge remain relatively limited. The emergence of digital humanities research concepts has promoted the integration of humanities disciplines and technology, triggering transformations in the construction and development thinking of ancient books literature databases [16], and bringing new opportunities for the application and development of classical book catalogs. In the process of humanities research gradually emphasizing "scientific" transformation, knowledge association, quantitative analysis, and mining represent the development direction for the deep development and utilization of ancient books literature [17], providing new concepts and unique creative thinking for the deep development and utilization of ancient books literature knowledge.

From the perspective of humanities research applications, analysis should be conducted from temporal and spatial dimensions, including the era of book formation, compilers' native places, and collection locations. The construction of large-scale classical books knowledge graphs helps researchers comprehensively observe ancient book version and format information, understand the evolution of ancient scholarship, examine version origins, and clarify transmission 脉络. Using algorithms on the classical books knowledge network enables computational analysis of academic and social relationships among compilers, mining deeper levels of cultural development and changes in ancient China. It also allows analysis of correlations among multiple dimensions such as compilers, compilation time, compilation methods, and version features, further revealing the rich knowledge hidden behind ancient book data and breaking through traditional single-data-source statistical analysis patterns. Through ontology knowledge or rule-based reasoning techniques, implicit knowledge in data can be obtained; through link analysis, hidden relationships between entities can be discovered; and through inconsistency detection techniques, noise and discrepancies in ancient book cataloging data can be identified. The spatial information visualization analysis function of classical book compilers provides new research methods for spatial environment analysis in literary geography and, more importantly, offers foundational data services for classical book researchers.

In the tide of ancient books digitization that combines tradition with modernity and presents both opportunities and challenges, we must preserve fine tradi-

tions while adapting to the development trends of the digital age. The ultimate foothold of classical books knowledge services is user service, requiring a complete platform to build a one-stop classical book catalog retrieval system for scholars [23], helping researchers collect and organize large-scale classical book catalogs. By continuously meeting various needs of different users, we can accelerate the innovation and upgrading of ancient book literature in content, technology, and research, construct knowledge bases for various types of classical book knowledge such as versions, formats, time, geography, persons, and compilation methods, and provide knowledge graph services, along with various large-scale classical book statistics, analysis, data mining, and knowledge reasoning services. Large-scale classical books knowledge graphs are also the foundation of machine intelligence.

3. Framework Construction for Classical Books Knowledge Graph

3.1 Knowledge Requirements Analysis for Classical Books

Classical book catalogs are not only a gateway to scholarship but also important materials for examining academic origins [18-19]. Applications centered on classical book catalogs mainly focus on the textual criticism of ancient book version origins [20]. Classical book catalogs are closely related to ancient Chinese academic culture, integrating the culmination of ancient Chinese literati's classical books. They provide publication information against spatiotemporal backgrounds, offering clues to the transmission of classical books and providing another perspective for observing the geographical distribution, composition, and changes of ancient literati, reflecting to some extent China's economic and cultural development and social changes. Through classical book catalogs, one can deduce the transmission and survival status of classical books through the ages and infer the evolution of ancient Chinese academic thought, as well as reflect the ideological culture and academic interests of various dynasties [21]. Compiler information in classical book catalogs serves as an important clue for studying the academic and social relationships among compilers. In recent years, classical book catalogs and compilers have become important research objects in literary geography. Analyzing the geographical distribution of literati through the ages to understand changes in China's ancient literary world and the evolution of ancient scholarship has become an important basis and method for literary geography research [22].

