The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/0264-0473.htm

A Chinese ancient book digital humanities research platform to support digital humanities research

Chih-Ming Chen
Graduate Institute of Library, Information and Archival Studies, National Chengchi University, Taipei City, Taiwan, and

Chung Chang
Research Center for Chinese Cultural Subjectivity, National Chengchi University, Taipei City, Taiwan

Received 26 October 2018
Revised 28 December 2018
11 February 2019
Accepted 12 March 2019

Abstract

Purpose – With the rapid development of digital humanities, some digital humanities platforms have been successfully developed to support digital humanities research for humanists. However, most of them have still not provided a friendly digital reading environment and practicable social network analysis tool to support humanists on interpreting texts and exploring characters' social network relationships. Moreover, the advancement of digitization technologies for the retrieval and use of Chinese ancient books is arising an unprecedented challenge and opportunity. For these reasons, this paper aims to present a Chinese ancient books digital humanities research platform (CABDHRP) to support historical China studies. In addition to providing digital archives, digital reading, basic search and advanced search functions for Chinese ancient books, this platform still provides two novel functions that can more effectively support digital humanities research, including an automatic text annotation system (ATAS) for interpreting texts and a character social network relationship map tool (CSNRMT) for exploring characters' social network relationships.

Design/methodology/approach – This study adopted DSpace, an open-source institutional repository system, to serve as a digital archives system for archiving scanned images, metadata, and full texts to develop the CABDHRP for supporting digital humanities (DH) research. Moreover, the ATAS developed in the CABDHRP used the Node.js framework to implement the system's front- and back-end services, as well as application programming interfaces (APIs) provided by different databases, such as China Biographical Database (CBDB) and TGAZ, used to retrieve the useful linked data (LD) sources for interpreting ancient texts. Also, Neo4j which is an open-source graph database management system was used to implement the CSNRMT of the CABDHRP. Finally, JavaScript and jQuery were applied to develop a monitoring program embedded in the CABDHRP to record the use processes from humanists based on xAPI (experience API). To understand the research participants' perception when interpreting the historical texts and characters' social network relationships with the support of ATAS and CSNRMT, semi-structured interviews with 21 research participants were conducted.

Findings – An ATAS embedded in the reading interface of CABDHRP can collect resources from different databases through LD for automatically annotating ancient texts to support digital humanities research. It allows the humanists to refer to resources from diverse databases when interpreting ancient texts, as well as provides a friendly text annotation reader for humanists to interpret ancient text through reading. Additionally, the CSNRMT provided by the CABDHRP can semi-automatically identify characters' names based on Chinese word segmentation technology and humanists' support to confirm and analyze characters' social network relationships from Chinese ancient books based on visualizing characters' social networks as a knowledge graph. The CABDHRP not only can stimulate humanists to explore new viewpoints in a humanistic research, but also can promote the public to emerge the learning interest and awareness of Chinese ancient books.

Originality/value – This study proposed a novel CABDHRP that provides the advanced features, including the automatic word segmentation of Chinese text, automatic Chinese text annotation, semi-automatic character social network analysis and user behavior analysis, that are different from other existed

The Electronic Library
Vol. 37 No. 2, 2019
pp. 314-336
© Emerald Publishing Limited
0264-0473
DOI 10.1108/EL-10-2018-0213

digital humanities platforms. Currently, there is no this kind of digital humanities platform developed for humanists to support digital humanities research.

Keywords Digital archives, Digital humanities, Automatic text annotation, Character social networks analysis, Chinese ancient books

Paper type Technical paper

1. Introduction

Digital humanities (DH) is work at the intersection of digital technology and humanities disciplines (Drucker *et al.*, 2014). It includes the systematic use of digital contents and tools in the humanities disciplines and makes possible new kinds of teaching and research. In recent years, DH has received particular attention in both the humanities and computer science fields. Particularly, academic libraries, such as Research Libraries UK and National Central Library of Taiwan, have started to develop infrastructure to support DH research, as well as close collaboration between librarians and researchers is disclosing new opportunities for librarians to actively collaborate in DH research (Cassella, 2017; Kamposiori, 2017). DH is an interdisciplinary field and is concerned with the intersection of computer science, knowledge management and a wide range of humanities disciplines (Steiner *et al.*, 2014). Humanities research involves the analysis, comprehension, production and sharing of a huge amount of digital sources. Moreover, humanists in the DH field rely heavily on using digital tools, such as text mining (Widlöcher *et al.*, 2015), annotation (Chen and Tsay, 2017; Sato *et al.*, 2016), social networks analysis (Uboldi *et al.*, 2013), geographic information systems (GIS) (Kallaher and Gamble, 2017) and natural language processing (NLP) technologies (Brooke *et al.*, 2015) to analyze data from digital sources to extract useful information, clues and find new knowledge. However, although a lot of digital tools have already been developed for supporting DH research, many of these tools do not properly fit the needs of the humanities community due to technical difficulties or scientific objectives (Picca and Egloff, 2017). Furthermore, different types of humanists have considerable differences in their knowledge of the collections, requiring varying levels of support, and every individual humanist has their own particular interests and priorities.

In recent years, Chinese ancient books are presenting an unprecedented challenge and opportunity due to the rise of international Chinese study around the world. To increase research efficiency and production, there have been a lot of international Chinese studies to explore issues about Chinese culture, religion, economy, and politics through Chinese ancient books or literatures with the support of DH tools. For example, Hsiang *et al.* (2012) was inspired by Imperial Court documents from the Ming and Qing dynasties because a memorial often quotes earlier Imperial Edicts as the directives for the actions being reported, while an Imperial Edict may cite earlier memorials as reasons for issuing the decree. Their study, thus, explored the citation relation among the Imperial Edicts and the memorials through the developed DH tool based on Taiwan-related Imperial Court documents from the Ming and Qing dynasties archived in the Taiwan History Digital Library (THDL) (http://thdl.ntu.edu.tw/index.html). Moreover, land deed research played a unique role in the modern Chinese history. Land deeds were the only proof of ownership in pre-1900 Taiwan. They are indispensable for the studies of Taiwan's social, anthropological and economic evolution. Based on a large amount of historical land deeds, Chen *et al.* (2013) developed a DH tool to discover two important relations – successive transactions and allotment agreements involving the same property – as well as presented a notion of land transitivity graph to capture the transitivity embedded in these land deeds' transactions. Also, Chan *et al.* (2015) presented a spatiotemporal model based on a geographical information system, spatial statistics and Shannon entropy for elucidating the evolution of market towns spanning and examining the spatial and hierarchical relationships of market towns from before 1550 to 1980 through a large quantity of historical Chinese ancient books.

To provide more convenient and user-friendly research environments for humanists engaging in international Chinese study, increasing effort has been made not only to make digital contents and tools available, but also to create complete virtual research environments (VREs) that provide interpretative frameworks for making sense of cultural artefacts (Steiner *et al.*, 2014). Such VREs support conceptualizing, visualizing and analyzing information, as well as humanists can collaboratively work on it. VREs usually do not consist of one monolithic technology, but cover a collection of tools assembled in one place to assist research tasks and processes (Steiner *et al.*, 2014). This study surveyed several current DH research platforms in Taiwan and around the world – including the CULTURA system (Steiner *et al.*, 2014), CText (https://ctext.org/zh), the Scripta Sinica database (http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm), TextGrid (Neuroth *et al.*, 2011), THDL (http://thdl.ntu.edu.tw/index.html), the CBETA Research Platform (CBETA-RP) (http://cbeta-rp.dila.edu.tw/) and the MARKUS semi-automatic text annotation system (http://dh.chinese-empires.eu/markus/beta/) – and compared their features to develop a more practicable DH research platform to support research. Among the developed DH platforms, Steiner *et al.* (2014) presented the CULTURA system and services and the two collections that have been used for testing and deploying the DH research environment. The services of the CULTURA system include personalized search tools, faceted search tools, annotators, social network visualization tools, and recommenders. Additionally, CText is a Chinese ancient text database, which contains the full text of various Chinese texts of philosophical, historical or linguistic interest from the pre-Qin era through to the Han Dynasty and beyond and makes the best possible use of modern information technology to aid humanities research. The Scripta Sinica database contains almost all of the important Chinese classics, especially those related to Chinese history. This database provides scholars, students and the general public with an excellent full-text database and search engine for the study of Chinese history and culture. Furthermore, TextGrid is a VRE that provides services for supporting researchers in the arts and humanities and tools for the analysis of text data and supports the curation of research data by means of grid technology (Neuroth *et al.*, 2011). The THDL covers about 80 percent of all primary Chinese historical materials about Taiwan before 1895. The primary functions of the THDL for supporting DH research include full-text search, techniques and interfaces for classifying and exploring a query result as a sub-collection, term frequency analysis and referential tools (Chen *et al.*, 2007). The CBETA-RP provides a user-friendly online reading interface with complete content and handy digital tools. Full-text search, dictionaries look up, person and place references, the statistics of term and relevant bibliography are also provided in the CBETA-RP, whereas MARKUS is a famous semi-automatic text annotation system and supports online text reading. However, most of the DH platforms currently developed have still not provided complete data analysis functions to support DH research. This study, thus, presents a Chinese ancient books digital humanities research platform (CABDHRP) which provides basic data analysis functions, such as full-text search with a single word or double words, term frequency analysis and post-query classification, as well as advanced data analysis functions, such as automatic text annotation, textual analysis, character social network analysis and user behavior analysis, to support DH research.

