Organization and Reconfiguration of Humanities Data in Libraries in the Context of Digital Humanities: Post-print

Authors: Ouyang Jian, Peng Songlin, Li Zhen

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] Data is one of the foundations and cores of digital humanities research. The organization and reconstruction of humanities data in libraries can not only improve the utilization rate of digital resources, but also expand library humanities data services, greatly promoting the development of digital humanities. It is also a concrete manifestation of knowledge-based professional services in libraries, conducive to providing higher-level domain services. [Method/Process] By analyzing the characteristics of humanities data in digital humanities research and the needs of humanities scholars for humanities data, it is argued that libraries need to conduct humanities data organization and reconstruction from the perspectives of completeness, computability, usability and reusability, discoverability, and accessibility of humanities data. [Results/Conclusion] To overcome the shortcomings of fragmentation and lack of systematicity caused by humanities data fragmentation, it is necessary to adopt data restoration and reconstruction methods to restore or rebuild the connections between knowledge embedded in humanities data, and employ means such as datafication, data fusion, data linking and publishing, ultimately achieving fine granularity of knowledge units, semantic knowledge organization, and visual knowledge presentation.

Full Text

Organization and Reconstruction of Library Humanities Data in the Context of Digital Humanities*

Ouyang Jian[1], Peng Songlin[2], Li Zhen[2]

[1] School of Literature, Guangxi University for Nationalities, Nanning 530006
[2] Guangxi Library, Nanning 530024

Abstract

[**Purpose/Significance**] Data is one of the foundations and core elements of digital humanities research. The organization and reconstruction of humanities data in libraries can not only improve the utilization rate of digital resources but also expand library humanities data services, thereby greatly promoting the development of digital humanities science. This represents a concrete manifestation of knowledge-based professional services in libraries and facilitates the provision of higher-level domain services. [**Method/Process**] This paper analyzes the characteristics of humanities data in digital humanities research and the data requirements of humanities scholars, arguing that libraries must approach humanities data organization and reconstruction from the perspectives of data integrity, computability, usability and reusability, and discoverability and accessibility. [**Result/Conclusion**] To overcome the fragmented and unsystematic problems caused by humanities data fragmentation, it is necessary to restore or reconstruct the connections between knowledge embedded in humanities data through methods such as datafication, data fusion, data association, and publication, ultimately achieving fine granularity of knowledge units, semantic knowledge organization, and visual knowledge presentation.

1. Opportunities and Challenges for Humanities Disciplines

Digital humanities provides new research methods and paradigms for traditional humanities and social sciences research [1], fundamentally transforming scholars' research approaches by enabling research objects to be presented as data and analyzed using digital technologies [2]. This has profoundly impacted how humanities scholars manage their disciplines and share their research [3], as researchers increasingly demand large-scale access to copyrighted or licensed data for various forms of computational research (such as text mining, data mining, and machine learning). In the process of digital humanities research, libraries serve as resource repositories. After years of development, libraries have played a crucial role in the opportunities and challenges facing humanities disciplines.

The evolution of the information environment represents not merely a conceptual leap but also the highlighting of practical needs. Libraries have achieved positive results in supporting humanities scholars' computational research and providing resources and services. Libraries have progressed from initially participating in digitization project collaborations to now negotiating text mining rights with researchers and vendors, hosting data outputs, and providing research spaces and venues for digital humanities. This transition from digital collections to digital data, from data management to data services, and from data presentation to data analysis presents both challenges and opportunities for libraries, becoming a catalyst for library transformation and a new growth

point for library services. How to organize and reconstruct these data into humanities data suitable for digital humanities research is the prerequisite and key for libraries to launch humanities data services.

2. Library Humanities Data Services in the Context of Digital Humanities

Reconstruction is an important foundation for libraries to provide digital humanities services. The construction of domain data resources has reached considerable scale, with libraries establishing massive digital resources in humanities fields that provide a series of information infrastructures for digital humanities research. Data and data services have become extensions of library services [4]. Under the trend of disintermediation, libraries urgently need to transform from digital collections to digital data, from data management to data services, and from data presentation to data analysis. For years, libraries have been important collaborators in digital humanities research [5], though often primarily as providers of information resources and services [6]. Digital humanities research is committed to promoting widespread access to and sharing of information resources, innovating data processing and research methods, facilitating scholarly communication, strengthening learning and teaching, and enhancing the public influence of cultural information resources—goals that align perfectly with the mission and development objectives of libraries, making libraries natural partners in digital humanities [7].