3.2 Classical Books Knowledge Graph Framework

Knowledge graphs have become a research hotspot in the field of knowledge organization in recent years, representing a new massive knowledge management and service model based on semantic networks [24]. The main purpose of constructing knowledge graphs is to acquire large amounts of machine-readable knowledge, forming a networked knowledge structure that enhances associations

between knowledge units, achieves users' subject retrieval needs, and truly realizes semantic retrieval [25]. In recent years, knowledge graphs have also begun to be applied in humanities research, particularly in museum artifact knowledge and intangible cultural heritage organization, broadening the storage dimensions and data presentation methods of traditional humanities data and achieving efficient and stable knowledge management. This study uses knowledge graph methods to construct a knowledge graph of Chinese classical books through the ages, associating, organizing, and reconstructing scattered classical book data to reveal the relationships between classical book knowledge units, laying a foundation for knowledge-oriented mining and computation, and helping scholars discover implicit knowledge.

Classical book knowledge consists of ancient book compiler information, collection locations, and various classical book catalog metadata. Through effective combination of fragmented classical book knowledge units, a systematic classical book knowledge base is ultimately formed. The construction of classical books knowledge graphs is divided into three parts: the demand layer, model layer, and application layer (see [Figure 1: see original paper]). From the demand layer perspective, classical books knowledge graph construction should be demand-oriented, understanding the needs of humanities research. In a unified system platform, multiple attribute data of research objects are organized in the form of knowledge graphs to form a new, more effective comprehensive dataset of the research object or to obtain new implicit knowledge. Analysis and mining are conducted using three common research dimensions in digital humanities: time, place, and relationships: analyzing the evolutionary trajectory of classical books with time as the main thread to reflect the development of classical literary academic concepts; analyzing and interpreting research objects from geographical space, including various spatial elements and their structures (combinations) and functions; analyzing relationships and structures among works, compilers, versions, etc., based on the attribute data of research objects. Therefore, ancient book works, compilers, temporal information, and geographical information are important components of classical book knowledge, providing researchers with multiple analysis perspectives including bibliography, time, geography, persons, versions, and responsibility methods.

The application layer consists of specific application service modules such as classical book knowledge query, knowledge analysis, and knowledge discovery, as well as public application service modules. Each application service module provides specific services: classical book knowledge query mainly serves general users' classical book knowledge retrieval; classical book knowledge analysis serves domain researchers' classical book analysis; and knowledge discovery utilizes the reasoning and computational advantages of knowledge graphs to assist scholars in discovering implicit knowledge within classical book knowledge.

4. Human-Machine Collaborative Classical Book Data Extraction and Multi-source Data Fusion

4.1 Classical Book Data Extraction and Cleaning

Classical books knowledge graphs primarily provide support for classical book research. Therefore, the data sources and their accuracy and authenticity are crucial to knowledge graph construction and form an important foundation. The classical book data in this study were mainly extracted from domestic and foreign ancient book bibliography network databases, national census data, professional materials from the publishing field, general domain knowledge graphs, online encyclopedias, and relevant web pages on the World Wide Web, primarily through data crawling. Crawling classical book catalog data is relatively simple; by using self-developed collection software and setting corresponding collection rules for different data sources, the collection of classical book catalogs from corresponding data sources can be completed (see [Figure 2: see original paper]).

Since most ancient book websites have reorganized the original cataloging data when publishing, with most being semi-structured data as shown in [Figure 3: see original paper], the collected data require varying degrees of cleaning. It is necessary to extract metadata such as title, dynasty, compiler, and compilation method. A supervised approach is adopted in the classical book catalog extraction process: first annotating a small amount of data, then performing machine learning, and finally using the learned model to clean the same type of data or data conforming to specific relationships. For example: "Du Gongbu Caotang Shijian Twenty-two Volumes (Tang) Du Fu compiled (Song) Lu! edited (Song) Cai Mengbi annotated (Qing) Fang Gonghu collated" can be cleaned into structured classical book data as shown in [Figure 4: see original paper].