Table I shows the function comparison of the seven DH research platforms. It was discovered that the Scripta Sinica database presents the richest full-text data collection, the CBETA-RP merely includes Buddhist texts and the MARKUS semi-automatic text

Table I. Function comparison of different digital humanities research platforms

Function comparison of different DH research platforms	CABDHRP	CULTURA	CText	Scripta Sinica database	THDL	CBETA research platform	MARKUS semi-automatic text annotation system
Full-text data collection	○	○	○	○	○	○	X
Automatic segmentation of Chinese words	○	X	X	X	X	X	X
Automatic annotation	○	○	○	X	○	X	○
Term search	○	○	○	○	○	○	○
External reference	○	○	○	○	○	○	○
Word frequency statistics	○	○	○	X	○	○	X
Bookmark note	X	X	X	○	○	X	X
Social network visualization tool	○	○	○	X	X	X	○
Recommender	X	○	X	X	X	X	X
Recording use process	○	X	X	X	X	X	X
Open API	X	X	○	X	X	○	○

Notes: “O” means that a feature is available; “X” means that a feature is not available

annotation system does not show the full-text data collection function. Currently, the CABDHRP merely contains 110 of the Ming Dynasty writings, but the archived contents are rapidly increasing with time. The CULTURA system consists of multiple distinct services, including personalized search tools, faceted search tools, annotators, social network visualization tools and recommenders (Steiner *et al.*, 2014). Besides the CABDHRP, the other six DH research platforms lack automatic segmentation of Chinese words; the automatic annotation function exists in the CABDHRP, CULTURA, CText, THDL and MARKUS, and each system presents the search function and external reference function. In terms of external reference, the Scripta Sinica database could connect to Qing Officials Query System, THDL presents variant databases, CBETA-RP merely includes Buddhist text databases and MARKUS shows the most external reference resources (Wikipedia, CBDB, TGAZ and ZDict). Only CABDHRP, CULTURA, CText, THDL and CBETA-RP show the function of word frequency statistics; only THDL presents the function of bookmark notes. The advanced features of the proposed CABDHRP that contain the automatic word segmentation of Chinese text, automatic Chinese text annotation based on LD, semi-automatic social network analysis of human–machine interaction and user behavior analysis based on recoding use process of humanists are different from the other six existing DH research platforms. Only CText, CBETA and MARKUS present the function of open API that can share their DH tool functions or archived contents with other DH research platforms. The advanced features in the proposed CABDHRP will be detailed in Section 3.2.

2. Literature review

2.1 Current states, future challenges and development trends of digital humanities

DH, formally known as *humanities computing*, is born of the encounter between traditional humanities and computational methods (Burdick *et al.*, 2012). Wang and Inaba (2009) indicated that there is no clear sub-discipline in DH, and DH is still expanding its research domain. Tang *et al.* (2017) also indicated that DH continues to expand, has become more inclusive and that its knowledge is integrated. Therón and Wandl-Vogt (2016) argued that the current key issues of DH involve the inclusion of citizens in the creation and consumption of the cultural resources offered, the volume and complexity of data sets, and available infrastructures. Cassella (2017) indicated that a very special relationship is developing between DH and academic library digital repositories where humanists are making more and more massive use of digital library collections to study and to conduct research.

Kaplan (2015) proposed three challenging issues on DH research derived from big data generated by a huge amount of digital contents. The first challenge is how to focus on the processing and interpretations of large cultural data sets, the second is how to concern digital culture at large, and the third is how to deal with the experience of big data. Moreover, many previous studies on DH have focused on the textual contents; however, new work that develops digital tools to analyze the visual layout and content of books and manuscripts based on computer vision techniques is emerging (Rushmeier *et al.*, 2015). Apparently, challenges and increasing research areas identified in DH include how to incorporate the different types of digital resources. In addition, Therón and Wandl-Vogt (2016) also indicated that the DH field is currently facing the challenge of proposing frameworks and systems that can be standardized for different disciplines with similar but heterogeneous data sets. Therefore, developing a DH research system that can simultaneously archive digital contents from different disciplines and heterogeneous data sets and perform data analysis by using a set of commonly used and standardized digital tools is definitely needed.

Upadhyay and Upadhyay (2017) claimed that DH needs to increase the development and demand in the field of computational science, social network theory and analysis, digital literacy, computational pedagogy, computational literacy, machine and deep learning, neuro and cognitive learning and analysis in the future. It is obvious that DH is an emerging and interdisciplinary research field aiming at enhancing and redefining traditional humanities scholarship through digital means.

2.2 Digital tools developed for supporting digital humanities research

DH aims to use a digital-based revolutionary new way to carry out enhanced forms of humanities research more effectively and efficiently. Martin-Rodilla and Gonzalez-Perez (2016) classified already existing developed works of digital tools that support DH research as manual, semi-automatic and automatic approaches. In the manual approaches, humanists performed a manual data analysis by using digital tools. In the semi-automatic approaches, digital tools offered suggestions based on data analysis results for humanists to further perform data analysis procedures. In the automatic approaches, digital tools automatically performed data analysis based on the computer algorithms for humanists to explore new clues or knowledge.

Currently, in the manual approaches, there have been some existing systems for manually annotating historical documents that provide functions to append information to historical iconographic or text documents to support DH research, such as SMART-GS (Hashimoto, 2014) and the collaborative annotation system (CAS) (Chen and Tsay, 2017). Chen and Tsay (2017) proposed a novel CAS with four types of multimedia annotations, including text, picture, voice and video annotations, which can be embedded into any HTML web page to enable humanists to collaboratively add and manage annotations on these web pages and provide a shared mechanism for discussing the shared annotations from humanists. Their study indicated that the proposed CAS has high potential in the applications of DH research, including discovering useful knowledge from a large amount of annotations on historical texts based on data mining technologies, exploring a lot of user-interactive data generated in the collaborative annotation process based on social networks analysis and recording the operation processes of humanities scholars on the CAS to evaluate humanities scholars' research behaviors. Moreover, Sato *et al.* (2016) also presented a web-based prototype system for collaboratively making annotations on historical documents by multiple humanities researchers who are geographically separated from each other. The proposed system can support multiple users making annotations to the same document simultaneously, as well as providing suggestions of annotations by using information, such as existing annotation strings and the surrounding words in the text. Also, Haslhofer *et al.* (2013) indicated that semantic tagging is a collaborative process in which a user selects and associates Web resources drawn from a knowledge context. Their study applied the semantic tagging technique in the specific context of online historical maps and allowed humanists to annotate and tag them. Compared to the label-based tagging technique, the semantic tagging technique not only does not affect the tag type and category distributions, but also leads to higher user satisfaction.

In the semi-automatic approaches, Picca and Egloff (2017) presented a new Python library called as the DH ToolKit (DHTK) to provide a fast and intuitive tool to exploit the semantic knowledge resources from linked open data (LOD). The key features of the DHTK include a DH orientation, ease of use, modularity, efficiency, extensibility and documentation. Uboldi *et al.* (2013) designed the Knot, which is a digital tool for exploring historical social networks, to provide humanists with an environment for exploring multi-dimensional and heterogeneous data, allowing them to discover and create explicit and implicit relationships between people, places and events. Moreover, Jackson (2017) used social network analysis to identify an additional role played by Duncan II Earl of Fife and applied the network density model to identify opinion leaders in medieval Scotland based on the People of Medieval Scotland database. Also, Brooke *et al.* (2015) introduced a software tool, GutenTag, which is aimed at giving literary researchers direct access to NLP techniques for the analysis of texts in the Project Gutenberg corpus. Harris *et al.* (2014) presented the anatomy of Samtla, which is an online integrated research environment designed in collaboration with historians and linguists, to facilitate the study of digitized texts written in any language. Samtla is fundamentally different from standard text search/mining systems which rely on the bag-of-words representation of text. In contrast, Samtla emphasizes the retrieval and discovery of fuzzy text patterns/motifs which are of critical importance to DH research. Furthermore, Luczak-Roesch *et al.* (2018) presented a novel tool for DH that leverages temporal data mining, network science, and visual analytics. Their preliminary analytical results show that this approach facilitates a new collaborative methodological practice that is a hybrid of close and distant reading. Pastorelli (2017) applied a hierarchical ascendant clustering algorithm to group the manuscripts of the Old Latin witnesses of the Gospel of John into clusters presented with the calculation of distances between readings, then between witnesses to enable the construction of trees illustrating the textual taxonomy obtained.

In the automatic approaches, Yohan *et al.* (2014) proposed a rule-based named entity recognition and classification system which can automatically recognize named entities in text and assign them with the appropriate categories for the Telugu language. Additionally, Kestemont and Gussem (2017) applied a layered neural network architecture from the field of deep representation learning to automatically solve two sequence tagging tasks, including part-of-speech tagging and lemmatization. The result of their study shows acceptable performance, which is on par with previously reported studies. Although there have been a lot of digital tools to be successfully developed to support DH research, Harris *et al.* (2014) argued that there is still a huge gap between what humanists actually want and what digital tools can do in terms of functionality and usability.

3. The developed CABDHRP

3.1 System architecture

Widlöcher *et al.* (2015) claimed that developing an integrated environment to support DH research is needed. Their study emphasized that the reading interface, information retrieval and data mining should not be regarded as totally separate tasks when developing an integrated environment to support DH research. Based on the perspective, this study developed a CABDHRP that includes three main parts, including a digital reading interface module, a basic information retrieval and analysis function module and an advanced text and social network mining module to support DH research. This study adopted DSpace (Chen *et al.*, 2012), an open-source institutional repository system, to implement the CABDHRP embedded with basic and emerging digital tools for supporting DH research. Although DSpace has sufficient functionalities for archiving, searching and managing different types of digital contents, the predefined metadata schema Dublin Core (DC) needs to be revised to archive digital targets, such as Chinese ancient books, maps, and music. Additionally, the user interface of DSpace for reading digital contents cannot satisfy humanities scholar needs due to a very primitive and simple interface. Therefore, the developed CABDHRP based on DSpace has to be implemented a user-friendly reader that can simultaneously display the scanned image of Chinese ancient text and the corresponding full-text and be integrated with many useful reading functions, such as automatic word segmentation, automatic text annotation and character social network analysis, on the reader. These functions are detailed in a later section. Currently, the CABDHRP has archived 110 Ming Dynasty writings with scanned images, metadata and full texts. To support DH research using Ming Dynasty writings in depth, having a larger body of Ming Dynasty writings will undoubtedly provide benefits in exploring potential research issues to the humanists who are interested in international Chinese study. Figure 1 shows the system architecture of the proposed CABDHRP in this study. The components of the system are explained in detail as follows:

- *Digital reading interface module*: The digital reading interface module simultaneously contains an automatic text annotation tool and a manual annotation tool, but the manual annotation tool has still not been integrated into the CABDHRP so far. An automatic annotation tool that can collect resources from different databases through LD for automatically annotating Chinese ancient texts was developed in this study to support DH research. It allows humanists to refer to resources from diverse databases when interpreting Chinese ancient texts, as well as provides a user-friendly text annotation reader for humanists interpreting ancient text through reading. Moreover, this study adopted asynchronous JavaScript (Ajax) and the model-view-controller (MVC) framework to implement a manual text annotation tool with four types of multimedia annotations, including text, picture, voice and video annotations, for knowledge creating, archiving and sharing services to support DH. The manual text annotation tool has been successfully applied in digital curation, crowdsourcing and DH because of existing strong relations among them (Chen and Tsay, 2017). Specifically, in addition to annotating a digital text, this tool also offers photo annotation functions that allow a user to give various types of annotations on a photo, such as an ancient map.
- *Basic information retrieval and analysis function module*: The basic information retrieval and analysis function module contains the term frequency analysis tool, geographic information analysis tool, intelligent search tool and user behavior recording tool. For textual corpus studied in DH, a search tool is an obvious necessity. The term frequency analysis tool can help humanists to observe the distribution state of a term searched from a collection. Moreover, spatial analysis has become one of the primary focus areas for many different types of humanities scholars. The geographic information analysis tool

