Journal of Chinese History (2020), 1–17
doi:10.1017/jch.2020.28

CAMBRIDGE
UNIVERSITY PRESS

UTILITIES

DocuSky, A Personal Digital Humanities Platform for Scholars*

Hsieh-Chang Tu[1], Jieh Hsiang[2]*, I-Mei Hung[1] and Chijiu Hu[1] Q1

[1]Research Center for Digital Humanities, National Taiwan University and [2]Department of Computer Science and Research Center for Digital Humanities, National Taiwan University
*Corresponding author. Email: jhsiang@ntu.edu.tw

Abstract

Q2

Introduction

DocuSky is a personal digital humanities platform for humanities scholars, which aims to become a platform on which a scholar can satisfy all her digital needs with no direct IT assistance. To this end, DocuSky provides tools for a scholar to download material from the Web and prepare (annotating, building metadata) her material, a one-click function to build a full-text searchable database, and tools for analysis and visualization. DocuSky advocates the separation of digital content and tools. Being an open platform, it encourages IT developers to build tools to suit scholars' needs, and it has already incorporated several popular Web resources and external tools into its environment. Interoperability is ensured through the format DocuXML.

In addition to describing the design principles of DocuSky, we will show its main features, together with several important tools and examples. DocuSky was originally developed for Sinological studies. We are enriching it work in other languages.

Rationale for Docusky[1]

The past two decades have seen many digitization projects, resulting in a large number of digital archives available over the Web. Many of these digital collections not only have high-quality digital content, but are also equipped with sophisticated tools for analyzing the content and exploring the textual contexts of (relationships between) documents. Some examples in Sinology are ctext.org[2] for the Chinese classics, CBETA[3] for the Buddhist canon, and THDL[4] (Taiwan History Digital Library) for Taiwanese historical documents.

*Department of Computer Science and Information Engineering, National Taiwan University, and Research Center for Digital Humanities, National Taiwan University.

[1] https://docusky.org.tw

[2] Donald Sturgeon, "Chinese Text Project," 2020, https://ctext.org.

[3] Chinese Buddhist Electronic Text Association, "Cbeta," 2020, https://www.cbeta.org/.

[4] Research Center for Digital Humanities of NTU, "Taiwan History Digital Library (THDL)," 2020, doi:10.6681/NTURCDH.DB_THDL/Text.

Although these digital archive systems significantly expanded a scholar's ability to access and analyze research material, they are not without drawbacks. There are two main problems. The first of these is the difficulty of dealing with a researcher's personal collection of data. A digital archive system is a *closed* system. Its content is collected and organized according to certain criteria, and tools are designed to work with its content. The content and tools interact within the confines of the system itself, while adding new content and new tools can only be done by the developers (or administrators) of the system. However, scholars often have their own personal collection of research material. Even if a tool in a specific archive system fits a scholar's need precisely, it is difficult to apply that tool to contents that are not already in the system. The second problem is the inability to cater to specific research problems. Scholars usually have their own research objectives. Different objectives lead to different emphases when collecting research material. On the other hand, the digital content of a digital archive system is often developed around a specific domain of interest, and its collection strategy is usually to make it as comprehensive as possible so that it can be used by scholars for different research purposes. In doing so, however, a digital archive system becomes inherently *data-centered*, and cannot cater to individual research needs.

To meet the individual needs of humanities scholars while maintaining the benefit of digital archive systems, we have developed a personal digital humanities platform called DocuSky.[5] DocuSky allows users to upload their own data and supplement them with additional data from certain web resource sites; it provides processing tools for the user to prepare/enrich the data (such as adding metadata or tagging) and a one-click function to build a full-text searchable database with textual context discovery. In addition to searching, retrieving, and exploring the built-in context discovery mechanisms in the database, the user can also use additional tools provided by DocuSky for analysis and visualization to explore further.

A central design philosophy behind DocuSky is the separation of content and tools. This allows the scholar (user) to have full control of the data she has in hand. To realize this methodology, we needed to design a way to represent the data and to let content and tools communicate with each other. We call this format DocuXML.[6] DocuXML provides an XML format of pre-defined metadata attributes as well as tags for annotation. Users can also define their own tags and metadata. Not only can content and tools be connected through DocuXML, it can also connect humanities scholars and IT researchers. A scholar can describe the tool that she needs, and as long as a tool uses a data format that is accepted by DocuSky, it can be used within the DocuSky platform freely. Another core technology in DocuSky is its ability to convert any file in DocuXML format into a searchable text database.