Relatively speaking, classical book compiler information is the focus and difficulty of data extraction. Classical book compilers were extracted from the compiler entries in the already-extracted classical book data, with 221,783 compilers identified after deduplication. In the constructed classical books knowledge graph, the compiler entity includes attributes such as dynasty, birth date, courtesy name, alternative name, posthumous name, occupation, native place, person tags, representative works, achievements, and official position, obtained mainly through three methods: structured data information extraction (such as the China Biographical Database (CBDB), *Dictionary of Chinese Historical Figures*, and other biographical dictionaries), semi-structured data information extraction (online encyclopedias), and unstructured data information extraction (web pages). These three types of data contain rich compiler attribute information; for example, CBDB contains considerable compiler information. The entire information extraction process is shown in [Figure 5: see original paper].

Only a small portion of data from CBDB and *Dictionary of Chinese Historical Figures* can match the compilers of ancient books, so most data need to be supplemented through online encyclopedias and network information. Using

compilers as keywords for retrieval, compiler information is extracted from encyclopedia search pages for supplementation. In encyclopedia websites, compilers are individual entities, with each entity page providing comprehensive introduction around one compiler. The webpage information structure is relatively fixed, allowing compiler information extraction through regular expression configuration of corresponding extraction templates. Encyclopedia content quality is relatively high, making encyclopedia websites the preferred choice for many knowledge graph constructions. For compilers that cannot be extracted and matched through structured or semi-structured data sources, search engines are used to find relevant web pages about classical book compilers. Since too many web pages are returned, a binary classifier needs to be constructed to determine whether the returned pages are introductory pages about ancient book compilers, and finally extract compiler information from those pages. Searching through online encyclopedias and search engines may result in issues such as polysemous classical book compilers, inconsistent compiler information attributes, unsplit multiple object attribute values, and non-uniform numerical attribute value formats. Therefore, data cleaning, name disambiguation, and data alignment are required. Particularly, many modern persons share the same names as ancient compilers; during processing, simple regular expressions can quickly determine whether they are ancient book compilers by judging the value range of birth and death years.

4.2 Multi-source Classical Book Data Fusion

Semantic linking and integration of heterogeneous knowledge resources are core components of knowledge graphs, requiring research on the association of heterogeneous data to transform it into knowledge networks with rich linking relationships. Data fusion involves using certain patterns and methods to synthesize multiple attribute data related to the same research object, forming a new, more effective comprehensive dataset that represents the research object. It integrates single or multi-source data of different categories, eliminates possible redundancy and contradictions among multi-source information, complements them, improves the reliability of research object information extraction, and enhances data use efficiency. Classical books knowledge graphs are also a project with highly fused data, containing multiple heterogeneous ancient book catalogs, person, and geographical data from different libraries' ancient book cataloging data, historical literature data, bibliographic materials, research results, and network data. These multi-source data need to be organized into an integrated whole and merged to support various research needs for knowledge expression and presentation. An important step in this process is data fusion, as shown in [Figure 6: see original paper].

Classical book data fusion mainly integrates structured, semi-structured, and unstructured classical book data such as ancient book catalogs, person, and place name data, as well as data from different sources. In classical books knowledge graph construction, multi-source data fusion mainly includes three

types: ancient book catalogs, classical book compilers, and place names. Ancient book catalogs contain important information such as versions, collection locations, and collection quantities. Due to different sources of ancient book catalogs and non-uniform cataloging rules, information originally belonging to the same classical book has some differences after collection, bringing difficulty to classical book data fusion. This study mainly uses title + compiler to determine whether they belong to the same classical book; when title and compiler are consistent, they are classified as the same classical book. Using this method, over 2.5 million classical book data entries were merged into 649,000+ different ancient books. Classical book compiler data mainly uses the combination of name + dynasty for compiler association; data meeting this combination condition are considered the same compiler, and relevant information about the corresponding compiler is extracted. Additionally, some compilers in ancient book catalogs use aliases or courtesy names; for example, the name Mohanzhai Zhuren corresponds to Feng Menglong, requiring mapping through a real name and courtesy name correspondence table. Place name data fusion mainly needs to handle changes in place names across different dynasties. Differences in place name changes across different data sources also lead to mismatches; for example, the ancient place name Jincheng refers to present-day Lanzhou, Gansu, which was called Jincheng in the Tang Dynasty, also known as Jincheng Prefecture, etc. In addition to automatic processing by computers, necessary manual intervention is also essential for data fusion (see [Figure 7: see original paper]). After extracting and obtaining relevant data from ancient book catalogs, person, and place name data, it is necessary to perform data transformation and establish structured data based on 梳理 and cleaning, achieving data integration and aggregation, and establishing a basic dataset. Meanwhile, redundant data from multiple data sources can be used to reasonably evaluate the accuracy of the knowledge graph. Redundant information can, on the one hand, improve the credibility of knowledge points, and on the other hand, provide reference for subsequent manual editing and verification, helping to eliminate ambiguities in ancient book works, versions, person names, and place names.