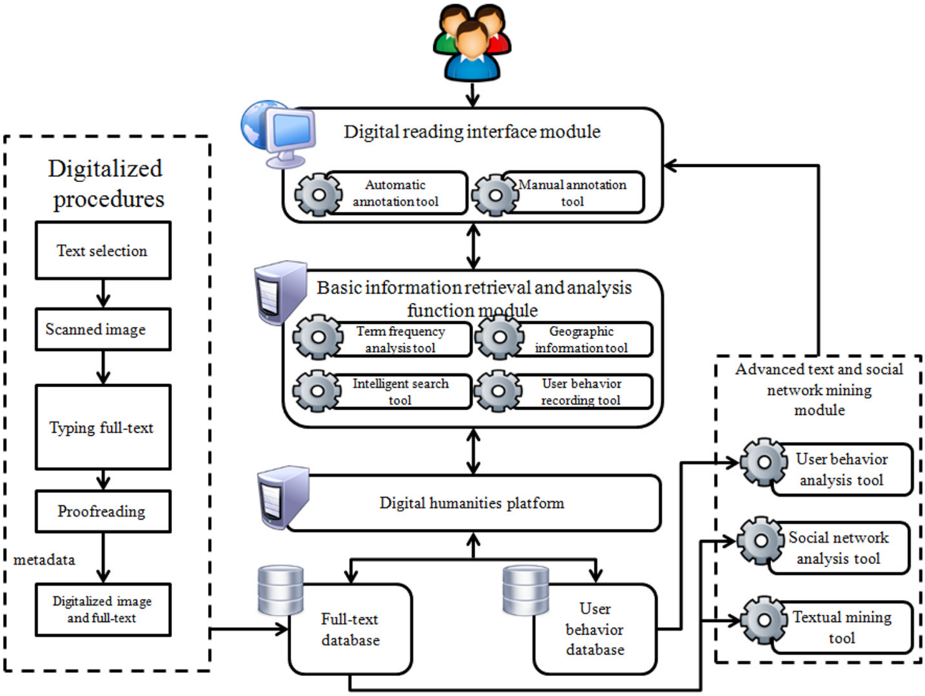

Digitalized procedures
Text selection
Scanned image
Typing full-text
Proofreading
metadata
Digitalized image and full-text
Digital reading interface module
Automatic annotation tool
Manual annotation tool
Basic information retrieval and analysis function module
Term frequency analysis tool
Geographic information tool
Intelligent search tool
User behavior recording tool
Digital humanities platform
Full-text database
User behavior database
Advanced text and social network mining module
User behavior analysis tool
Social network analysis tool
Textual mining tool

Figure 1.
The system architecture of the developed CABDHRP

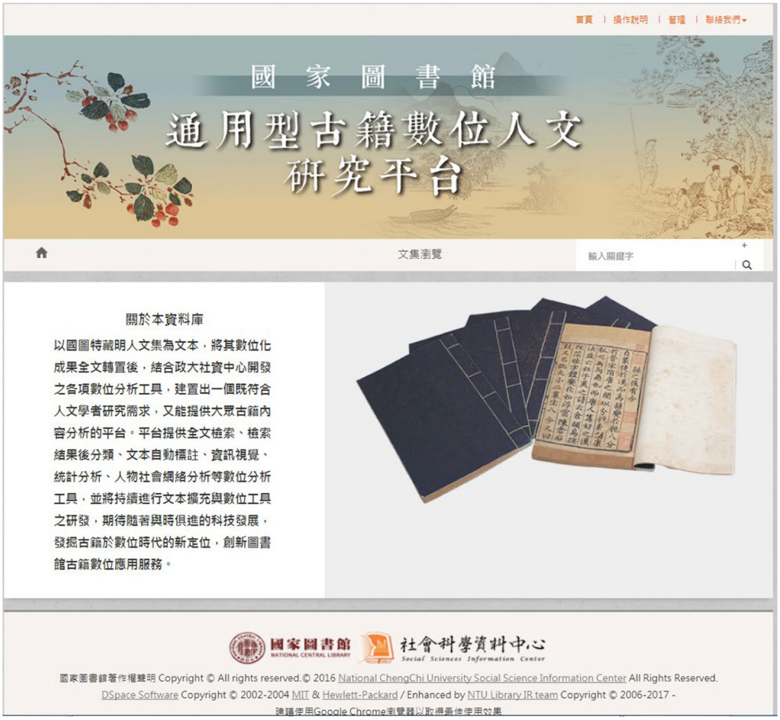
aims to interpret data through spatial relationships. This study has started to apply the TGAZ API, which is a read-only interface designed to search the contents of the China Historical GIS place name database, to implement the geographic information analysis tool for supporting the spatial analysis of Chinese ancient books. The intelligent search tool contains full-text search and metadata search with post-query classification according to a humanist's search query. Full-text search aims to help humanists find the documents that contain a specific term, whereas metadata search aims to help humanists find the documents with a specific metadata attribute. Also, because a huge amount of humanists' behaviors of operating the CABDHRP are worthy of collecting, this study thus developed the user behavior recording tool based on experience-API (xAPI), which is a new-generation learning process record standard developed by advanced distributed learning (ADL) (Tin Can API, 2015). Recording the operation processes of humanities scholars on the CABDHRP aims to evaluate humanities scholars' research behaviors for improving the usability of system functions and is not an effort to invade the privacy of system users. "Actor", "verb" and "object" are the record formats generated by the learning process recorded through the xAPI. The *actor* is the executor of behaviors and could be an individual or a group; for example, a user. The *verb* is the type of action performed by the actor; for example, clicked or logged in. The *object* is the entity that interacts with the actor; for example, an image. The xAPI provides the flexible design architecture to allow self-defined verb names for the recorded learning process.

- *Advanced text and social network mining module*: The advanced text and social network mining module contains the user behavior analysis tool, social network analysis tool, and textual mining tool. The user behavior analysis tool aims at purposively, accurately and authentically recording humanist's behaviors while operating the CABDHRP for DH research by using the xAPI technology. Such records can include details, such as log-in usage time, the system's operating behavior, text viewing time and interactions with other humanists. The social network analysis tool aims to help humanists examine the relationships between the characters extracted from the writings of Chinese ancient books via visualizing characters' social networks. This tool provides a knowledge graph, which describes and stores characters' relationships as triples, as well as represents the knowledge graph as a multi-relational graph consisting of characters as nodes and several considered relationship attributes as different types of edges. In other words, a knowledge graph represents some concepts (e.g. people, places, events) and their semantic relationships (Haslhofer *et al.*, 2018; Krause *et al.*, 2016). A knowledge graph could be vehicles for formalizing and connecting findings and insights derived from the analysis of possibly large-scale digital corpora in the DH domain (Haslhofer *et al.*, 2018). The textual mining tool aims to help humanists explore implicit information or knowledge from the Chinese ancient books based on data mining technologies, including classification, clustering, association rule and prediction.

3.2 The system's main functions

This section mainly focuses on introducing the main functions provided by the developed CABDHRP. Figure 2 shows the home page of the CABDHRP on Ming Dynasty Chinese ancient books. The CABDHRP can not only aid a humanist to explore new facets in the humanistic research, but also encourage the emergence of a learning interest and awareness of Chinese ancient books in the public. Figure 3 shows the full-text search and post-query classification results according to a humanist's search query. Undoubtedly, providing full-text search is the easiest way to help humanists find the documents containing a particular item in a large data set, either in isolation or together with other browsing options, such as faceted search. The post-query classification aims to classify the documents of the resulting set of a query according to several predefined dimensions, such as year and author. Figures 4 and 5 show the comparison charts of term frequency distribution of two terms simultaneously searched by a humanist on year and collection, respectively. This function can help humanists observe the variation of two terms' frequency distribution with year or collection.

When a humanist enters the reading interface of the CABDHRP for text reading, the system would segment the entire text into more than two terms based on the Jieba Chinese parser, a lexicon-based Chinese word segmentation system and display these segmented terms highlighted in blue. When a humanist moves the mouse cursor to the term highlighted in blue, the system would change the term so that it is highlighted in red to remind that the term has available annotation information from the databases with LD for the humanist (Figure 6). The system would then present the annotation content after the humanist clicks on the term highlighted in red. On the top of the annotation, the annotations of five default databases (including Wikipedia, CBDB, TGAZ, Moedict and EC dictionary) could be mutually referred to. It would not be displayed when there is no LD in the five default databases for a term. When a humanist clicks on a term with available annotation to view the automatic annotation content, the annotations from different LD resources are displayed in the default order of the system. The user could rate whether such annotation is helpful or not based on their self-judgment. The humanist would enhance the annotation score if they rated the annotation as a useful database; on the other hand, the annotation score would be reduced if they rated the annotation as not helpful. The system would calculate the accumulated annotation scores of different LD



關於本資料庫

以國圖特藏明人文集為文本，將其數位化成果全文轉置後，結合政大社資中心開發之各項數位分析工具，建置出一個既符合人文學者研究需求，又能提供大眾古籍內容分析的平台。平台提供全文檢索、檢索結果後分類、文本自動標註、資訊視覺、統計分析、人物社會網絡分析等數位分析工具，並將持續進行文本擴充與數位工具之研發，期待隨著與時俱進的科技發展，發掘古籍於數位時代的新定位，創新圖書館古籍數位應用服務。