As repositories of knowledge, information, and data, libraries have accumulated rich experience in data management and services such as data storage, organization, text mining, and metadata standards [8]. Digital humanities research provides libraries with a unique opportunity to successfully engage in interdisciplinary data management activities and establish close partnerships with academic groups in humanities and computer science [9].

Digital humanities evolved from humanities computing, with its most distinctive feature being quantitative analysis assisted by computers. Data is one of the foundations and cores of digital humanities research, making data-driven research mainstream in the field [10]. Digital humanities research requires humanities data to be integrated, granular, interconnected, and computable. With the development of digital libraries, vast amounts of humanities materials—including books, newspapers, journals, photographs, paintings, musical scores, ancient texts, images, and videos—have been digitized, forming large-scale, diverse, and valuable digital resources. Digital documents, databases, and retrieval systems have gradually become foundational platforms for humanities research. Although libraries have rich digital resources in philosophy, history, literature, linguistics, arts, and anthropology, the reality that library humanities data remains scattered, isolated, and closed has constrained libraries' role in digital humanities research.

The texts, images, audio, and metadata that deeply index and describe them

stored in libraries are often the research objects of digital humanities scholars. However, digitized information resources have not truly changed how users utilize documents; digital documents cannot transition from "reading" to "analysis." Therefore, current library participation in digital humanities research remains limited. As primary data managers and providers for digital humanities, libraries need to provide necessary humanities data for scholars, liberating them from the tedious work of data collection, organization, and authentication.

3. Characteristics of Humanities Data and Scholars' Needs in Digital Humanities Research

Humanities data consists primarily of computable digital encodings that can be processed by computers, comprising formatted data, texts, images, audio, and video. Library humanities data organization services and applications should be demand-oriented, thoroughly investigating researchers' data needs according to the characteristics of humanities data. Following the data acquisition, annotation, comparison, sampling, interpretation, and presentation methods in digital humanities research, libraries should extract, fuse, and reorganize existing data resources to form data needed for humanities research, actively creating foundational humanities data platforms that facilitate communication and creativity among researchers.

3.1 Characteristics of Humanities Data in Digital Humanities Research In traditional humanities research, scholars spend most of their time collecting and organizing relevant materials. Humanities research lacks research teams and is mostly conducted by individuals, with research materials accumulated over long periods. Combined with scholars' limited digital technology skills, this results in lengthy data construction cycles and highly individualized data. Different scholars often have vastly different interpretations of the same materials, making consensus difficult to achieve. Additionally, records often vary greatly in quality, volume, subject matter, recording methods, and level of detail, resulting in frequent substantial incompleteness in the data [20]. At the macro level, this determines that humanities data is chaotic and fragmented, presenting as unstructured, disorganized, and implicit, with diverse forms.

From a micro perspective, humanities data has two dimensions: the first describes data structure (clear and explicit), and the second describes data size and degree of variation [12]. Most humanities data consists of formatted tables with clear data structures and explicit attribute values. However, humanities data also has its particularities. Text in books or manuscripts or visual elements in paintings are analog, non-discrete data that are difficult to analyze or transform computationally. Language, texts, paintings, and music have symbolic systems beyond physically measurable dimensions, and analysis of these dimensions depends on semantics and pragmatics—meaning in context relies on contextual interpretation and manual annotation, thus potentially having multiple meanings. This increases the complexity of the relationship between

researchers and their research objects.

Computability and quantifiability are another major characteristic of digital humanities research. The core goal of digital humanities is to integrate modern information technology into humanities fields, thereby changing how knowledge is acquired, annotated, compared, sampled, interpreted, and presented. Through design, computational analysis, and visualization, digital humanities reshapes and transforms humanities knowledge, providing scholars with more possibilities and clues for differentiated, regular, macro-level, and trend-based research, thus expanding academic frontiers and potential.