5. Implementation and Application of Classical Books Knowledge Graph

Classical book knowledge has both material attributes as a physical carrier recording knowledge content and spiritual attributes as materialized thinking and solidified knowledge. Therefore, the construction of classical books knowledge graphs must include both the external physical attributes and the internal implicit knowledge of ancient books, achieving multi-dimensional association of ancient book literature features to reach interoperability and shared use of classical book knowledge.

5.1 Entity Types and Attributes of Classical Books Knowledge Graph

The core of classical books knowledge graphs consists of bibliographic information, versions, compiler information, and place name information related to collection locations and compilers' native places. Based on the usage scenarios of classical books knowledge graphs, four entity types are determined: Work, Person, Version, and Place. The *Functional Requirements for Bibliographic Records* defines the concept of a work as abstract, representing unique intellectual or artistic creation; for ancient books, this refers specifically to a classical book compiled by a compiler, i.e., a specific bibliographic item. A version refers to various different copies of a book formed through multiple transcriptions, engravings, or other methods; one "work" can correspond to multiple "versions," and one "version" can have multiple copies with different collectors. Person corresponds to the compiler of a work. Place refers to the version collection location and the compiler's native place. The entity types of classical books knowledge graph are shown in .

Attributes are important metadata corresponding to entities in classical books knowledge graphs. Each work contains attributes such as title, compiler, compilation method, classification, and work formation era. The work formation era refers to the compiler's era, generally determined by the death year of the compiler [26]; for individual authors, the dynasty can be determined by referring to their life activities, book completion era, and traditional cataloging. Versions can be described through engraving time and version type (format), and also include unique compilers and compilation methods. Person is an important object of relationships in classical books knowledge graphs, containing attributes such as courtesy name, alternative name, birth year, death year, representative works, achievements, tags, occupation, native place, and index year (generally selecting the death year as the index year; when the death year is unknown, selecting the time point of their mentioned appointment or activity events in literature as the index year). Place mainly contains attributes such as country, province, city/county, name, and GIS information.

5.2 Entity Relationship Types of Classical Books Knowledge Graph

Entity relationship refers to the relationship between entities within a certain time period. There are multiple relationships between entities in classical books knowledge graphs, mainly including relationships between works and compilers, works and versions, versions and collection locations, and compilers and native places. There are compilation methods between works and compilers. By analyzing the compilation methods of collected ancient book catalogs, we found over 2,000 types. For the convenience of research and statistical analysis, it is necessary to normalize compilation methods. According to the compilation method requirements in *Zhonghua Guji Zongmu Cataloging Rules*: "Generally, record according to what is stated at the beginning of the main text volume; compilation methods with the same or similar nature as stated in the original book can be appropriately merged without strictly following the original state-

ment." Additionally, the following processing is applied: Compilation methods such as writing, authoring, narrating, studying, imitating, and discussing are all classified as "compiled"; Those who compile and organize previous authors' works are classified as "edited"; Those who collect and arrange previous authors' works are classified as "compiled"; Those who copy and compile relevant materials to form specialized books are classified as "compiled and revised"; otherwise, the original compilation method is used to establish the relationship between works and compilers. The relationship between work entities and version entities is "work-version," while the relationship between versions and collection locations is "collected in," and the relationship between compilers and native places is "born in." Some works are sub-items of other works. Therefore, except for the multiple relationships between works and compilers, other relationships between entities are relatively fixed. The entity relationships and types of classical books knowledge graph are shown in . Entities are connected through relationships to form the classical books knowledge concept graph (see [Figure 8: see original paper]).

