From collection resources to intelligent data: Construction of intelligent digital humanities platform for local historical documents of Shanghai Jiao Tong University

Yin Qian, Zhuoyuan Xing and Xiaohua Shi
Shanghai Jiao Tong University, China

Abstract

Local historical documents originated from daily life of people belong to special collection resources that were not published publicly. They are valuable assets of universities and libraries. At present, most documents had only finished digitalization or partial datalization work. However, the requirements of deep knowledge mining in documents data, providing visual analysis, and effectively supporting the research of historic humanities scholars had not been fully met. Taking the local historical documents project of Shanghai Jiao Tong University as an example, using relevant techniques of digital humanities (DH), the in-depth analysis and utilization research of documents data were carried out. On the one hand, the core database of the documents was established based on standardizing metadata cataloguing and establishing metadata association. On the other hand, based on the core database, an intelligent DH system platform was constructed. The platform is to realize full-field retrieval and display of the documents, text analysis, association analysis, statistics, and visual presentation of knowledge. In addition, in the process of using the platform for research, humanities scholars can continuously expand the data dimensions and the relationships between data, achieve intelligent supplementation of documents data and platform self-learning. The concept of DH has led to a new direction of database construction and platform development. In the exploration and practice of DH, libraries should continue to widen thinking, improve service and innovation capabilities, and provide better research perspectives, research environments, research support, and research experience for humanities scholars.

Correspondence:
Xiaohua Shi, Library, Shanghai Jiao Tong University, Shanghai, China.
E-mail:
xhshi@sjtu.edu.cn

1 Introduction

Local historical documents originated from daily life of specific area people belong to special collection resources that were not published publicly. For a long time, university history departments and libraries have paid more attention to seek and collect resources than to integrate the resource scientifically, and this will lead to that, these precious documents cannot be fully and systematically revealed and utilized. At

Digital Scholarship in the Humanities © The Author(s) 2020. Published by Oxford University Press on behalf of EADH.
All rights reserved. For permissions, please email: journals.permissions@oup.com
doi:10.1093/llc/fqaa027

present, most documents had only finished digitalization or partial datalization work. However, the requirements of deep knowledge mining in documents data, providing visual analysis, and effectively supporting research process of historical humanities scholars had not been fully met. In this article, we introduce the local historical documents project in Shanghai Jiao Tong University (SJTU), which takes advantage of relevant techniques of digital humanities (DH) and fulfils in-depth analysis of intelligent documents data. According to the research needs, we proposed a data centralized framework with resources construction, database construction, and intelligent platform construction, to provide further research perspective, research environment, research support, and research experience for historic humanities scholars.

2 Introduction of DH

In era of digital network, academic collections in libraries become more and more similar. The unique academic, historic, and valuable special resources have become an important factor for the sustainable special collection strategy in libraries. These are most precious wealth of libraries and also the embodiment of the advantages and competitiveness of libraries. The competition of library resources in the future will be the competition of special collections with 'informal publications' as the core (Wanguo and Ying, 2018) .

Library stores a large number of digital special collection documents, but all documents data need to be interpreted by humanities scholars to have professional meaning. In process of historic humanities research, scholars usually face various difficulties. They have limited ability to obtain available information related to their own research projects from a large number of resources and are hard to deal with a large-scale data. The significance of DH support lies in deep mining and intelligent analysis of large scale texts, and the main object of DH research is humanities data resources.

DH emerged in the 1990s and accompanied with the changes in research methods of humanities scholars. The application of digital technology enriches the methodological system of humanities research, broadens the horizon of humanities research, changes the traditional humanities research environment, and more importantly, provides novel research methods, tools, and platforms for traditional humanities research.

The work by Manuel Perez-Garcia (2011, 2013, 2019) presents the most updated state of art of databases in DH. He summarized the application of new technologies, software coding, and computer analysis in the social sciences and humanities mainly in the field of economic history, and explains how the support of new technologies has helped historians to develop their research over the last few decades. Manuel Perez-Garcia proposed the application of both database and genealogical programmes for the southern Europe family studies as a methodological tool. He also designed a new multi-relational database to test the 'industrious revolution' hypothesis and present the 'vicarious consumption' theory. Kantabutra *et al.* (2014) have been the inventor of Intentionally Linked Entities, which is a database system for representing dynamic social networks, narrative geographic information, and general abstractions of reality.