3.2 Humanities Scholars' Data Needs Digital humanities research places unique demands on humanities data. The organization and construction of humanities data are largely determined by disciplinary norms and methodologies, requiring the involvement of humanities literacy. Only by understanding the characteristics of humanities data and meeting the research needs of humanities scholars can the effectiveness of humanities data be ensured. Integrated and fused digital materials and data are the foundation of digital humanities research. Humanities scholars' research patterns have shifted from primarily "reading" documents to "analyzing" them, transforming descriptive content into analyzable data as an auxiliary means for humanities research—namely, data-based research [21].

Space and time are dual dimensions upon which human survival and development depend and have been focal topics for philosophers throughout history [22]. Most objects of humanities research are closely integrated with time, requiring analysis of evolution, formation, and development processes along temporal lines, with in-depth understanding of diachronic content changes. From a spatial perspective, research objects are analyzed and interpreted through geographical space, while spatiotemporal analysis examines the distribution and combination of spatial positions and changes over time. Relationship analysis synthesizes temporal and spatial analysis, emphasizing the fixed connections and mutual influences of relationships or structures between things in time and space. Digital humanities research perspectives primarily focus on the relationships between research objects in space, time, and their interrelations, giving digital humanities research multi-perspective characteristics. Therefore, humanities data must be multidimensional, capable of describing object characteristics from temporal, spatial, and relational perspectives.

The multidimensional integration of humanities research data first requires fusing data from similar research purposes, comprehensively analyzing results from similar studies to obtain new concepts and elevate understanding to a new level. Second, it requires fusing research data from different categories and purposes. Humanities scholars have always valued case studies, which are micro-level research where the most basic value is material completeness—meaning a given case's humanities data should have as comprehensive coverage as possible. With the development of digital environments, increasingly more newly produced in-

formation resources are available, and more resources are being digitized, providing ample sources for humanities data construction.

4. Basic Elements of Library Humanities Data Organization and Reconstruction for Digital Humanities Research

Information scientist L. Floridi defines data as the most basic units, which can only be considered information when they have identifiable structure and meaning [23]. Data can be represented in many different forms, with the special characteristic of being discrete rather than continuous. Data in humanities can be considered as the selective digital representation and description of certain aspects of a given object that machines can understand and read [13]. The processing, organization, and interpretation of humanities data are determined by disciplinary norms and methodologies. The processability of humanities data enables measurement and identification at the macro level through micro data, allowing digital humanities scholars to widely apply humanities data to visualization and data mining [24].

The characteristics of humanities data in digital humanities research and scholars' needs constitute the basic elements of library humanities data organization and reconstruction, primarily including humanities data integrity, computability, usability and reusability, and discoverability and accessibility.

4.1 Humanities Data Integrity Humanities research has obvious empirical characteristics, with the academic tradition of selecting authentic, reliable, and reasonably scoped research materials [25]. Humanities researchers are extremely concerned about the authenticity of materials and their traceability. In the context of digital humanities, library humanities data organization and reconstruction must first achieve humanities data integrity [14]. Integrity has two meanings: first, achieving traceability of humanities data throughout their lifecycle of collection, processing, conversion, and publication; second, referring to the coverage extent of a certain type of humanities data.

Humanities research places great emphasis on the authenticity and reliability of materials. Protecting the integrity of humanities data collection, processing, and conversion, as well as the traceability of key materials, enables critical addressability of humanities data. This allows researchers to understand why certain data are included or excluded, why certain transformations are performed, who performed these transformations, and to access the code and tools used for these transformations. Like web archives, this key addressability concept is crucial throughout the research process. Researchers hope to select, evaluate, and trace original materials according to humanities research needs. Integrity is mainly reflected in three aspects: source, processing, and presentation. Source refers to the origin of humanities data, ensuring traceability of original materials. Processing refers to scholars' data processing and transformation, requiring data consistency. Presentation refers to methods and tools for presenting processed data. Collection sources, processing, and presentation methods directly

affect scholars' trust in the data. Projects like the China Biographical Database (CBDB) and Chinese Civilization in Time and Space (CCTS) retain source record information during data construction.

4.2 Humanities Data Computability Digital humanities is the continuation and development of humanities computing. Quantitative analysis of computable humanities data is the core of digital humanities and a distinctive feature that separates it from traditional humanities research. Quantitative analysis involves quantifying the characteristics of research objects, evolving from single-variable statistical descriptions of events with numerical characteristics to quantitative studies of multiple non-numerical things or events [26]. Objects involved in computation must have clear measurement attributes, requiring humanities data and knowledge to be granular with independent attributes. This necessitates fine-grained metadata processing and annotation from multiple perspectives to reveal formal and content features of documents.