5.3 Implementation of Classical Books Knowledge Graph

Knowledge graphs express "entity-attribute" and attribute values (statement) using a triple model. Currently, knowledge graph storage mainly includes linked data, graph databases, and relational databases [27]. After comprehensive comparison of the advantages and disadvantages of various knowledge graph storage methods, this study chose the graph database Neo4j for classical books knowledge graph storage. In Neo4j, knowledge units consist of vertices (Vertex), edges (Edge), and properties (Property), stored in the form of triple (S, P, O) data. Therefore, mapping needs to be established between classical books knowledge concepts and Neo4j storage. In Neo4j, node types correspond to classical books concept classes, i.e., nodes are divided into four instance types: Work, Person, Version, and Place. Each node corresponds to corresponding works, compilers, versions, and place instances, with each instance's attributes indicated by property names and values. Edges correspond to relationships between instances, with edge properties indicating the relationship types between entities (see). According to the mapping rules from the data model to database data, the cleaned, fused, and normalized data are converted into corresponding datasets and imported into Neo4j to implement the classical books knowledge graph (see [Figure 9: see original paper]).

The final classical books knowledge graph consists of over 2.5 million Chinese classical books through the ages collected from 743 libraries and research institutions worldwide, including 649,549 ancient book entities (Work instances), 221,783 responsible persons (Person instances), 1,498,383 ancient book versions (Version instances), and 13,960 place name nodes (Place instances). These four types of nodes and their relationships form a massive classical books knowledge graph. The nodes, attributes, and edges form a three-dimensional, multi-dimensional, and multi-purpose ancient books knowledge association network,

providing a relatively comprehensive description of the main extant Chinese classical books bibliographic information worldwide. This offers researchers a one-stop platform for mining the hidden knowledge behind massive ancient book bibliographic data, greatly enhancing ancient books knowledge service functions.

5.4 Applications of Classical Books Knowledge Graph

As a foundational knowledge service platform, classical books knowledge graphs can first provide basic classical book knowledge services for the general public, enabling them to understand traditional classical book knowledge through simple graphs and enhancing cultural dissemination effects.

Second, classical books knowledge graphs broaden the application scope of classical books. Multi-dimensional classical books knowledge graphs provide advanced services such as deep knowledge mining and knowledge reorganization for professional researchers. With classical books knowledge graphs, implicit knowledge in classical book data can be analyzed. Particularly in ancient book version comparison and origin examination, classical books knowledge graphs have significant advantages, enabling quick understanding of multi-dimensional correlations among version features and binding characteristics. They also allow analysis of book formation era, collection location, and collection quantity (see [Figure 9: see original paper]) to obtain quantitative historical distribution of academic development and research focuses.

Third, classical books knowledge graphs provide rich foundational research data services for related humanities research. Organized from different knowledge dimensions such as ancient book entities, classical book compilers, ancient book versions, and place name nodes, classical books knowledge graphs describe classical books from multiple perspectives, providing powerful multi-dimensional analysis functions for related research. With these data, more in-depth research can be conducted. For example, compilers of the same ancient book usually have certain relationships. At the levels of version, edition, impression, and copy, compiler information in ancient book catalogs serves as an important clue for studying academic and social relationships among compilers. Quantitative analysis of this compiler cataloging information can yield numerous relationships such as academic cooperation, academic inheritance, and social interactions. For instance, through interactive operations on the cooperation network, other compilers who have direct or indirect cooperation with compiler Wu Jianren can be discovered (see [Figure 10: see original paper]).