Wachowicz and Owen (2013) implement to historical research as well as Manuel Perez-Garcia, and propose a knowledge space representation for helping construct formal and linguistic knowledge in geographically integrated history. Kaplan (2015) tries to represent big data research in DH as a structured research field and intends to draw a map for big data DH as big cultural datasets, digital culture, and digital experiences. Manovich (2015) presents a number of core concepts from data science that are relevant to digital art history and the use of quantitative methods to study any cultural artefacts or processes in general.

DH promotes the paradigm change of humanities research (Ling, 2018). At the same time, processing, experience, and achievements of humanities research on special collection resources can in turn promote the improvement of digital technology.

3 Collection and Collation of Local Historical Documents Resources

With increasing number of discoveries and collections of local historical documents, thousands of confused documents will bring more difficulties for users to use them. Scholars could utilize these resources better only through orderly organization. Effective use must be based on the premise of good documents collation. With abundant resources preservation, collation, and organizational experience, libraries have become an indispensable force in the process of collecting and collating local historical documents.

3.1 Introduction to the overall situation of local historical documents

Local historical documents, also known as folk historical documents, folk documents, and so on, mainly come from the daily life of local people. All documents are found and obtained by means of folk collection. These documents were not published or reorganized. This kind of characteristic is close to the nature of archives, they are words and other forms of materials produced in the course of people's daily activities. Their main forms include contract documents, litigation documents, village regulations, account books, diaries, letters, singing books, scripts, religious ritual books, prescriptions, daily miscellaneous books, etc., which cover a wide range of social, economic, political, and cultural fields of folklife (Zhengman, 2004). Scholars can get a glimpse of folk historical memory and restore the rich and colourful life of civil society.

SJTU began to engage in local historical documents collating and research in 2007. After 2012, the collected documents were handed over to the library for preservation, restoration, and collation. By 2018, more than 350,000 pieces of local historical documents had been collected (Fig. 1). These documents were mainly from Zhejiang, Anhui, Fujian, and Jiangxi, which are four provinces in China. And the collection had covered 36% of these four provinces (Fang *et al.*, 2015). At present, they were the most systematic documents group that could reflect the traditional society in Southeast China. According to the characteristics of documents from different sources, libraries adopted different measures of collation and preservation.

3.2 Digital construction of documents

In recent years, 'The trend of synchronization between historical research and digital era is becoming more and more obvious' (Hang, 2014). Digital resources have become an important basis for academic exchanges and research. As a basic work to provide raw resources for historical research, libraries have carried out many large-scale digital processing projects of local historical documents. With transforming the physical form into electronic form to store and use, libraries are responsible for long-term preservation, accelerate the popularization of humanistic knowledge, and provide better support for research of humanities scholars.

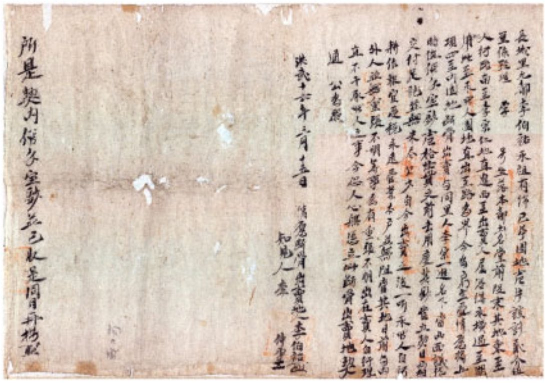

Fig. 1 An example of local historical documents

We divide the digitalization of local historical documents into two levels: long-term preservation level and data service level. The main difference between the two is that long-term preservation level will use higher scanning resolution, while data service level will take into account the speed of access and server pressure to adopt lower scanning resolution. According to the characteristics of different forms of documents to develop the corresponding technical standards and formats. The digitization has no technical obstacle itself, but it requires more effective onsite management. If documents damaged seriously, they need to be repaired first and then scanned (Baoguo *et al.*, 2017).