Most humanities data consists of formatted tables with clear data structures and explicit attribute values. However, humanities data also has its particularities. Text in books or manuscripts or visual elements in paintings are analog, non-discrete data that are difficult to analyze or transform computationally. Language, texts, paintings, and music have symbolic systems beyond physically measurable dimensions, and analysis of these dimensions depends on semantics and pragmatics—meaning in context relies on contextual interpretation and manual annotation, thus potentially having multiple meanings. Humanities data computability and quantifiability are major characteristics of digital humanities research. The core goal is integrating modern information technology into humanities fields to change how knowledge is acquired, annotated, compared, sampled, interpreted, and presented. Through design, computational analysis, and visualization, digital humanities reshapes humanities knowledge, providing scholars with more possibilities for differentiated, regular, macro-level, and trend-based research.

4.3 Humanities Data Usability and Reusability Humanities data is one of the foundations of digital humanities research. Although digital humanities is interdisciplinary, it is also highly specialized, making humanities data often extremely specific and limiting its application scenarios. Achieving universality and applicability of humanities data is a goal for libraries. To enable constructed humanities data to be applied to more research scenarios, usability and reusability are crucial for digital humanities. The form of storage and publication of humanities data objects directly affects their usability and reusability.

Humanities data is typically instantiated in certain formats. A set of universal formats and data structures can better support digital humanities research and teaching. With the establishment of various standards for humanities data documentation, disciplinary research and teaching have been greatly promoted. The Text Encoding Initiative (TEI) defines a series of universal standards for

electronic text materials and is widely used in text-based humanities research worldwide [27]. Although TEI is considered a core format for advanced users, the data preserved in its underlying XML files is often less suitable for digital humanities. Therefore, in recent years, humanities data has increasingly catered to scholars' actual needs, developing collection and conversion strategies to facilitate transformation between various data types in digital humanities [28]. During organization and construction, it is necessary to determine at the functional level which humanities data tables are most welcomed by researchers and to convert them according to common data format requirements of digital humanities research tools and methods, generating more user-friendly formats like Access or Excel to improve usability.

Currently, most humanities data is produced by research projects. After project completion, team dissolution becomes a major issue for data maintenance and updates, making long-term storage and curation key factors for usability and reusability. The rise of data curation services provides guarantees for long-term preservation, exchange, and broader access and reuse of digital humanities research data [29].

4.4 Discoverability and Accessibility Humanities datasets vary enormously, with numerous mixed data causing data silos to persist. The primary purpose of humanities data construction is to serve humanities scholars, making discoverability, accessibility, and availability crucial for library data services. Humanities research often requires comparing, sampling, and interpreting across multiple datasets, demanding strong interconnectivity between humanities data. Therefore, large-scale fusion of relevant humanities data and fine-grained, relational reconstruction of resources have become key focuses for libraries supporting humanities research [30].

To support accessibility and availability, data revelation during construction, organization, and management is essential, primarily reflected in internal descriptions such as cataloging, indexing, ontologies, and semantics. These describe data characteristics and various relationships. External discovery tools like browsing, navigation, and retrieval are equally indispensable, directly connecting researchers with data. Google's Dataset Search tool (https://toolbox.google.com/dataset-search) demonstrates the importance of data retrieval, enhancing discoverability. The value of reproducibility and transparency in humanities data is increasingly recognized, with the focus of organization and development being data access. Current access methods vary greatly, from simple static web pages, XML files, and XSL files to GitHub and FTP access, and increasingly to Application Programming Interfaces (APIs) that simplify data acquisition.

5. Models and Methods for Library Humanities Data Organization and Reconstruction for Digital Humanities Research

5.1 Basic Concepts In digital library concepts, libraries primarily organize digital information resources that exist in highly structured ways according to disciplinary classification systems, resource-centered, with strict hierarchical structures forming static "pyramid-shaped" information architecture (see [Figure 1: see original paper]). For humanities scholars, such resources remain at the document level, unable to transition from "reading" to "analysis," with rigid hierarchical structures failing to meet disciplinary research needs. This differs significantly from digital humanities requirements for integrity, computability, usability, reusability, and discoverability.