The relationship between literature and geographical environment is interactive. Analyzing the geographical distribution patterns of Chinese literati through the ages is an important content of literary geography research, with compilers of ancient books literature serving as the main subjects of analysis. In traditional research, studying literature from a geographical space perspective and parsing spatial information in texts is a complicated task. This classical books knowledge graph contains multi-dimensional data of related literati, enabling ancient

literary geography research using the native place attribute of compilers and assisting scholars in analyzing the geographical distribution of literati through the ages. Furthermore, using information such as the formation era of related ancient books literature in this knowledge graph, the evolution of ancient Chinese scholarship can be further examined.

Classical books knowledge graphs represent an attempt at deep development and utilization of ancient books literature, holding significant importance for enhancing ancient books catalog knowledge services. Classical books knowledge graphs can semantically annotate and link these information resources, establishing knowledge-centered resource semantic integration services.

Knowledge graphs represent a best practice in the field of knowledge engineering [28]. Their graph-based structure is more conducive to the representation of classical book knowledge and the association of various knowledge units, facilitating the storage, retrieval, and knowledge services of classical books. This study constructed a massive knowledge graph of Chinese classical books through the ages [29] using four types of nodes—ancient book entities, classical book compilers, ancient book versions, and place names—and their relationships. The nodes, attributes, and edges form a three-dimensional, multi-dimensional, and multi-purpose ancient books knowledge association network, greatly enhancing ancient books knowledge service functions.

This classical books knowledge graph basically completes classical book data acquisition, automatic annotation, and segmentation by computer, and on this basis, completes information extraction work and data semantic standardization. However, due to insufficient manual data review, some problems inevitably exist in the computer processing of data, and data quality needs to be improved subsequently. Some compiler attribute data also need to be supplemented and perfected. Meanwhile, the research depth and height of classical books knowledge graphs require further research and exploration; for example, intelligent question answering and knowledge reasoning functions of classical books knowledge graphs await further in-depth research and development.

References

[1] Gao Luming. Classical Bibliographies and Their Functions [J]. 文史知识, 1981(5): 105-109. [2] Ju Mingku. Ancient Books Digitization and Traditional Philology [J]. Journal of Tsinghua University (Philosophy and Social Sciences Edition), 2011, 26(5): 154-158, 161. [3] Yang Qinghu. Concepts and Issues of Digital Philology [J]. Heilongjiang Chronicles, 2013(13): 203. [4] Yu Manling, Yu Zhuohua. Experience in Compiling Ancient Books Indexes with Electronic Computers [J]. Journal of Sun Yat-sen University (Philosophy and Social Sciences Edition), 1988(4): 95-96. [5] Lin Zhongxiang. Practice and Theory of Compiling the "Complete Collection of Ancient and Modern Books" Index [J]. Journal of Guangxi University (Philosophy and Social Sciences Edition), 1994(2): 94-102. [6] Zhang Qiyu. An Example of Ancient Books