However, electronic documents will not enhance the use value, except for their ease of dissemination. DH is not a simple humanities digitalization, which contains a large number of new topics to be solved by humanities scholars and digital technicians. Turning resources into intelligent data, they can integrate into the scientific research process and support scientific research and innovation. This is common goal of the history departments and libraries.

4 Intelligent Data Construction of Local Historical Documents

The rapid development of big data brings out an important concept: smart data. In the field of DH, smart data can be understood as a kind of information that is meaningful to humanities scholars (Kobielus, 2016). Puschmann and Bastos (2015) put forward how the DH have embodied smart data and big data concepts and approaches, which demonstrate an emerging and significant change in terms of methodology.

Smart data has strong semantic expression ability and association ability. Zeng (2019) introduces a number of semantic enrichment methods and efforts that can be applied to libraries, archives, and museums (LAMs) data at various levels, aiming to support deeper and wider exploration and use of LAM data in DH research. It shifts the focus of big data from 'big' to the essence of data at the knowledge level, which can fully represent the semantic attributes and characteristics of data resources. By effectively organizing and extracting the data of local historical documents, using digital technology and combining with the knowledge of humanities scholars themselves, the data could match the needs of humanities research intelligently, and make researcher's judgement, decision-making and behaviour more wisely. Smart data has played and will continue to play a huge role in the field of DH.

4.1 Self-built metadata scheme

Metadata structure determines the way in which documents are retrieved and used. Good metadata construction will transform the use of local historical documents from 'reading' to 'analysis'. In construction of the local historical document database, how to establish a reasonable metadata structure and clarify internal relations of different documents, such as time sequence, geographical distribution, and interpersonal network, should be the issues that humanities scholars and technicians need to consider.

Taking contract documents in local historical documents as an example, library tried to combine the knowledge of archival science and library science, used special metadata design methods to extract resource characteristics and user needs, and designed a set of metadata specifications applicable to contract documents.

We investigated the demands of historical humanities scholars for resources and paid more attention to correlation between resources, people, and families. We analysed the document resources, extracted the resource attributes, and summarized three major attribute modules: External Physical Characteristics, Content Features, and Identity Recognition Features. Based on reusing of DC, 18 metadata elements were defined, including 4 custom elements (Jie *et al.*, 2017), as shown in the Table 1.

The ultimate goal of self-built metadata is to construct a local historical documents data service platform. After collating and cataloguing, we will form a relatively systematic and complete collection of resources, which is the basic data resources to support research and writing of humanities history.

4.2 Realizing multiple associations of resource data

Through metadata, descriptive text can be transformed into analysable data, and data use will be more inclusive and flexible. And metadata can implement contextual association, people and events correlation analysis, and relevance analysis with other documents. In order to clarify the internal relations of different documents, such as time series, geographical layout, event correlation, and interpersonal network, multiple related elements were set up in the specification of describing metadata.

4.2.1 *Resource physical association*

The file number element is used as the unique identity of each local historical document. The file number is fixed before and after digitalization.

4.2.2 *Chinese–Western calendar association*

Local historical documents use the Chinese year number to record the date, and it is hard for users to clearly understand local area-specific time. A Gregorian date is set to correspond to it, which could provide a perspective to observe the changes of local economy and society on the time axis. At same time, the events recorded in documents can be placed in the domestic and international position of the same period for comparison and calculation.

4.2.3 *Family association*

Local historical documents have accumulated many generations of documents of some rural families. There are close connections between different families and also inheritance. By setting the family element to record the administrative division and family information, we try to gather the documents of the same region and the same family together.

4.2.4 *Geographical association*

The combination of geographical element and family element can more clearly reflect the geographical relationship between the documents, helping researchers to further explore the geographic information of the resources (Xin *et al.*, 2018).

4.2.5 *Character association*

Most local historical documents record the civil behaviour of individuals, families, and organizations. Therefore, the element of the character had been set. And through the statistics of the identity information of the character, the family, social, and economic relations of the relevant persons can be related.