Digital humanities scholars typically focus on thematic research requiring comprehensive data coverage, necessitating personalized data customization. Humanities data should shield researchers from data structures and storage, serving research topics and enabling convenient, rapid access to relevant data. Data should form a "web-like" structure centered on research topics (see [Figure 2: see original paper]), where organization and reconstruction transform data into "smart data" reflecting specific disciplines or research areas, establishing connections to meet scholars' multidimensional analytical needs.

To overcome the fragmented and unsystematic problems caused by data fragmentation, libraries must restore and reconstruct connections between knowledge embedded in humanities data. Organization and reconstruction mainly take two forms: humanities data restoration and humanities data reconstruction. Restoration rebuilds original systematic data and knowledge structures according to existing disciplinary knowledge systems, focusing on datafication and establishing relationships between data. This involves completing granular deep indexing and description of texts, images, and audio-visual materials to form original complete humanities data structures, then establishing associations based on their relationships (see [Figure 3: see original paper]). Reconstruction does not strictly follow original disciplinary knowledge systems but organizes and rebuilds personalized research topic structures according to scholars' needs, discovering previously unknown relationships and knowledge in original data to better solve real problems and facilitate knowledge innovation (see [Figure 4: see original paper]). The Venice Time Machine project exemplifies this by reconstructing a thousand years of Venetian history from discrete knowledge in digitized ancient maps, monographs, manuscripts, and musical scores [31].

5.2 Methods for Organization and Reconstruction Humanities data organization and reconstruction differ from traditional digital library integration. Integration, also called digital resource integration, uses various technologies to integrate autonomous, distributed, and heterogeneous resources into a unified environment for "one-stop" retrieval. In contrast, humanities data organization and reconstruction involve analysis, synthesis, transformation, and publi-

cation to construct complete, authoritative humanities datasets and establish connections between them. This includes not just digitization but multi-angle granular deep indexing and metadata description, datafication, data fusion, and knowledge association, ultimately achieving fine granularity of knowledge units, semantic knowledge organization, and visual knowledge presentation.

5.2.1 Datafication Digital humanities research includes digitization, datafication, data management, and computational analysis. Digitization as the final product of digital libraries still has distance from the computability required by digital humanities. Therefore, it is necessary to further convert electronic forms into recognizable texts and analyzable data for further measurement. Datafication is fundamental work in digital humanities, with its core task being the reorganization of document content into new text or data structures established by users—namely, the structuralization of documents [19]. Digital humanities research directly manages data in collection and annotation, while comparison, sampling, and computational analysis depend on data analysis and interpretation.

Datafication includes optical character recognition (OCR) to make document resources suitable for text analysis and mining—an initial stage. It also includes reorganizing document content into quantifiable forms for tabular analysis [32]. Currently, most document datafication remains at the OCR stage, with automatic recognition of manuscripts and ancient texts still facing significant technical challenges. With digital humanities development, more content reorganization and formalization work has emerged, such as the China Biographical Database (CBDB) developed by Harvard University, Peking University, and Taiwan's Academia Sinica; the China Historical Geographic Information System (CHGIS) by Harvard's Center for Geographic Analysis and Fudan University's Institute of Historical Geography; Shanghai Jiao Tong University's "Chinese Local Historical Documents Database"; and multiple databases from National Taiwan University's Digital Humanities Research Center. These projects transform needed document content into quantifiable, analyzable forms from a digital humanities perspective, achieving multi-angle refinement and association to meet requirements for integrity, computability, usability, reusability, and discoverability.

5.2.2 Data Fusion Traditional quantitative analysis typically involves deep tracking and analysis of single data sources, where analysts have control and deep understanding of data sources and structures. Digital humanities research particularly emphasizes data reusability and multi-perspective sampling analysis, making the formation of effective multi-perspective analytical datasets a necessary challenge and foundation for humanities research in the big data era. Datafication only achieves the mapping of traditional digital humanities materials into the digital world for computer storage, processing, and display. The multidimensionality of humanities data requires organization through information and knowledge units to construct simulated domain application environments,

making data fusion indispensable.

Data fusion enables humanities scholars to easily handle diverse, multi-source data for multi-dimensional mining and analysis, helping discover new patterns and values. After years of development, digital library resource platforms have stored large amounts of computable foundational data as important sources for digital humanities. Therefore, data reuse and reorganization are crucial tasks. Fusing different humanities data from different libraries is essential for digital humanities research.