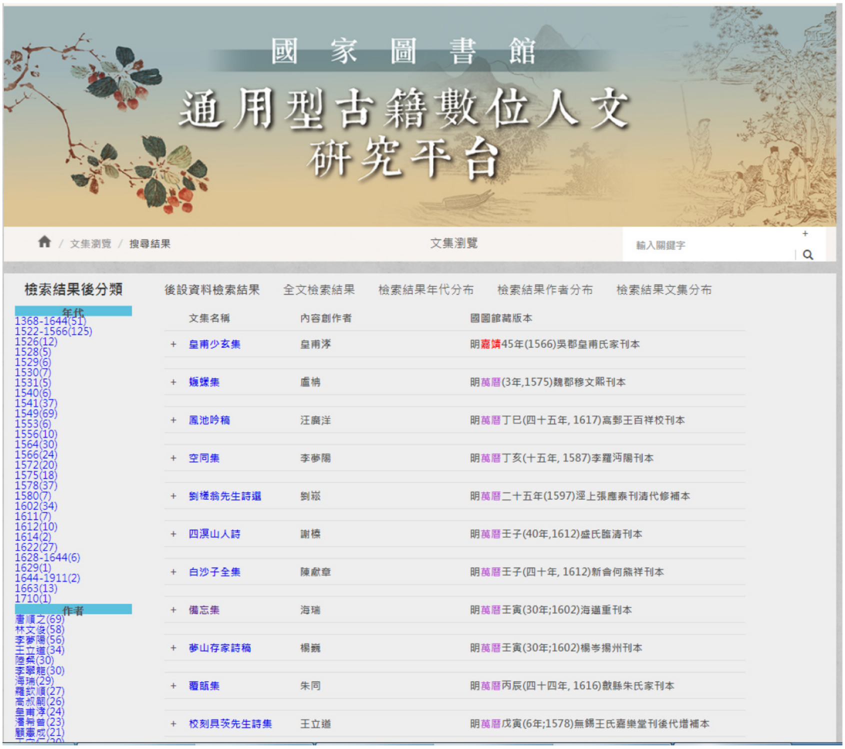
國家圖書館著作權聲明 Copyright © All rights reserved.© 2016 National ChengChi University Social Science Information Center All Rights Reserved.
DSpace Software Copyright © 2002-2004 MIT & Hewlett-Packard / Enhanced by NTU Library IR team Copyright © 2006-2017 -
建議使用Google Chrome瀏覽器以取得最佳使用效果

Figure 2.
The home page of the CABDHRP on Ming Dynasty Chinese ancient books

resources. When the humanist re-clicks on the annotation, the annotations from different LD resources would be ordered according to the accumulated annotation scores; ones with higher annotation scores are displayed in higher priority (Figure 6).

The ATAS would be affected by the correctness of Chinese word segmentation and some unknown terms that cannot be identified by the Chinese word segmentation system. Currently, the accuracy of Chinese word segmentation for the texts of Ming Dynasty ancient books has still not been good enough due to a lot of one-gram words with meaning in Chinese ancient texts and lack of a Chinese word segmentation lexicon that can parse Chinese ancient texts well. Therefore, the ATAS would not show automatic annotations based on LD from different databases for the unknown terms appearing in a text. In this case, the humanist could select a term without annotations in the system. When there are annotations for the term, the humanist could consider adding the term as a new term. After clicking on the "Add new term button", the system would pop up a window for confirmation, and then the new term is added to the parser lexicon of the Chinese word segmentation system Jieba so that the correctness of Chinese word segmentation can be gradually increased because of the human–computer interaction (HCI). The system would precede correct word segmentation and automatic annotation of the term in the next loading for the reader (Figure 7).

Applying social network analysis for DH research aims to examine the relationships between individuals, organizations or groups that interact with each other. These relationships represented by social networks are often visualized to help humanists understand how they are tied together and assist in drawing conclusions and raising further research questions. Thus, the CABDHRP also develops a character social network relationship map tool (CSNRMT) that can



年代
1368-1644(51)
1522-1566(125)
1526(12)
1528(5)
1529(6)
1530(7)
1531(5)
1540(6)
1541(37)
1549(69)
1553(6)
1556(10)
1564(30)
1566(24)
1572(20)
1575(18)
1578(37)
1580(7)
1602(34)
1611(7)
1612(10)
1614(2)
1622(27)
1628-1644(6)
1629(1)
1644-1911(2)
1663(13)
1710(1)

作者
唐順之(69)
林文俊(58)
李夢陽(56)
王立道(34)
陸粲(30)
李攀龍(30)
海瑞(29)
羅欽順(27)
高叔嗣(26)
皇甫涍(24)
潘希曾(23)
顧憲成(21)

	文集名稱	內容創作者	國圖館藏版本
+	皇甫少玄集	皇甫涍	明嘉靖45年(1566)吳郡皇甫氏家刊本
+	蟣蝶集	盧柟	明萬曆(3年,1575)魏郡穆文熙刊本
+	凰池吟稿	汪廣洋	明萬曆丁巳(四十五年, 1617)高郵王百祥校刊本
+	空同集	李夢陽	明萬曆丁亥(十五年, 1587)李羅沔陽刊本
+	劉槎翁先生詩選	劉崧	明萬曆二十五年(1597)涇上張應泰刊清代修補本
+	四溟山人詩	謝榛	明萬曆壬子(40年,1612)盛氏臨清刊本
+	白沙子全集	陳獻章	明萬曆壬子(四十年, 1612)新會何熊祥刊本
+	備忘集	海瑞	明萬曆壬寅(30年;1602)海邁重刊本
+	夢山存家詩稿	楊巍	明萬曆壬寅(30年;1602)楊孝揚州刊本
+	覆瓿集	朱同	明萬曆丙辰(四十四年, 1616)歙縣朱氏家刊本
+	校刻具茨先生詩集	王立道	明萬曆戊寅(6年;1578)無錫王氏嘉樂堂刊後代增補本

Figure 3.
A full-text search and post-query classification results of a humanist's search query

semi-automatically assist humanists to more efficiently and accurately explore the characters' social network relationships from texts for useful research findings through HCI. The CSNRMT developed on the CABDHRP provides a user-friendly interface and functions, allowing humanists interpreting characters' social network relationships with the support of a character social network relationship map, reading interface, text and external search and notebook provided by the system, in the text interpretation process. The overall system interface of the CSNRMT is shown in Figure 8, and the functions are explained below.

3.2.1 The user interface of the CSNRMT. Users can view the character social network relationship in texts by using the user interface of the CSNRMT as shown in Figure 9. The orange and blue nodes in the figure are all characters in the text automatically determined by the system. The orange character nodes are the characters appearing in the text of the current reading text paragraph, and the purple character nodes are the characters appearing in other text paragraphs. For example, one of character nodes appearing in the CSNRMT is shown in Figure 10, and each pair of character node will display two character names and the relationship attribute between the two character names. The system would pre-judge the character's relationship as unknown based on the character relationship analyzer. The edition interface of character social network relationship is shown as Figure 11. By clicking on the line of character relationship in the character social network relationship map, the character social network relationship matrix compiler is enabled for editing the character social network relationship. The edited relationship would be immediately updated on the character social network relationship map.

3.2.2 The user interface for reading a text. The reading interface of a text, as shown in Figure 12, is integrated with the character social network relationship map and external search

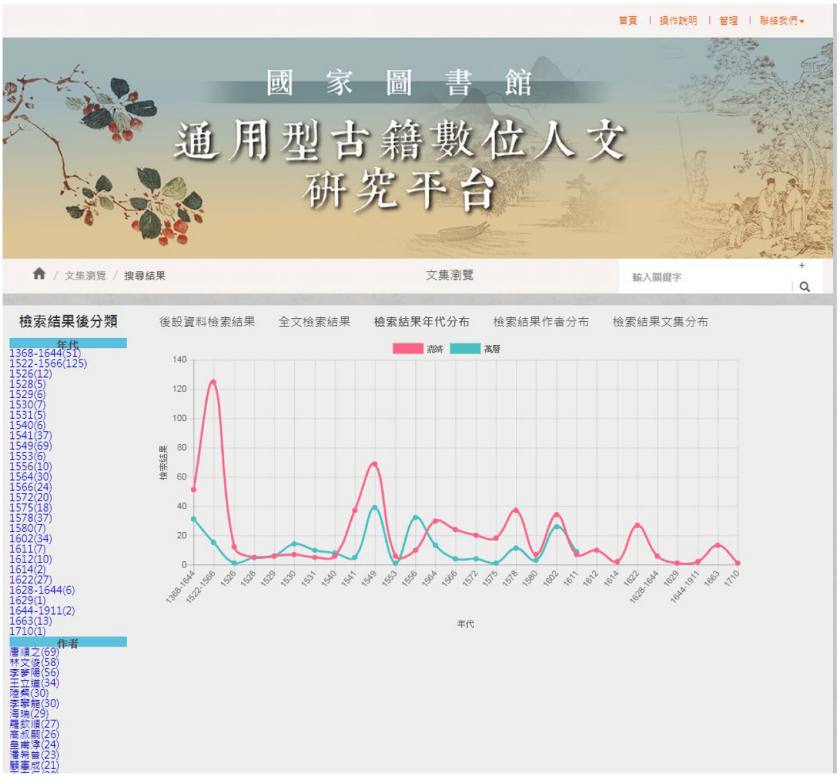

首頁 | 操作說明 | 管理 | 聯絡我們
國 家 圖 書 館
通用型古籍數位人文
研究平台
文集瀏覽 / 搜尋結果
文集瀏覽
輸入關鍵字
檢索結果後分類
後設資料檢索結果
全文檢索結果
檢索結果年代分布
檢索結果作者分布
檢索結果文集分布
年代
1368-1644(51)
1522-1566(125)
1526(12)
1528(5)
1529(6)
1530(7)
1531(5)
1540(6)
1541(37)
1549(69)
1553(6)
1556(10)
1564(30)
1566(24)
1572(20)
1575(18)
1578(37)
1580(7)
1602(34)
1611(7)
1612(10)
1614(2)
1622(27)
1628-1644(6)
1629(1)
1644-1911(2)
1663(13)
1710(1)
作者
唐順之(69)
林文俊(58)
李夢陽(56)
王立道(34)
陸粲(30)
李攀龍(30)
海瑞(29)
羅欽順(27)
高叔嗣(26)
皇甫涍(24)
潘希曾(23)
顧憲成(21)
嘉靖
萬曆
140
120
100
80
60
40
20
0
檢索結果
1368-1644
1522-1566
1526
1528
1529
1530
1531
1540
1541
1549
1553
1556
1564
1566
1572
1575
1578
1580
1602
1611
1612
1614
1622
1628-1644
1629
1644-1911
1663
1710
年代

Figure 4.
A comparison chart of term frequency distribution of two terms searched by a humanist based on year

function. The character social network relationship map would correspond to the changes according to the text paragraphs in the reading interface. When a user clicks on a character node appearing on the paragraph in the character social network relationship map, the location of the character in the text is highlighted on the reading interface for the user analyzing the character social network relationships. Marking a word by using a mouse could link with the external search function for querying extra information about the word. Detailed operation mechanisms for the function are explained in the text search and external search.

3.2.3 The user interface of text search and external search. The text search allows users searching characters or words in a text and will highlight the search results in the text. The external search allows users searching required information through the default external databases for interpreting the character social network relationships in a text. It provides external search through various databases, such as CBDB, Moedict, Wikipedia, Baidu Baike, Google, Zdic, dictionary of Chinese Character Variants and Kangxi Dictionary, as shown in Figure 13. A user simply needs to mark a word in a text, and the word will be automatically delivered to the external search box for searching.