5 Intelligent DH Platform Construction

5.1 Consideration on platform construction

Bassett et al. (2017) argued some urgent questions, with the recent turn towards what has come to be called ‘platformisation’, that is the construction of a single digital system that acts as a technical monopoly within a particular sector, and it is certainly the case that the implications of machine learning infrastructures and their black-boxed techniques for sorting, classification, and ordering large amounts of data. Puschmann and Bastos (2015) compared two academic networking platforms, HASTAC and Hypotheses, to show the distinct ways in which they serve specific communities in the DH in different national and disciplinary contexts.

By investigating several DH platforms and sorting out the needs of historic humanities scholars, from the perspective of application, the intelligent DH platform should have the following characteristics:

5.1.1 *Retrieve and discover data resources*

Helping humanities scholars to obtain the information they need from a large number of local historical documents resources is the basic function of the platform. System can provide the retrieval of all the metadata fields and feed the retrieval results back to the scholars.

5.1.2 *Text analysis and statistics based on data mining*

Local historical documents have a lot of content to be revealed and reused, but they are all hidden inside the entity and need to be deeply explored and analysed. Based on the multiple associations of data, time span distribution, geographical distribution, and family distribution of the documents can be statistically analysed, and this information can also be used to perform association analysis and reveal the potential value of the documents.

5.1.3 *Preserve and manage scientific research data*

In process of research, humanities scholars will produce a lot of research data. These data are various in form, complex in variety, and numerous in number. Research data can be properly preserved and revealed, which can form a complete research context of scholars and their teams. Humanities scholars can set version numbers for their research outcomes based on time nodes to form a comparable and traceable outcome set.

5.1.4 *Analyse and display research data*

Traditional data analysis and visualization tools need to use more professional knowledge or require higher learning costs, develop an easy-to-use toolset in the platform, help researchers analyse the research data and sort out the research results, and present the results in a more intuitive and visual way for analysis and discussion by the research team members, which in turn promotes the development of the entire research process.

5.1.5 *Open data service support*

Without data interoperation, DH platform will lack driving force of development, and the value of data will not be fully utilized. It is an effective way to improve the vitality of resource by sorting out the underlying data to obtain data catalogues and providing data support services at database level, data interface level, or data application level. At the same time, researchers can also add and modify data under the data review process of the platform to ensure the accuracy and completeness of the data.

5.1.6 *Application of artificial intelligence*

The application of artificial intelligence in the field of DH platform can be tried from the following two aspects: intelligent recommendation of information resources through user behaviour data analysis and resource learning algorithm with supervised feedback. Scholars mark resources in their research process, which can more accurately index the attributes and labels of resources, and also can expand the dimension of data and the association between data.

5.1.7 *Community communication*

Users are accustomed to sharing and exchanging information in a community-based manner. All kinds of resource service platforms have certain community functions. For example, book reviews, ratings, labels, etc. of academic resources. The support of community communication can enhance the communication between users of the platform and enhance the utilization and activity of the system.

5.2 Platform architecture design

Based on the above considerations, the architecture of the intelligent DH system platform is shown in Fig. 2.

5.2.1 *Resource collection*

Get data from different data sources. There are mainly four types of data sources: Library's Own Special Reservation, Purchase of Commercial Databases, Open Access Data, and Co-Construction Project Data.

5.2.2 *Resource storage*

Resource data needs to be translated, removed duplicates, cleaned, and combined because of the different data source, different ways of obtaining data, different metadata coding, and different resource description. According to our metadata scheme of local historical documents, the metadata was coded uniformly and stored in the metadata repository. These data will be preserved for a long time. In the future, we can identify metadata entities, build knowledge ontology, and publish linked data to realize the purpose of automatic association, reuse, and sharing with external resources.

5.2.3 Data processing

Apply multiple DH methods to data processing. Based on the metadata repository, several data processing and intelligence analysis technology can be used. For example, index establishment, image segmentation, character recognition, text analysis, data mining, and so on. Many hided information will be shown to the higher level after data processing step.