Building a Full-Text Searchable Database in Docusky

Scholars may amass files to the point of unmanageability on a local hard disk. A key feature of DocuSky is its ability to transform that set of documents into a full-text searchable database. Because the database building function of DocuSky requires files in DocuXML format, DocuSky provides facilities for converting several formats into DocuXML. Documents in text format, are input into the *text2DocuXML* tool and

[5] Hsieh-Chang Tu and Jieh Hsiang, "Docusky Collaboration Platform," Research Center for Digital Humanities of NTU, 2018–2020, https://docusky.org.tw.

[6] Research Center for Digital Humanities of NTU, "DocuXML 1.2 Scheme," http://docusky.org.tw/DocuSky/documentation/docs/DocuXml-1.2-Scheme.html.

with just one click it generates a file with the documents in DocuXML format. When that is uploaded to the *db-builder* tool in DocuSky it produces a full-text searchable database with the original file names as natural delimiters separating the texts. The user can also declare separate groupings of files (such as using the folder names where the files are stored) so that each grouping becomes a different corpus and can serve as an attribute for post-classification of search results. This simple, powerful construct allows the user to create a full-text searchable personal database with a few clicks.

It is also straightforward to convert a spreadsheet (in Microsoft Excel or csv format using the the *xls/csv2DocuXML* tool to transform the spreadsheet into DocuXML format. In this case the user also must declare the correspondence of her own attributes with the metadata attributes in DocuXML. Then, using the same *db-builder* tool, she can immediately create a personal database of the spreadsheet. In addition to full-text search, such a database has an additional feature of using the metadata attributes to post-classify the query result. For instance, if "year" is a metadata attribute, the query result can be ordered chronologically. We call such relationships *contexts from metadata*.

Another important scholarly activity is to annotate or tag person names, place names, or any chosen set of terms. This is particularly important for a language like Chinese which lacks natural delimiters such as the word-separating whitespace in Western languages. Among the tagging tools provided in the platform, DocuSky encourages the use of Markus,[7] a popular tool for annotation developed at Leiden University, as the main tagging/annotation tool. Once documents are annotated using Markus, the resulting files can be converted into DocuXML format via the *M2D* conversion tool. The *db-builder* tool can then build a personal database from the files. The database built from texts with Markus tagging has the additional feature of *contexts from tagging*: relations that present the textual contexts of tags. (For instance, the frequencies of person names appearing in a query result, or the locations on a map if geographic coordinates are given in tags.)

Textual Context Discovery

These textual contexts reflect the design methodology of THDL. Most retrieval systems are based on the precision/recall model, which treats documents as independent entities. Scholars, however, usually consider documents as related and try to find the relationships (context) among documents[8]. Thus, instead of ranking the documents in the query result, THDL assumes the documents in the query result are related and tries to find the textual contexts (relationships). The aforementioned *contexts from metadata* and *contexts from tagging* are two important types of textual contexts incorporated in THDL. Other contexts include *statistical contexts* (n-gram analysis, for example) and *semantic contexts*. Semantic contexts are domain-dependent. In THDL we provided

[7] Hou Ieong Brent Ho and Hilde De Weerdt, MARKUS. Text Analysis and Reading Platform. 2014–, https://dh.chinese-empires.eu/beta/. Funded by the European Research Council and the Digging into Data Challenge.

[8] J. Hsiang and C.A. Weng. "Multiple-Contextualization: Problems and Challenges on Digital Archives," *Essential Digital Humanities: Defining Patterns and Paths*, edited by Jieh Hsiang (Taipei: NTU Press, 2012), pp. 25–60.

two such contexts: the Citation Graph of Imperial Decrees and Memorials[9] and Land Transaction Relation of land deeds[10].

Except for domain-dependent semantic contexts, all the textual contexts are incorporated into any database generated from DocuSky. Another very useful statistical context is *Simple Concordance Analysis*, which allows the user to query a keyword with a prefix or postfix number of words, or any string that begins with a keyword and ends with another. For example, if one designates a keyword with a prefix of 2, the system will retrieve all terms with the keyword and the preceding two characters (with their frequencies).

A Working Example

In the following we show an example of a text database, generated by DocuSky, that was jointly created with Dr. Michael Stanley-Baker[11] when he was working at the Max Planck Institute in Berlin. The example is *Bencaojing jizhu* (本草經集注), a sixteh-century work by Tao Hongjing (陶弘景). The book contains 730 herbal medicines and is among the most studied Chinese medicine books. The full text of *Bencaojing jizhu* was downloaded from Wikisource.[12] Stanley-Baker defined a set tags, such as #DiseaseName, #DrugName, #Drug_Action, each with a list of associated terms. Using Markus he tagged these terms in the text; person names (using CBDB[13] and BSPAD[14]) and place names (using CHGIS[15]) are also tagged. The metadata are simply the original chapters and sections of the source book. The tagged file was then converted to DocuXML (with one click) and fed to *db_builder* (another click) to create the following database.