3.2.4 The user interface of notebook. The user interface of notebook is shown in Figure 14 and allows users recording found information during interpreting a text, reducing the time for switching between windows to complete the interpretation of the character social network relationships under a single interface.

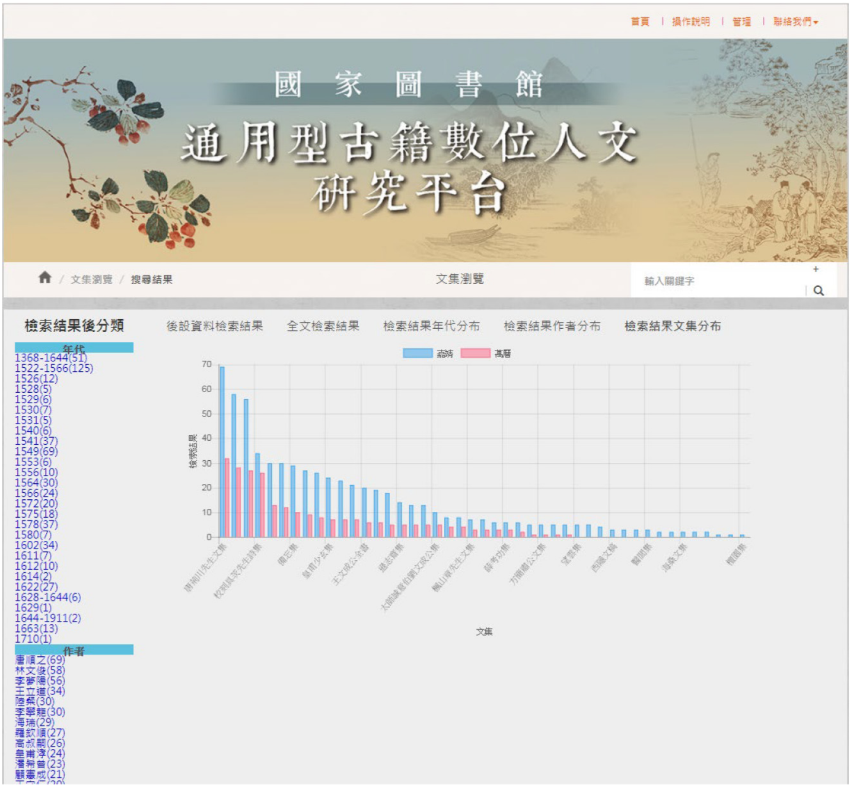
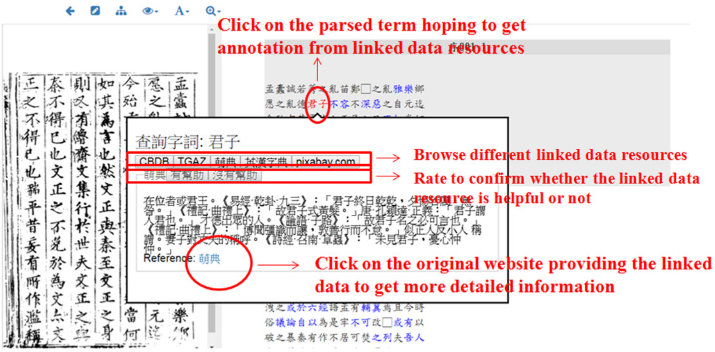

首頁 | 操作說明 | 管理 | 聯絡我們
國家圖書館
通用型古籍數位人文研究平台
文集瀏覽 / 搜尋結果
文集瀏覽
輸入關鍵字
檢索結果後分類
後設資料檢索結果
全文檢索結果
檢索結果年代分布
檢索結果作者分布
檢索結果文集分布
年代
1368-1644(51)
1522-1566(125)
1526(12)
1528(5)
1529(6)
1530(7)
1531(5)
1540(6)
1541(37)
1549(69)
1553(6)
1556(10)
1564(30)
1566(24)
1572(20)
1575(18)
1578(37)
1580(7)
1602(34)
1611(7)
1612(10)
1614(2)
1622(27)
1628-1644(6)
1629(1)
1644-1911(2)
1663(13)
1710(1)
作者
唐順之(69)
林文俊(58)
李夢陽(56)
王立道(34)
陸粲(30)
李攀龍(30)
海瑞(29)
羅欽順(27)
高叔嗣(26)
皇甫涍(24)
潘希曾(23)
顧憲成(21)
嘉靖
萬曆
70
60
50
40
30
20
10
0
檢索結果
文集

Figure 5.
A comparison chart of term frequency distribution of two terms searched by a humanist based on collection


查詢字詞: 君子
CBDB | TGAZ | 萌典 | 找漢字典 | pixabay.com
萌典 | 有幫助 | 沒有幫助
Reference: 萌典

Click on the parsed term hoping to get annotation from linked data resources

Browse different linked data resources

Rate to confirm whether the linked data resource is helpful or not

Click on the original website providing the linked data to get more detailed information

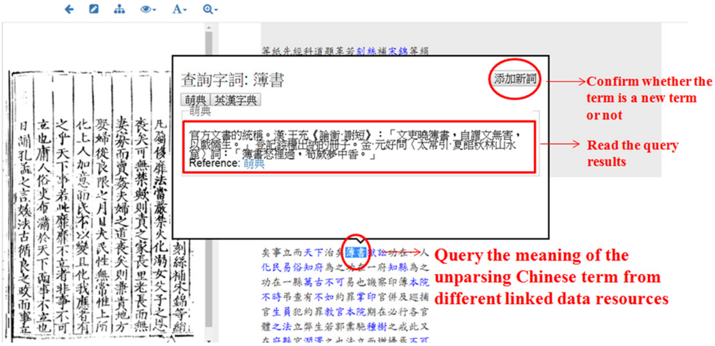
Figure 6.
The user interface for reading Chinese ancient text on the CABDHRP

3.2.5 The user interface for exporting character social network relationship. When recording the character social network relationships, the relationships between characters are recorded as a triple-dimensional matrix that contains two character names and their corresponding relationship attribute and can be optionally exported as a CSV (comma-separated value) file for Microsoft Excel import or an RDF (resource description framework) file as shown in Figure 15 so that a humanist can further conduct social network analysis via social network analysis software, such as UCINET or Pajek. Therefore, the character social network relationships with RDF format generated by the CSNRMT can easily be published as LOD to provide machine-readable information to support web service in the future.

4. Research methodology

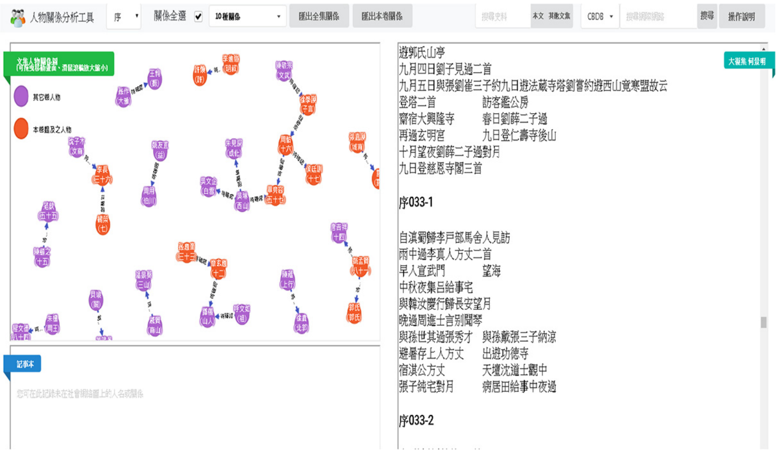
4.1 Research design

To examine the effects of the ATAS and CSNRMT developed on the CABDHRP for assisting DH research, two experiments were designed in this study to examine the effectiveness of the research participants on interpreting Chinese ancient text and characters' social network relationships. In the first experiment, research participants were invited to interpret a paragraph of ancient text of the Ming Dynasty's writings with the support of the ATAS within 20 min, and then write down their reading abstracts. In addition, an historical expert who is a university professor of the Department of History was invited to select an ancient text with a lot of characters' relationships from the Ming Dynasty collection archived in the CABDHRP as the experimental text. In the second experiment, research participants were invited to read an ancient text with the assistance


查詢字詞: 簿書
添加新詞
萌典
英漢字典
萌典
Confirm whether the term is a new term or not
Read the query results
Query the meaning of the unparsing Chinese term from different linked data resources

官方文書的統稱。漢·王充《論衡·謝短》：「文吏曉簿書，自謂文無害，以戲儒生。」登記錢糧出納的冊子。金·元好問〈太常引·夏館秋林山水圖〉詞：「簿書愁裡過，笱歛夢中香。」
Reference: 萌典

Figure 7.
The user interface for adding a new term into the lexicon for enhancing Chinese word segmentation performance on the CABDHRP


人物關係分析工具
序
關係全選
10種關係
匯出全集關係
匯出本卷關係
搜尋史料
本文
其他文集
CBDB
搜尋網際網路
搜尋
操作說明
文集人物關係圖
其它類人物
本卷題及之人物
大復集 何景明
記事本
您可在此記錄未在社會關係圖上的人名關係

遊郭氏山亭
九月四日劉子見過二首
九月五日與張劉崔三子約九日遊法藏寺塔劉嘗約遊西山竟寒盟故云
登塔二首　　訪客鑑公房
齋宿大興隆寺　　春日劉薛二子過
再過玄明宮　　九日登仁壽寺後山
十月望夜劉薛二子過對月
九日登慈恩寺閣三首

序033-1

自滇蜀歸李戶部馬舍人見訪
雨中過李真人方丈二首
早入宣武門　　望海
中秋夜集吕給事宅
與韓汝慶行歸長安望月
晚過周進士言別聞琴
與孫世其過張秀才　　與孫戴張三子納涼
避暑存上人方丈　　出遊功德寺
宿淇公方丈　　天壇沈道士觀中
張子純宅對月　　病居田給事中夜過