5.2.4 *Resource utilization*

The basis for data utilization is access control, and any data call should be made under access control protection. On this basis, the system can provide various types of services for humanities scholars, such as one-stop search service, full-field search, and acquisition services, and provide recommended resources. Another important aspect of data utilization is data sharing. Through data interface, the data is opened to third-party systems for use, improving system interoperability.

5.3 Platform function module design

The function of the DH platform is not only storage and retrieval of data, the main purpose is to make the resource truly usable by humanities scholars, and become useful data, fresh data, and intelligent data. And on this basis, it provides a better research environment and research support for humanities scholar, helping them to reorganize knowledge, discover problems, bring new research perspectives, and provide decision-making basis for future works. The functions of the platform are divided and modularized from humanist dimension, library dimension, and DH technology dimension, as shown in Fig. 3.

5.3.1 *Scholar application module*

It is a functional module based on the operation of humanities scholars (Fig. 4). There are three main functions as follow.

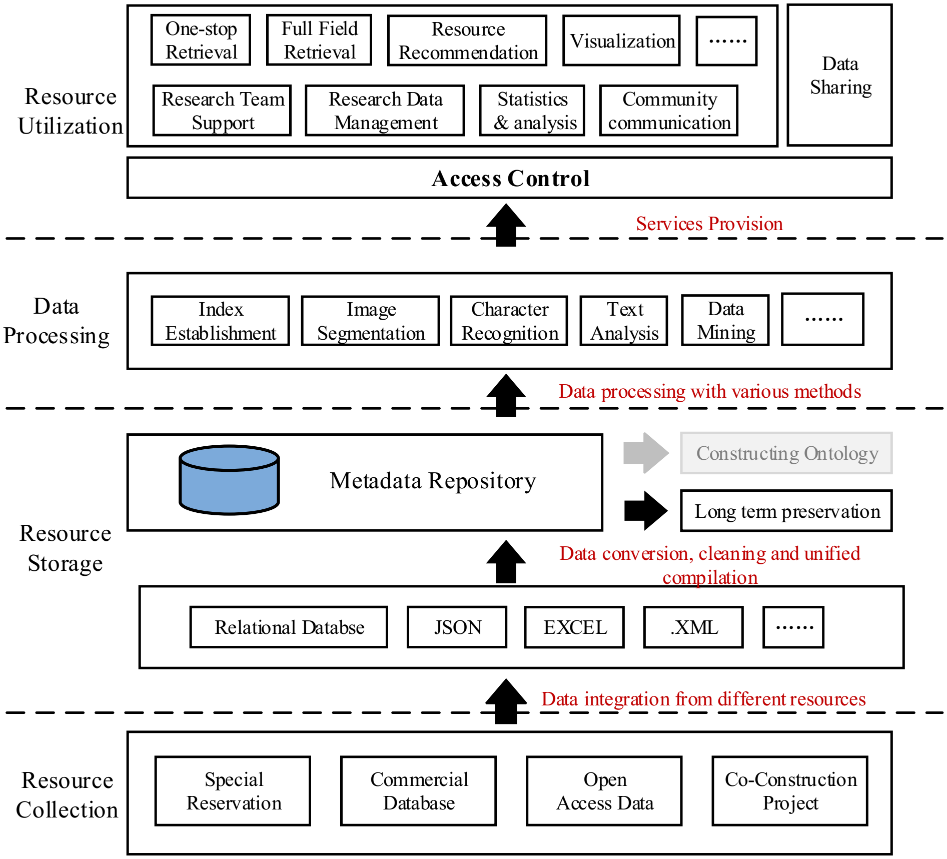

Fig. 2 Platform architecture

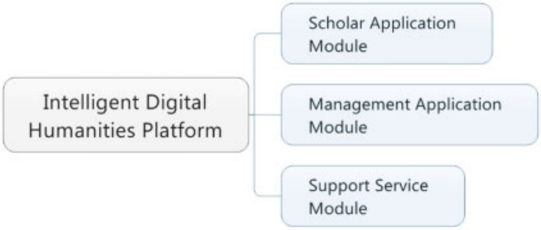
Intelligent Digital Humanities Platform
Scholar Application Module
Management Application Module
Support Service Module

Fig. 3 Function module design of platform

5.3.1.1 *Resource utilization.* Scholars can search, view, and download the documents, and can also store the documents in their research document management module for centralized manage and use. When scholars find that the documents information is missing, wrong, or can provide relevant annotations, they can submit feedback to the system to help the system improve the quality and utilization of the documents.