Humanities data fusion uses certain patterns and methods on multiple attribute data related to the same research object to generate a new, more effective comprehensive dataset or obtain new implicit knowledge. It integrates single or different types of multi-source data to eliminate redundancy and contradictions while complementing information, improving timeliness and reliability of information extraction and data use efficiency. The process involves connecting required multi-source databases, acquiring relevant data, studying and understanding the data, cleaning and transforming it, and establishing structures to achieve data combination, integration, and aggregation.

Fusion forms include heterogeneous fusion (structured, semi-structured, and unstructured data), multi-source fusion (data from different disciplines and sources), and multi-modal fusion (text, images, audio, etc.) (see [Figure 5: see original paper]). Fusion levels include data-level fusion (direct combination of raw data), feature-level fusion (feature extraction and comprehensive analysis), and decision-level fusion (semantic-based fusion for joint inference) (see [Figure 6: see original paper]) [33].

5.2.3 Data Association and Publication Accessibility and usability of humanities data are major challenges in digital humanities research. Scholars often need to transfer research queries from one dataset to another or enable cross-dataset queries. However, the diversity of humanities research results and the organization of research around individual or small-group efforts make data access, sharing, and reuse seem unattainable. Linked data technology is most likely to fill this gap and improve limitations in accessing and reusing humanities data [34].

Humanities datasets are formed by establishing associations between various data, ranging from ancient maps to bibliographic records, paintings, audio-visual materials, ancient text analysis, and illustrated facts, all with close relationships requiring aggregation, integration, and cross-searching. Semantic technologies and linked data enable large-scale collaborative and aggregated research in digital humanities. Linked data and knowledge graphs enhance machine readability and understanding, building explicit semantic ontologies to achieve literature knowledge content revelation, enhancing data reuse and interconnection with external data, and forming an interconnected, decentralized global knowledge network. Shanghai Library has conducted extensive practice in applying linked

data technology to genealogical and historical geographic data [36], enabling the transformation from library-centered knowledge organization systems to cross-domain publicly available and accessible knowledge graphs, improving usability and reusability.

Domain-specific knowledge structure development and management are crucial elements in humanities. Knowledge organization systems have centuries of tradition in libraries, used for organizing resources and facilitating discovery and retrieval in metadata description. With linked data, these systems have undergone digital transformation and entered the internet era. Since Google Knowledge Graph's emergence in 2012 [37], knowledge graphs have attracted widespread attention and become a research hotspot, bringing new transformations to knowledge organization in libraries.

Knowledge graphs are a recent research focus in knowledge organization—a new massive knowledge management and service model based on semantic networks, aiming to describe concepts, entities, events, and their semantic relationships in the objective world [38]. The main purpose is to acquire large amounts of machine-readable knowledge, entity-attribute-value pairs, with entities interconnected through relationships to form network structures that enhance associations between knowledge units and enable semantic retrieval [39].

Knowledge graph technology has long been used to manage authoritative data about places or people in specific domains, following patterns similar to library knowledge organization schemes. They are increasingly published according to linked data principles and connected with other web knowledge graphs using shared semantic concepts [40]. Knowledge graphs are inherently semantic representations with clear semantic web characteristics. Their construction for digital humanities research enables semantic annotation and linking of information resources, generating new comprehensive datasets or implicit knowledge for research objects with multiple attributes. Cross-domain knowledge graphs open new research opportunities.

In recent years, knowledge graphs have gained increasing attention in digital humanities projects domestically and internationally, demonstrating tremendous application prospects. The authors have been constructing ancient book knowledge graphs, organizing nearly 2 million ancient book entries from China, Japan, Europe, and America based on elements closely related to ancient books (compilers, birthplace, time period, compilation methods, collection institutions) and building multidimensional relationships from temporal, spatial, and relational perspectives. This forms an ancient book knowledge graph [41], facilitating semantic indexing, classification, querying, computational analysis, and visualization of ancient book knowledge. From a digital humanities application perspective, strong knowledge associations help examine version origins and evolution patterns. Graph analysis can reveal correlations across multiple dimensions including responsible parties, compilation time, methods, and version characteristics, uncovering rich cultural and historical knowledge hidden behind ancient book data. This breaks through traditional single-source statistical analysis

models, uses rule-based reasoning to obtain implicit knowledge, and provides new research methods for literary geography spatial analysis through visualization of responsible parties' spatial information, enhancing the value of ancient book catalog knowledge services.