序033-2

Figure 8.
The overall system interface of the CSNRMT

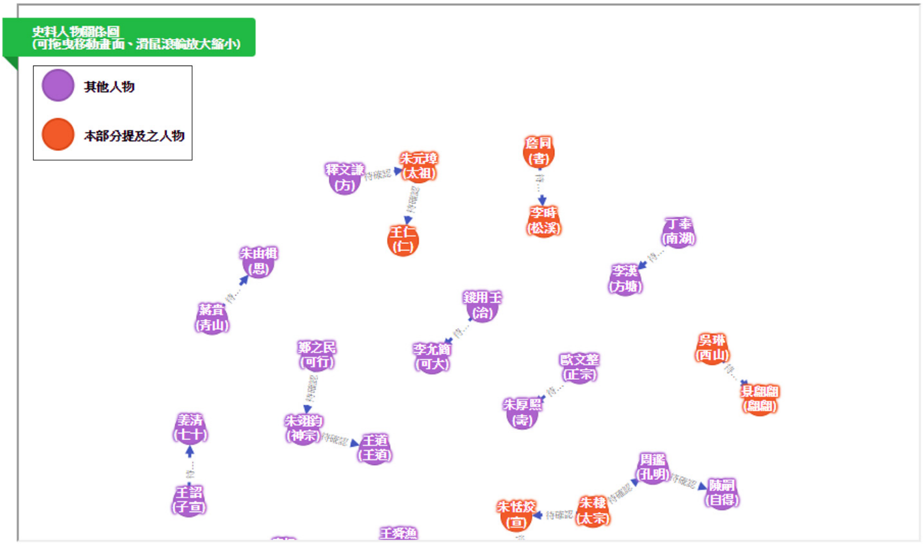

史料人物關係圖
(可拖曳移動畫面、滑鼠滾輪放大縮小)
其他人物
本部分提及之人物
釋文謙(方)
朱元璋(太祖)
王仁(仁)
詹同(書)
李時(松溪)
丁奉(南湖)
李漢(方塘)
朱由楫(思)
蔣貴(青山)
錢用壬(治)
李允簡(可大)
鄭之民(可行)
歐文整(正宗)
朱厚熙(壽)
吳琳(西山)
吳翩翩(翩翩)
姜清(七十)
王詔(子宣)
朱翊鈞(神宗)
王道(王道)
周鑑(孔明)
陳嗣(自得)
朱怡炆(宣)
朱棣(太宗)
王舜漁
待確認

Figure 9.
The user interface of the character social network relationship map

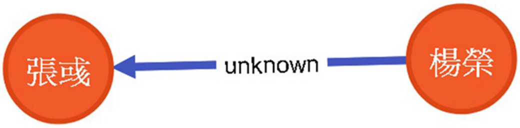
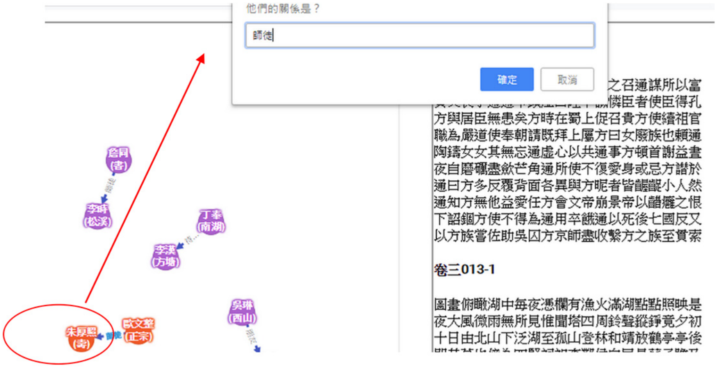

張彧
unknown
楊榮

Figure 10.
The social network relationship between two characters


詹同(書)
李時(松溪)
丁奉(南湖)
李漢(方塘)
吳琳(西山)
朱厚熙(壽)
歐文整(正宗)
確定
取消

他們的關係是？

師徒

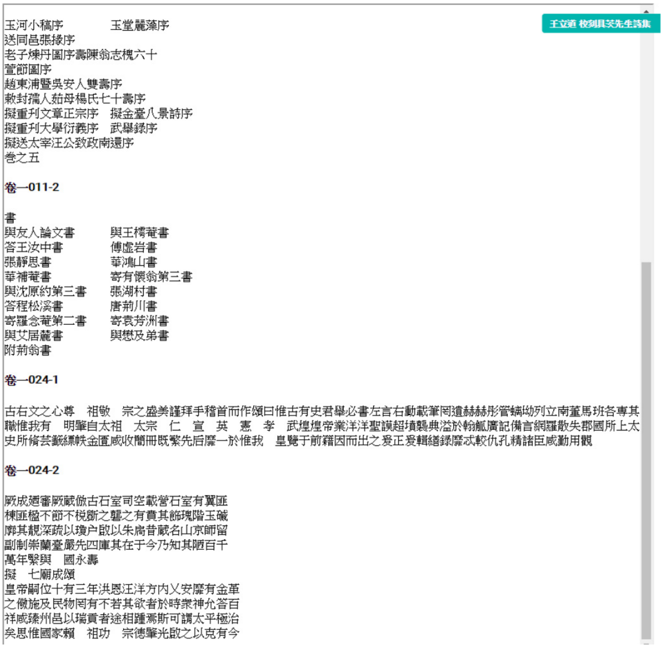
之召通謀所以富
憐臣者使臣得孔
方與居臣無患矣方時在蜀上促召貴方使續祖官
職為嚴道使奉朝請既拜上屬方曰女廢族也賴通
陶鑄女女其無忘通虛心以共通事方頓首謝益晝
夜自磨礪盡斂芒角通所使不復愛身或忌方譖於
通曰方多反覆背面各異與方昵者皆齷齪小人然
通知方無他益愛任方會文帝崩景帝以齟齬之恨
下詔錮方使不得為通用卒餓通以死後七國反又
以方族嘗佐助吳囚方京師盡收繫方之族至貫索

卷三013-1

圖畫俯瞰湖中每夜憑欄有漁火滿湖點點照映是
夜大風微雨無所見惟聞塔四周鈴聲鏦錚竟夕初
十日由北山下泛湖至孤山登林和靖放鶴亭亭後

Figure 11.
The edition interface of the character social network relationship

of CSNRMT for interpreting characters' social network relationships within 40 min, and then write down the social relationships of interpreting characters. Semi-structured in-depth interviews were then conducted after the research subjects completed the reading abstracts and the interpretation of characters' social network relationships. The interview questions included: "Are there any difficulties or questions when using the



玉河小稿序　　　玉堂麗藻序
送同邑張掾序
老子煉丹圖序壽陳翁志槐六十
萱節圖序
趙東浦暨吳安人雙壽序
敕封孺人茹母楊氏七十壽序
擬重刋文章正宗序　擬金臺八景詩序
擬重刋大學衍義序　武舉錄序
擬送太宰汪公致政南還序
卷之五

卷一011-2

書
與友人論文書　　　與王檉菴書
答王汝中書　　　傅虛岩書
張靜思書　　　華鴻山書
華補菴書　　　寄有懷翁第三書
與沈原約第三書　張湖村書
答程松溪書　　　唐荊川書
寄羅念菴第二書　寄袁芳洲書
與艾居麓書　　　與懋及弟書
附荊翁書

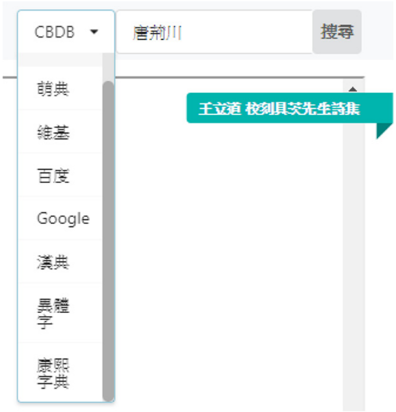
卷一024-1

古右文之心尊　祖敬　宗之盛美謹拜手稽首而作頌曰惟古有史君舉必書左言右動載筆罔遺赫赫彤管螭坳列立南董馬班各專其
職惟我有　明肇自太祖　太宗　仁　宣　英　憲　孝　武煌煌帝業洋洋聖謨超墳軼典溢於翰觚廣記備言網羅散失郡國所上太
史所脩芸籤縹帙金匱咸收簡冊既繁先后靡一於惟我　皇覽于前籍因而出之爰正爰輯繕錄靡忒較仇孔精諸臣咸勤用觀

卷一024-2

厥成廼審厥藏倣古石室司空載營石室有翼匪
楝匪楹不節不梲斲之礱之有賁其飾瑰階玉礣
廓其靚深疏以璦户啟以朱扃昔藏名山京師留
副制崇蘭臺嚴先四庫其在于今乃知其陋百千
萬年繄與　國永壽
擬　七廟成頌
皇帝嗣位十有三年洪恩汪洋方内乂安靡有金革
之儆施及民物罔有不若其欲者於時衆神允答百
祥咸臻州邑以瑞貢者途相踵焉斯可謂太平極治
矣思惟國家賴　祖功　宗德肇光啟之以克有今

Figure 12.
The user interface for reading a text



唐荊川

Figure 13.
The user interface for external search

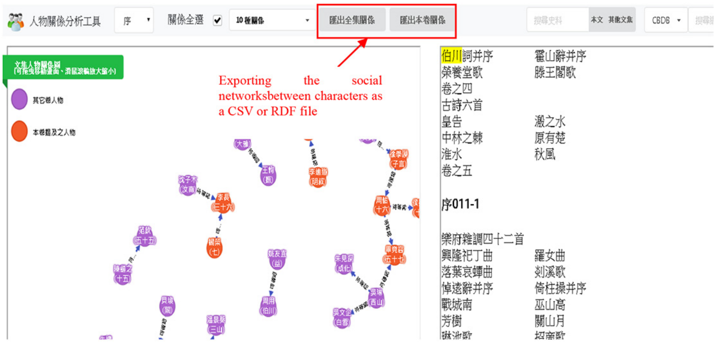
functions of the ATAS and CSNRMT in the system?", "Is the overall system design clear and understandable?", "Is the human-computer interaction design in the system perceived as easy to use?" and "Are there any opinions or suggestions for the future development of the system?".

4.2 Research subjects

The research participants are the students who are able to interpret the ancient texts of Ming Dynasty and explore characters' social network relationships. In consideration of cost, time and location, a total of 21 students from the departments of Chinese

記事本
您可在此記錄未在社會網絡圖上的人名或關係

Figure 14.
The user interface for a notebook


人物關係分析工具
序
關係全選 ☑
10 種關係
匯出全集關係
匯出本卷關係
搜尋史料
本文
其他文集
CBDB
文集人物關係圖
(可拖曳移動畫面、滑鼠滾輪放大縮小)
● 其它卷人物
● 本卷提及之人物
伯川詞并序
榮養堂歌
卷之四
古詩六首
皇告
中林之棘
淮水
卷之五
序011-1
樂府雜調四十二首
興隆祀丁曲
落葉哀蟬曲
悼遠辭并序
戰城南
芳樹
霍山辭并序
滕王閣歌
濈之水
原有楚
秋風
羅女曲
剡溪歌
倚柱操并序
巫山高
關山月

Exporting the social networksbetween characters as a CSV or RDF file

Figure 15.
The user interface for exporting the character social network relationship

Literature and History in a national university in Taipei City, Taiwan, who were willing to participate in the experiment, were sampled to conduct two experiments, respectively, with the support of the ATAS and CSNRM for a target task related to DH research. One task was to interpret a paragraph of an ancient text of the Ming Dynasty's writings with the support of the ATAS, and then write down the reading abstracts. Another one is to read the ancient text with the assistance of CSNRMT for interpreting characters' social network relationships. The 21 students included 3 undergraduates, 14 graduates and 4 doctoral students.