5.3.1.2 *Research projects management.* Scholars can set up their own research projects on the platform and manage them. They can research in the name of individuals or teams. They can manage the research documents, such as modify the metadata, add data associations, add tags to data, and so on. These contents can be selected to be visible only to themselves, visible to team members, or visible to public. Scholars can also manage the research outcomes of the project, and they can choose to disclose or not disclose these outcomes.

5.3.1.3 *Personal information management.* Scholars can manage their personal information.

5.3.2 *Management application module*

This module is set up to manage the whole platform (Fig. 5). Its functions mainly include the following.

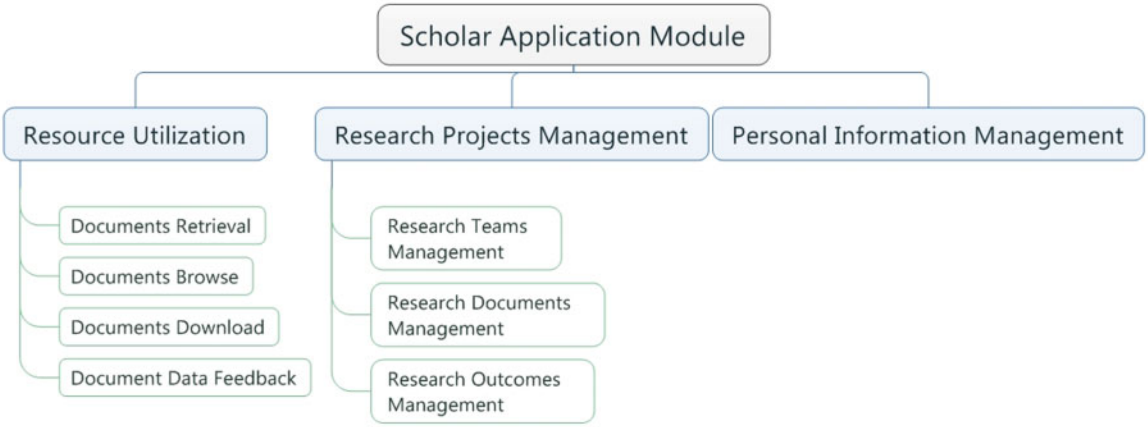
Scholar Application Module
Resource Utilization
Research Projects Management
Personal Information Management
Documents Retrieval
Documents Browse
Documents Download
Document Data Feedback
Research Teams Management
Research Documents Management
Research Outcomes Management

Fig. 4 Scholar application module

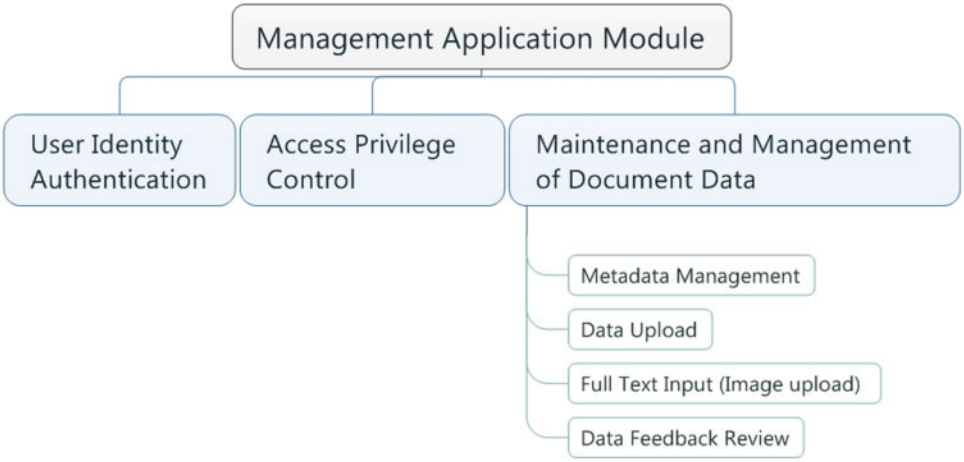
Management Application Module
User Identity Authentication
Access Privilege Control
Maintenance and Management of Document Data
Metadata Management
Data Upload
Full Text Input (Image upload)
Data Feedback Review

Fig. 5 Management application module

5.3.2.1 *User identity authentication.* All use of platform must be based on user identity authentication. After log in, system will determine whether user has right to use the platform.