6. Conclusion

Data is one of the foundations and cores of digital humanities research. In digital humanities research processes, libraries as resource repositories have massive digital resources that become important sources of humanities data. Under the trend of disintermediation, libraries urgently need to transform from digital collections to digital data, from data management to data services, and from data presentation to data analysis. This presents both opportunities and challenges for libraries, becoming a catalyst for transformation. The organization and reconstruction of humanities data for digital humanities research is key. Libraries must understand the characteristics of humanities data and scholars' needs, approaching organization and reconstruction from the perspectives of integrity, computability, usability and reusability, and discoverability and accessibility. To overcome the fragmented and unsystematic problems caused by data fragmentation, libraries must restore or reconstruct connections between knowledge embedded in humanities data using datafication, data fusion, data association, and publication, ultimately achieving fine granularity of knowledge units, semantic knowledge organization, and visual knowledge presentation. This not only improves digital resource utilization but also expands library humanities data services, greatly promoting digital humanities development and embodying knowledge-based professional services for higher-level domain services.

References

[1] Wang Xiaoguang. Digital Humanities: Concepts, Current Status, and Reflections [EB/OL]. [2018-06-26]. http://meeting.lib.szu.edu.cn/conference/zh-hans/information?v=07000003.
[2] Leonard P. Mining large datasets for the humanities [EB/OL]. [2018-08-26]. http://library.ifla.org/930/1/119-leonard-en.pdf.
[3] Lyman P. Challenges to digital libraries and digital humanities [M]. Translated by Zhu Changhong. Guangxi: Guangxi Normal University Press, 2010.
[4] Qin Jian. Data and data services: Extension of library services [EB/OL]. [2018-06-26]. http://society.library.sh.cn/.
[5] Special report: digital humanities in libraries [EB/OL]. [2018-07-09]. https://americanlibrariesmagazine.org/2016/01/04/special-report-digital-humanities-libraries/.
[6] Schaffner J, Erway R. Does every research library need a digital humanities center? [EB/OL]. [2018-07-27]. http://oclc.org/research/publications/library/2014/oclcresearch-digitalhumanitiescenter-2014-overview.html.

[7] Spiro L. Why digital humanities [EB/OL]. [2018-07-01]. http://digitalscholarship.files.wordpress.com/2011/1
5.pdf.
[8] ACRL. Top trends in academic libraries [EB/OL]. [2018-07-01]. http://crln.acrl.org/content/75/6/294.full#xref-ref-68-1.
[9] Vandegrift M, Varner S. Evolving in common: creating mutually supportive relationships between libraries and the digital humanities [J]. Journal of library administration, 2013(53): 67-78.
[10] Data driven: digital humanities in the library [EB/OL]. [2018-07-15]. http://memory.loc.gov/.
[11] American memory: remaining collections [EB/OL]. [2018-12-15]. http://memory.loc.gov/.
[12] HathiTrust research center data capsule v1.0: an overview of functionality [EB/OL]. [2018-12-13]. http://dspace.handle.net/2022/18936.
[13] Big? Smart? Clean? Messy? Data in the humanities [EB/OL]. [2018-07-18]. http://dhinthelibrary.wordpress.com/.
[14] Padilla T. Humanities data in the library: integrity, form, access and discovery [J]. Public services quarterly, 2014, 10(4): 324-35.
[15] Humanities data: a hands-on approach [EB/OL]. [2018-12-13]. http://digital.humanities.ox.ac.uk/dhoxss/2016/workshops/dhcurationshops/dhcurations.
[16] Flanders J, Munoz T. An introduction to humanities data curation [EB/OL]. [2018-07-09]. http://dlib.org/dlib/march16/padilla/03padilla-print.html.
[17] Shaping humanities data: use, reuse, and paths toward computationally amenable cultural heritage collections [EB/OL]. [2018-12-13]. https://dh2017.adho.org/abstracts/670/670.pdf.
[18] Xu Liheng. The datafication of Tang dynasty figures: a glimpse into the China Biographical Database (CBDB) [M]// Bao Weimin, Liu Houbin. Tang-Song Historical Review, Vol. 3. Beijing: Social Sciences Academic Press, 2017: 20-32.
[19] Zhao Siyuan. Digitization, datafication, and text mining of local historical documents: taking the “Chinese Local Historical Documents Database” as an example [J]. Qing history research, 2016(4): 26-35.
[20] Xu Liheng. Big data of Chinese historical figures [J]. Communications of CCF, 2018, 14(4): 19-24.
[21] Liu Wei, Ye Ying. Discussion on the technical system and theoretical structure of digital humanities [J]. Journal of library science in China, 2017, 43(5): 32-41.
[22] Mei Xinlin. Literary geography: theoretical construction based on the dimension of “space” [J]. Zhejiang social sciences, 2015(3): 122-136, 160.
[23] Floridi L. Information: a very short introduction [M]. Oxford: Oxford University Press, 2010: 22-25.
[24] Padilla TG, Higgins D. Library collections as humanities data: the facet effect [J]. Public services quarterly, 2014, 10(4): 324-35.
[25] Li Xingdong, Liu Xiaohong. Behind the demand for authenticity in scientific research materials: on the value pursuit of humanities research [J]. Contemporary educational science, 2007(11): 43-44.