Index—Introducing the Index Database of the Electronic Edition of "Complete Collection of Ancient and Modern Books" [J]. Library Journal, 2000(5): 48-49. [7] Bao Juxiang. Automatic Compilation of Ancient Books Catalog Indexes—Taking "Chinese Ancient Books Index Database" as an Example [J]. China Index, 2013, 11(1): 25-29. [8] He Lin, Cao Ling. Research on the Construction of Agricultural Ancient Books Ontology and Its Retrieval Mechanism [J]. New Technology of Library and Information Service, 2006(12): 37-39, 53. [9] Luo Chenguang, Shan Chuan, Wang Shan. Preliminary Study on the Construction of Ancient Books Knowledge Base Based on Ontology [J]. New Technology of Library and Information Service, 2007(4): 8-11. [10] "Chinese Classical Books Through the Ages Catalog Analysis System" Passed National Technical Appraisal [EB/OL]. [2020-03-21]. http://pkunews.pku.edu.cn/xwzh/2009-11/02/content_{161068}.htm. [11] Gan Shenghong. Bringing to Life the Words Written in Ancient Books—Exploration and Practice of Ancient Books Knowledge Services by Zhonghua Book Company [EB/OL]. [2020-05-21]. https://news.artron.net/20180522/n1004873.html. [12] Song Denghan, Zhou Di, Li Mingjie. Design of Chinese Ancient Books Version Resource Description Based on RDA (I) [J]. Library, 2010(4): 51-53. [13] Song Denghan, Zhou Di, Li Mingjie. Design of Chinese Ancient Books Version Resource Description Based on RDA (II) [J]. Library, 2010(5): 49-52. [14] Deng Zhonghua, Huang Xin, Lu Yingjun, et al. On the Construction of Chinese Ancient Books Version Ontology Library [J]. Document, Information & Knowledge, 2014(4): 80-87, 93. [15] Xia Cuijuan, Lin Haiqing, Liu Wei. Research and Design of Chinese Ancient Books Data Model for Evidence-Based Practice [J]. Journal of Library Science in China, 2017, 43(6): 16-34. [16] Fan Jia. Connotation of "Digital Humanities" and Deep Development of Ancient Books Digitization [J]. Library Science Research, 2013(3): 29-32. [17] Xu Qing, Shi Xiangshi, Wang Wei. Deep Development of Ancient Books Digital Resources [J]. Library and Information Service, 2007, 51(3): 95-97, 79. [18] Fu Rongxian. On the Essence of Zhang Xuecheng's Concept of "Distinguishing Academic Schools and Examining Origins" [J]. Journal of Academic Libraries, 2016, 34(2): 111-117. [19] Lu Xin. Viewing the Function of Chinese Classical Bibliography from "Distinguishing Academic Schools and Examining Origins" [J]. Journal of Library Science in Jiangxi, 2008(1): 9-11. [20] Wang Guoqiang. Evaluation of "Distinguishing Academic Schools and Examining Origins"—Revaluation of the Value of Chinese Classical Bibliography [J]. Journal of Zhengzhou University (Philosophy and Social Sciences Edition), 1991(3): 77-82. [21] Zhao Tao. Academic Origin of the History Section in Ancient Book Catalogs and the Historical Direction of Evolution of Ancient Chinese Historiography [J]. Journal of Northwest University (Philosophy and Social Sciences Edition), 2015, 45(2): 26-32. [22] Zeng Daxing. Geographical Distribution of Chinese Literati Through the Ages [M]. Beijing: The Commercial Press, 2013. [23] Cao Xin. "One-Stop Ancient Books Catalog Retrieval System" is the Trend of the Times [N]. Xinhua Book News, 2017-02-17(12). [24] Li Juanzi, Hou Lei. Review of Knowledge Graph Research [J]. Journal of Shanxi University (Natural Science Edition), 2017, 47(3): 454-459. [25] Liu Qiao, Li Yang,

Duan Hong, et al. Survey on Knowledge Graph Construction Technology [J]. Journal of Computer Research and Development, 2016, 53(3): 582-600. [26] Cataloging Rules for the General Catalog of Chinese Ancient Books [EB/OL]. [2020-03-09]. https://max.book118.com/html/2017/0916/134109490.shtm. [27] Chen Tao, Liu Wei, Shan Rongrong, et al. Research on the Application of Knowledge Graphs in Digital Humanities [J]. Journal of Library Science in China, 2019, 45(6): 34-49. [28] Zhao Yiming. Is Knowledge Graph a Knowledge Organization System? [J]. Document, Information & Knowledge, 2017(5): 2. [29] Chinese Ancient Books Basic Data Analysis Platform [EB/OL]. [2020-11-09]. http://121.201.35.124:88.

Author Contributions:

Ouyang Jian: Overall framework and functional design, system development, paper conception and writing.
Liang Zhufang: Data collation and correction.
Ren Shuhua: Paper revision.

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