5.3.2.2 *Access privilege control.* Setting access restrictions on document resources. For example, only allowing intra-school access or limit IP segment access.

5.3.2.3 *Maintenance and management of document data.* Managing and maintaining core data of the platform, such as metadata management, data update, and upload, full-text input, review data feedback submitted by scholars, and update corresponding content of the database after confirmation.

5.4.3 *Support service module*

This module serves the DH platform itself and assists users to make full use of platform functions (Fig. 6).

These functions include collect and analysis user behaviour data, personalized resource recommendation, statistics of various documents data, providing analysis and visualization tools, and effective analysis and visualize scholars research data and research

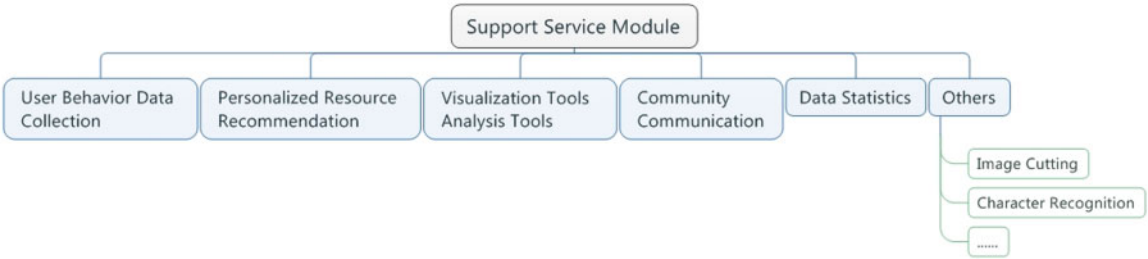

Fig. 6 Support service module

Table 1 Local historical documents elements set

Attributes	ID	Element	Reuse	Element description
Content features	1	Title	dc:title	Title of local historical documents
	2	Character	custom element	All the important persons and their roles in the original text
	3	Family	custom element	Administrative divisions and family information
	4	Cause	dc:description	Events or acts recorded in document, such as litigation, trading, tax-paid behaviour, etc.
	5	Geographic information	dc:description	Geographical information in document content
	6	Area code	dc:spatial	Code information of a certain region obtained by querying the codebook
	7	Document date	dc:data	The date (in Chinese year number) that the document was generated
	8	Gregorian date	dc:data	Gregorian calendar year, corresponding to the Chinese year number
	9	Object	custom element	The object of the transaction, land, houses, goods, rights, etc.
	10	Amount	custom element	Amount due to transfer of property rights
Physical features	11	Number of pages	dc:extent	Quantity of documents
	12	Size	dc:extent	Document size
	13	Material	dc:format	Materials of documents
	14	Note	dc:description	Other important information about the physical form of local documents
Identity recognition features	15	Type	dc:type	Types of local documents
	16	Identifier	dc:identifier	The serial number naturally generated when the document is recorded
	17	File number	dc:identifier	Unique number for each document
	18	Language	dc:language	Language information

results, community communication and some other features.

6 Conclusion

DH can provide a new research paradigm and a new academic perspective for the traditional humanities and is the systematic extension of library humanities service. Library will play an important role in resource organization, resource preservation (management), technology support, and platform service in the process of subject research support.