[26] Quantitative analysis in humanities: a tentative exploration [J]. Theoretical studies in literature and art, 1992(1): 43.
[27] Sperberg-McQueen M. Text encoding and enrichment [M]// The humanities computing yearbook 1989-90. Oxford: Oxford University Press, 1991.
[28] Schreibman S, Siemens R, Unsworth J. A companion to digital humanities [M]. Oxford: Blackwell, 2004.
[29] Poole AH. A greatly unexplored area: digital curation and innovation in digital humanities [J]. Journal of the Association for Information Science & Technology, 2017, 68(7): 1-10.
[30] Li Xin, Zhang Yi, Wang Zhili. Digital humanities research needs for integrating heterogeneous special collections in libraries [J]. Digital library forum, 2017(11): 48-53.
[31] Abbott A. The 'time machine' reconstructing ancient Venice's social networks [J]. Nature, 2017, 546(7658): 341-344.
[32] Schönberger. Big data: a revolution that will transform how we live, work, and think [M]. Translated by Zhou Tao. Hangzhou: Zhejiang People's Publishing House, 2013: 104.
[33] Ouyang Jian. Multi-source data fusion for digital humanities research [EB/OL]. [2018-12-13]. http://society.library.sh.cn/adls2016.
[34] Hoekstra R, Meroño-Peñuela A, Dentler K, et al. An ecosystem for linked humanities data [EB/OL]. [2018-12-13]. https://www.semanticscholar.org/paper/An-ecosystem-for-Linked-Humanities-Data-Hoekstra-Mero%C3%B1o-Pe%C3%B1uela/e-82d876d5e7ef8a09430d38ee340da86acf0d0c?tab=abstract.
[35] Linked open data for cultural heritage and digital humanities [EB/OL]. [2018-07-09]. https://ontotext.com/linked-open-data-cultural-heritage/.
[36] Xia Cuijuan, Liu Wei, Chen Tao, et al. Development practice of genealogical linked data service platform [J]. Journal of library science in China, 2016, 42(3): 27-38.
[37] Singhal A. Introducing the knowledge graph: things, not strings [EB/OL]. [2018-07-21]. https://goo.gl/U168iz.
[38] Li Juanzi, Hou Lei. Survey of knowledge graph research [J]. Journal of Shanxi University (natural science edition), 2017(3): 454-459.
[39] Liu Qiao, Li Yang, Duan Hong, et al. Survey of knowledge graph construction techniques [J]. Journal of computer research and development, 2016, 53(3): 582-600.
[40] Haslhofer B, Isaac A, Simon R. Knowledge graphs in the libraries and digital humanities domain [EB/OL]. [2018-12-13]. https://arxiv.org/pdf/1803.03198.pdf.
[41] Ouyang Jian. Construction of large-scale ancient book knowledge graph for digital humanities research [R]. Beijing: Library Society of China, 2017.

Author Contributions

Ouyang Jian: Topic selection, structural conception, and paper writing;
Peng Songlin: Paper revision;

Li Zhen: Paper revision.

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