5. The use assessment of the ATAS and CSNRM

5.1 The use assessment of the ATAS

To understand the research participants' using perception when interpreting the texts with the support of the ATAS, semi-structured interviews with the 21 research participants were conducted. The results are summarized as follows. First, this study found that different opinions appear on the function of the automatic Chinese word segmentation. Most of the research participants agreed with the accuracy of automatic annotation based on Chinese word segmentation in interpreting the contents of the ancient texts and accelerating the reading speed. Other research participants considered that the automatic Chinese word segmentation did not provide much assistance, as they could segment the terms by themselves, or incorrect Chinese word segmentation might affect the text interpretation. Second, the research participants expressed that the function to link to source websites allows them to view more useful information. Particularly, the complete annotation information would be more easily browsed on source websites than on the ATAS. Importantly, most of the research participants strongly expressed the necessity of text segmentation when reading ancient texts without punctuations such as the texts in the Ming Dynasty's writings. They also regarded that the importance of text segmentation is higher than word segmentation in smoothing reading. A lot of the research participants expressed that there are many variations in Chinese ancient texts and suggested that the variations dictionary of the Ministry of Education of Taiwan could be included for the search, or the pre-processing of variations could be proceeded before uploading texts to websites. Finally, most of the research participants considered that the ATAS should increase LOD sources from other databases. They suggested that increasing online databases or dictionaries which are often used for interpreting Chinese ancient texts, such as Chinese dictionary (http://dict.revised.moe.edu.tw/cbdic/), the Scripta Sinica database (http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm) and Zdic (www.zdic.net/), should be considered.

5.2 The use assessment of the CSNRM

To understand the research participants' using perception when interpreting the historical character social network relationships with the support of the CSNRMT, the semi-structured interviews with the 21 research participants were conducted. The results are summarized as follows. First, most interviewees considered that the CSNRMT developed in this study presents a clearly understandable system interface, clear marks among function blocks and good mutual reference among system functions. Second, the character social network relationship map could present possible characters and their relationships in the text to help users browse the historical characters' relationships before interpreting the text so as to understand the possible relationships mentioned in the text and allow users to not miss characters and their relationships in the text interpretation process. However, many interviewees expressed that some characters in other dynasties would be identified as the characters in the Ming Dynasty in the character social network relationship map. For instance, if the first founder of the Song Dynasty was identified as the first founder of the Ming Dynasty, then the result would seriously affect the effectiveness to interpret characters' social network relationships in the text. Also, some interviewees indicated that the characters named Hsiung sir and Chang sir simultaneously appearing in a text could not be identified accurately in the developed character social network relationship map because they are the short forms of two character names; however, such characters might play critical roles in the character social network relationship map. Moreover, a lot of interviewees indicated that the system could search characters through the internal search and various external search functions, including Wikipedia and CBDB, to largely assist in interpreting characters' social network relationships in a text. The internal text search function provided by the system allows users clicking on a character in the character social network relationship map to rapidly mark the location of the character appearing in a text. It also allows users to discover useful information related to the character through the external search functions for interpreting characters' social network relationships. Finally, the note function and reading interface developed in this system are presented on the same interface, allowing users reading text and making notes at the same time to reduce the switch times between the reading text interface and note interface. They, therefore, could pay more attention to interpreting characters' social network relationships in a text.

6. Discussion

The Jieba Chinese parser was applied to the ATAS for automatic Chinese word segmentation. However, the Jieba Chinese parser is not particularly developed for processing Chinese ancient texts, so the word segmentation of Chinese ancient texts could not achieve an accuracy as high as modern Chinese texts. Although the ATAS offers the function of adding new terms to improve the accuracy of Chinese word segmentation through semi-automatic adjustment of HCI, the growth of the Chinese word segmentation lexicon size is still too slow to significantly increase the accuracy of word segmentation of Chinese ancient text. Namely, developing a Chinese word segmentation system that can identify Chinese ancient texts well is an urgent issue for the ATAS. Currently, there is no such Chinese word segmentation system that has been well developed for Chinese ancient texts. To overcome this problem, developing a Chinese word segmentation lexicon to support a Chinese word segmentation system based on analyzing a large amount of the Ming Dynasty's ancient texts collected in the developed CABDHRP is practicable. This has been considered as our future work. More importantly, most of the research participants strongly expressed the necessity of text segmentation when reading Chinese ancient texts without punctuations like the Ming Dynasty's writings. They regarded that the text segmentation is more important than word segmentation while reading a Chinese ancient text for DH research. Actually, text segmentation will also influence the accuracy of Chinese word segmentation of an ancient text (Pak and Teh, 2016). Therefore, developing a text segmentation scheme with high accuracy for Chinese ancient texts is a more urgent issue than word segmentation, particularly in using the ATAS to support reading Chinese ancient text without punctuations.

Additionally, the character names' identification in the CSNRMT mainly focuses on the Ming Dynasty's characters collected in CBDB. That is, the character names are simply identified based on the lexicon of the Ming Dynasty's character names from CBDB. Some research participants indicated that some character names might be wrongly identified, thus causing serious interference in the reading process. Additionally, some other research participants indicated that some characters before the Ming Dynasty are not identified by the system, while such characters might be critical in presenting primary characters' social network relationships. Moreover, the problem of cross-dynasty characters with the same name might result in mistake. For instance, an interviewee mentioned that the first emperor of the Song Dynasty is identified as the first emperor of the Ming Dynasty due to the same posthumous name. Obviously, the characters' social network relationships identified by the CSNRMT should be more accurate and broad to conform to the text characteristics, types and humanists' needs. Namely, the character name recognition technology of the Ming Dynasty writings should be further improved based on discovering the rules that can correctly identify character names by using advanced machine learning technologies, such as the hidden Markov model (Bikel *et al.*, 1997) or conditional random field (Gu *et al.*, 2015). Particularly, human–machine cooperation based on active learning (Settles, 2012) may help improve the accuracy of the character name recognition technology of the Ming Dynasty writings.

Finally, a variety of DH projects around the world are working to build digital tools for facilitating the research by humanists, scholars and educators, but these tools are often "one-offs" from a technology perspective. That is, they focus on a single activity or set of activities and function independently from one another. To provide more complete and powerful functions in supporting DH research, integrating the functions of the CABDHRP with the digital tools developed by other DH platforms should be considered. However, integrating existing digital tools into the CABDHRP presents high challenges because they are built using a variety of programming languages and system level dependencies, and in most cases, they were never designed with interoperability in mind (Bialock *et al.*, 2018). This study strongly pleads with designers of such systems that the issues of function interoperability and data exchange between different DH platforms should be considered in the design process. Generally, using open API to provide system functions is a practicable way to achieve the goal of function interoperability. Furthermore, there have been a lot of methods that can be used to achieve the goal of data exchange between different DH platforms, such as via files, an SQL database, LOD or web services (Baierer *et al.*, 2016).

7. Conclusion and future work

In recent years, with the advancement of information and communication technologies, the retrieval and use of Chinese ancient books is presenting an unprecedented challenge and opportunity. This study successfully developed the CABDHRP, which can archive any subject's Chinese ancient digital writings and provide several emerging digital tools, including the ATAS and CSNRMT, to support DH research. Many interviewees indicated that the user interface and functions of the ATAS and CSNRMT developed in this study are practicable in supporting DH research and easy to use, the functions are clearly marked and the integration among functions is good. The CABDHRP not only provides innovative and forward-looking applications in the DH field, but also encourages the emergence of a learning interest and awareness of Chinese ancient books in the public. More importantly, this study tried to develop a systematic and standardized DH research platform with novel and practicable digital tools in a virtual environment for different humanities disciplines with similar but heterogeneous data sets.

Additional studies are warranted. First, text segmentation technology that divides digital text into meaningful sentences is regarded more important than word segmentation that divides digital text into meaningful words when reading Chinese ancient text without punctuations. Therefore, the future work of enhancing the CABDHRP to support DH research more effectively should focus on developing an automatic text segmentation algorithm to insert appropriate punctuations into Chinese ancient texts for promoting the reading comprehension performance of humanists. In this work, the deep learning derived from neural networks, such as the bidirectional LSTM (long short-term memory), has a very high potential to be applied to solving this problem (Wang *et al.*, 2017). Second, annotations typically facilitate text reading. Compared to the automatic text annotation system developed in the CABDHRP, manually collaborative annotations of digital texts can get more correct annotated contents from humanists and allow humanists to share their knowledge with each other. Most importantly, the annotated contents can help readers obtain a deeper and broader understanding while reading a text compared to digital content without annotations. Thus, future work for this study will integrate the manually collaborative reading annotation system with a reading annotation and interactive discussion scaffold developed in our previous work into the CABDHRP to improve humanists' reading performance for DH research (Chen and Chen, 2014). Third, the character name recognition technology for the Ming Dynasty writings should be further improved through discovering the rules that can correctly identify character names by using machine learning technologies. Fourth, opinion leaders in the two-step flow of the communication model were the brokers between mass media and the public, as well as were the key persons in information transfer (Li *et al.*, 2013). The future work for this study should develop an opinion leader and maximum influencer identification algorithm for the CSNRMT based on centrality, structure hole or PageRank to help humanists perceive useful characters' relationship clues for discovering new insights. Finally, developing a dashboard system for real-time analysis of humanists' behavior processes on the CABDHRP should be considered in the future. If the behavior process records of humanists interpreting text by using the CABDHRP can be analyzed in real time, then the CABDHRP can provide real-time feedback with more effective information for supporting humanists' text interpretation.

References

Baierer, K., Dröge, E., Eckert, K., Goldfarb, D., Iwanowa, J., Morbidoni, C. and Ritze, D. (2016), "DM2E: a linked data source of digitised manuscripts for the digital humanities", *Semantic Web – Interoperability, Usability, Applicability, an IOS Press Journal*, Vol. 2014 No. 1, pp. 1-13.