In this article, we expound how to collect, organize, and utilize special collection resources, such as local historical documents in SJTU, by DH thoughts, and propose an intelligent DH data framework to effectively support humanities research. We hope that these useful works will provide some inspiration and reference for institutions and individuals who use the same type of special collection resources. At the same time, in the road of exploration and practice in DH, libraries should not only do important job in protection, development and open utilization of special collection resources, but also constantly expand research ideas, enhance service, and innovation capabilities with various data innovation efforts, and seize opportunities to enhance academic and social status of libraries.

Acknowledgements

This work was supported by the Social Science Planning of Shanghai (Grant No. 2018BTQ002), and the authors wish to thank the reviewers of this paper who offered a number of useful recommendations.

References

Baoguo, Z., Xiaohua, S., and Xin, W. (2017). Research of quality control in the process of the large-scale book digitization. *Research on Library Science*, 2017(4): 51–5.

Bassett, C., Berry, D. M., Fazi, M. B., Pay, J., and Roberts, B. (2017). Critical digital humanities and machine-learning. DH 2017. https://dh2017.adho.org/abstracts/509/509.pdf (accessed 14 June 2019).

Fang, L., Jin, C., and Xin, W. (2015). Planning and practice for the historical documents digitalization in recent library holding of Shanghai JiaoTong University. *Journal of Academic Library*, **33**(2): 77–83.

Hang, L. (2014). Digitization of historical documents calls for active involvement of scholars. *Chinese Social Science Today*, 2 April 2014 (A02).

Jie, Z., Fang, L., and Meng, T. (2017). The design and use of metadata application profile for local historical property contracts. *Library and Information Service*, **61**(8): 106–11.

Kantabutra, V., Owens, J. B., and Crespo-Solana, A. (2014). Intentionally-linked entities: a better database system for representing dynamic social networks, narrative geographic information, and general abstractions of reality. In Crespo Solana, A. (ed.), *Spatio-temporal Narratives: HGIS and the Study of Trading Networks (1500 - 1800).* Cambridge Scholars Press, pp. 56–78. http://www.c-s-p.org. (accessed 16 June 2019)

Kaplan, F. (2015). A map for big data research in digital humanities. *Frontiers in Digital Humanities*, **2**: 1.

Kobielus, J. (2016). The evolution of big data to smart data. http://www.dataversity.net/big-data-smart-data-big-drivers-smart-decisionmaking/ (accessed 12 June 2019).

Ling, Z. (2018). Review of the library and digital humanities forum. *Journal of Academic Library*, **36**(2): 5–10.

Manovich, L. (2015). Data science and digital art history. *International Journal for Digital Art History*, 2015(1): 13–33.

Perez-Garcia, M. (2011). New technologies applied to family history: a particular case of Southern Europe in the eighteenth century. *Journal of Family History*, **36**(3): 248–62.

Perez-Garcia, M. (2013). *Vicarious Consumers: Trans-national Meetings between the West and East in the Mediterranean World (1730–1808).* London: Routledge.

Perez-Garcia, M. (2019). Consumption of Chinese goods in southwestern Europe: a multi-relational database and the vicarious consumption theory as alternative model to the industrious revolution (eighteenth century). *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, **52**: 15–36.

Puschmann, C. and Bastos, M. (2015). How digital are the digital humanities? An analysis of two scholarly blogging platforms. *PLoS One*, **10**(2): e0115035.

Wachowicz, M. and Owens, J. B. (2013). The role of knowledge spaces in geographically-integratedhistory. In von Lünen, A. and Travis, C. (eds.), *History and GIS: Epistemologies, Considerations and Reflections*, Chapter 9. Dordrecht, Heidelberg, New York, London: Springer, pp. 127–44.

Wanguo, L. and Ying, H. (2018). A summary of the academic seminar in digital resource construction and knowledge service in 2017. *Journal of Academic Library*, **36**(1): 12–17.

Xin, W., Jie, Z., and Meng, T. (2018). Study of pathway of information organization and display in deed document from geographic aspect. *New Century Library*, 2018(4): 55–9.

Zeng, M. L. (2019). Semantic enrichment for enhancing LAM data and supporting digital humanities. *El profesional de la información*, **28**(1):1–35.

Zhengman, Z. (2004). Folk historical documents and cultural inheritance research. *Southeast Academic*, **S1**: 293–6.