Bialock, T., Childress, D., Ikuura, H. and Nagasaki, K. (2018), "Towards unified descriptive practices for Japanese classical texts: TEI, IIIF, and the UCLA toganoo collection of esoteric Buddhism", in *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*, pp. 78-79.

Bikel, D., Schwarta, R. and Weischedel, R. (1997), "An algorithm that learns what's in a name", *Machine Learning*, Vol. 34 Nos 1/3, pp. 211-231.

Brooke, J., Hammond, A. and Hirst, G. (2015), ""Gutentag: an NLP-driven tool for digital humanities research in the project Gutenberg corpus", in *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, *Association for Computational Linguistics*, Denver, CO, pp. 42-47.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. and Schnapp, J. (2012), *Digital Humanities*, The MIT Press, Cambridge, MA.

Cassella, M. (2017), "New trends in academic library partnerships: academic libraries and digital humanities", in *Proceedings of the IATUL Conferences*, pp. 1-9.

Chan, T.C., Pai, P.L., Shaw, S.L. and Fan, I.C. (2015), "Spatiotemporal evolution of market towns in the Jiangnan area during the Ming-Qing dynasties of China", *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, Vol. 48 No. 2, pp. 90-102.

Chen, C.M. and Chen, F.Y. (2014), "Enhancing digital reading performance with a collaborative reading annotation system", *Computers and Education*, Vol. 77, pp. 67-81.

Chen, C.M. and Tsay, M.Y. (2017), "Applications of collaborative annotation system in digital curation, crowdsourcing, and digital humanities", *The Electronic Library*, Vol. 35 No. 6, pp. 1122-1140.

Chen, S.P., Hsiang, J., Tu, H.C. and Wu, M.C. (2007), "On building a full-text digital library of historical documents", *Lecture Notes in Computer Science: Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers, Hanoi*, Vol. 4822, pp. 49-60.

Chen, C.M., Chen, Y.T., Hong, C.M., Liao, C.W. and Huan, C.M. (2012), "Developing a Taiwan library history digital library with reader knowledge archiving and sharing mechanisms based on the DSpace platform", *The Electronic Library*, Vol. 30 No. 3, pp. 426-442.

Chen, S.P., Huang, Y.M., Hsiang, J., Tu, H.C., Ho, H.I. and Chen, P.Y. (2013), "Discovering land transaction relations from land deeds of Taiwan", *Literary and Linguistic Computing*, Vol. 28 No. 2, pp. 257-270.

Drucker, J., Kim, D., Salehian, I. and Bushong, A. (2014), "Introduction to digital humanities: concepts, methods, and tutorial for students and instructors", *UCLA, California*, available at: http://dh101.humanities.ucla.edu/?page_id=13 (accessed 28 January 2018).

Gu, C., Tian, X.P. and Yu, J.D. (2015), "Automatic recognition of Chinese personal name using conditional random fields and knowledge base", *Mathematical Problems in Engineering*, Vol. 2015, Article ID 480879.

Harris, M., Levene, M. and Zhang, D. (2014), "The anatomy of a search and mining system for digital humanities", in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries, IEEE Press, Piscataway, NJ*, pp. 165-168.

Hashimoto, Y. (2014), "SMART-GS web: a HTML5-powered, collaborative manuscript transcription platform", *Japanese Association for Digital Humanities Annual Conference*, *Ibaraki, Japan*.

Haslhofer, B. Antoine, I. and Rainer, S. (2018), "Knowledge graphs in the libraries and digital humanities domain", available at: https://arxiv.org/pdf/1803.03198.pdf

Haslhofer, B., Robitza, W. and Lagoze, C. (2013), "Semantic tagging on historical maps", *Proceedings of the 5th Annual ACM Web Science Conference*, *Paris, France*, pp. 148-157.

Hsiang, J., Chen, S.P., Ho, H.I. and Tu, H.C. (2012), "Discovering relationships from imperial court documents of Qing dynasty", *International Journal of Humanities and Arts Computing*, Vol. 6 Nos 1/2, pp. 22-41.

Jackson, C. (2017), "Using social network analysis to reveal unseen relationships in medieval Scotland", *Digital Scholarship in the Humanities*, Vol. 32 No. 2, pp. 336-343.

Kallaher, A. and Gamble, A. (2017), "GIS and the humanities: presenting a path to digital scholarship with the story map app", *College and Undergraduate Libraries*, Vol. 24 Nos 2/4, pp. 559-573.

Kamposiori, C. (2017), "The role of research libraries in the creation, archiving, curation, and preservation of tools for the digital humanities, RLUK report", available at: www.rluk.ac.uk/wp-content/uploads/2017/07/Digital-Humanities-report-Jul-17.pdf

Kaplan, F. (2015), "A map for big data research in digital humanities", *Frontiers in Digital Humanities*, Vol. 2 No. 1, pp. 1-7.

Kestemont, M. and Gussem, J.D. (2017), "Integrated sequence tagging for medieval Latin using deep representation learning", *Journal of Data Mining and Digital Humanities, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*, available at: https://jdmdh.episciences.org/3835

Krause, S., Hennig, L., Moro, A., Weissenborn, D., Xu, F., Uszkoreit, H. and Roberto Navigli, R. (2016), "Sar-graphs: a language resource connecting linguistic knowledge with semantic relations from knowledge graphs", *Journal of Web Semantics*, Vol. 37-38, pp. 112-131.

Li, Y., Ma, S., Zhang, Y., Huang, R. and Kinshuk, (2013), "An improved mix framework for opinion leader identification in online learning communities", *Knowledge-Based Systems*, Vol. 43, pp. 43-51.

Luczak-Roesch, M., Grener, A. and Fenton, E. (2018), "Twenty thousand leagues above the book: an interactive visual analytics approach to literature", in *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, *ACM*, *New York, NY*, pp. 94-97.

Martin-Rodilla, P. and Gonzalez-Perez, C. (2016), "Understanding user behavior in textual analysis: a thinking aloud approach for digital humanities research contexts", in *the 4th International Conference on Technological Ecosystem for Enhancing Multiculturality*, pp. 269-276.

Neuroth, H., Lohmeier, F. and Smith, K.M. (2011), "TextGrid: virtual research environment for the humanities", *International Journal of Digital Curation*, Vol. 6 No. 2, pp. 222-231.

Pak, I. and Teh, P.L. (2016), "Text segmentation for analysing different languages", in *The First EAI International Conference on Computer Science and Engineering, 11-12 November 2016*, *Penang, Malaysia*.

Pastorelli, D. (2017), "A classification of manuscripts based on a new quantitative method: the old Latin witnesses of John's gospel as text case", *Journal of Data Mining and Digital Humanities, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*, available at: https://jdmdh.episciences.org/3783

Picca, D. and Egloff, M. (2017), ""DHTK: the digital humanities toolkit", *CEUR Workshop Proceedings*, pp. 81-86.

Rushmeier, H., Pintus, R., Yang, Y., Wong, C. and Li, D. (2015), "Examples of challenges and opportunities in visual analysis in the digital humanities", available at: http://graphics.cs.yale.edu/site/sites/files/dh-ei-SUBMIT.pdf

Sato, T., Goto, M., Kimura, F. and Maeda, A. (2016), "Developing a collaborative annotation system for historical documents by multiple humanities researchers", *International Journal of Computer Theory and Engineering*, Vol. 8 No. 1, pp. 88-93.

Settles, B. (2012), “Active learning”, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Vol. 6 No. 1, pp. 1-114.

Steiner, C.M., Agosti, M., Sweetnam, M.S., Hillemann, E.C., Orio, N., Ponchia, C., Hampson, C., *et al.* (2014), “Evaluating a digital humanities research environment: the CULTURA approach”, *International Journal on Digital Libraries*, Vol. 15 No. 1, pp. 53-70.

Tang, M.C., Cheng, Y.J. and Chen, K.H. (2017), “A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses”, *Scientometrics*, Vol. 113 No. 2, pp. 985-1008.

Therón, R. and Wandl-Vogt, E. (2016), “New trends in digital humanities”, in García-Peñalvo, F.J. (Ed.), *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM ‘16), ACM, New York, NY*, pp. 945-947.

Tin Can API (2015), “What is the tin can API?”, available at: https://tincanapi.com/overview/

Uboldi, G., Caviglia, G., Coleman, N., Heymann, S., Mantegari, G. and Ciuccarelli, P. (2013), “Knot: an interface for the study of social networks in the humanities”, in *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI (CHItaly ‘13), ACM, New York, NY*, p. 9.

Upadhyay, S. and Upadhyay, N. (2017), “Future directions and a roadmap in digital computational humanities for a data driven organization”, *Procedia Computer Science*, Vol. 122, pp. 1055-1060.

Wang, X. and Inaba, M. (2009), “Analyzing structures and evolution of digital humanities based on correspondence analysis and co-word analysis”, アート・リサーチ, Vol. 9, pp. 123-134.

Wang, X., Wang, M. and Zhang, Q. (2017), “Realization of Chinese word segmentation based on deep learning method”, in *Proceedings of AIP Conference*, 1864, 020150-1–020150-6.

Widlöcher, A., Bechet, N., Lecarpentier, J.M., Mathet, Y. and Roger, J. (2015), “Combining advanced information retrieval and text-mining for digital humanities”, in *The ACM Symposium on Document Engineering*, Lausanne, Switzerland, pp. 157-166.

Yohan, P.M., Sasidhar, B., Basha, S., A.H. and Govardhan, A. (2014), “Automatic named entity identification and classification using heuristic based approach for Telugu”, *International Journal of Computer Science Issues*, Vol. 11 No. 1, pp. 173-180.

About the authors

Chih-Ming Chen is currently a professor in the Graduate Institute of Library, Information and Archival Studies, and he is also the University Librarian at National Chengchi University, Taipei, Taiwan. He received BSc and MSc degree from the Department of Industrial Education at National Taiwan Normal University in 1992 and 1997 and received PhD degree from the Department of Electronic Engineering at National Taiwan University of Science and Technology in 2002. His research interests include digital library, digital humanities, e-learning, data mining, machine learning and intelligent agents on the Web. Chih-Ming Chen is the corresponding author and can be contacted at: chencm@nccu.edu.tw

Chung Chang received BSc degree from the Department of Information and Library Science at Fu Jen Catholic University in 2016 and received master’s degree from the Graduate Institute of Library, Information and Archival Studies in 2018. He is now an IT engineer of the Research Center for Chinese Cultural Subjectivity at National Chengchi University, Taipei, Taiwan. Her research interests include DH and digital library.

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**

Reproduced with permission of copyright owner. Further reproduction prohibited without permission.