Figure 1 shows the resulting database, in which the left column is the classification of results from a query for the tag #Drug_Action.

Clicking on "tag cloud" shows the same post-classification in tag cloud form (Figure 2).

Clicking on the "show places on the map" button produces Figure 3, with all locations names shown on the map.

By clicking on a specific place name, its location will be highlighted on the map. Figure 4 shows the same interface by clicking on Lantian (藍田).

The following example shows the efficacy of the *Simple Concordance* context. It is a common practice in modern Chinese to put book titles in double brackets. For example, The Analects (論語) of Confucius will be presented as 《論語》. This convention was also used in the full-text of Wikisource. We take advantage of this property, and issue a

[9] Jieh Hsiang, Shih-Pei Chen, Hou-Ieong Ho, and Hsieh-Chang Tu, "Discovering Relationships from Imperial Court Documents of Qing Dynasty." *International Journal of Humanities and Arts Computing* 6.1–2 (2012), 22–41.

[10] Shih-Pei Chen, et al, "Discovering Land Transaction Relations from Land Deeds of Taiwan," *Literary and Linguistic Computing* 28.2 (2013), 257–270. Available at www.researchgate.net/publication/260021449_Discovering_land_transaction_relations_from_land_deeds_of_Taiwan.

[11] Michael Stanley-Baker, "Medicine and Religion in China," 2020, https://michaelstanley-baker.com/posts/.

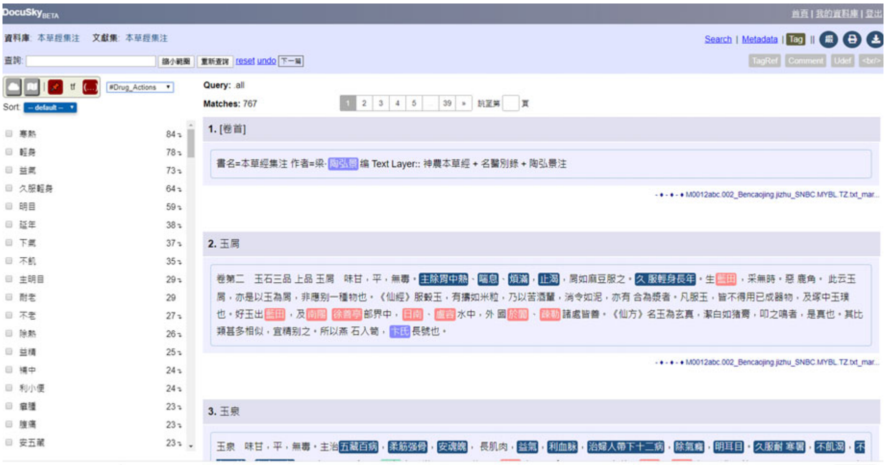
[12] https://zh.wikisource.org/zh-hant/ zh-hant/本草經集注.

[13] "Chinese Bibliographic Database." Harvard University, 201,. https://projects.iq.harvard.edu/chinesecbdb/home.

[14] DILA, "Buddhist Studies Authority Database Project, DILA, 2008, https://authority.dila.edu.tw/.

[15] CHGIS Harvard Office and HGIS Center, Fudan University, "Chinese Historical GIS," 2001, http://chgis.fas.harvard.edu/.

Fig. 1 - Colour online, B/W in print



DocuSky BETA　首頁 | 我的資料庫 | 登出

資料庫: 本草經集注　文獻集: 本草經集注

查詢: 縮小範圍 重新查詢 reset undo 下一篇

Search | Metadata | Tag || TagRef Comment Udef <br/>

#Drug_Actions

Sort: -- default --

Query: .all

Matches: 767

1 2 3 4 5 ... 39 » 跳至第 頁

□ 寒熱	84
□ 輕身	78
□ 益氣	73
□ 久服輕身	64
□ 明目	59
□ 延年	38
□ 下氣	37
□ 不飢	35
□ 主明目	29
□ 耐老	29
□ 不老	27
□ 除熱	26
□ 益精	25
□ 補中	24
□ 利小便	24
□ 癰腫	23
□ 腹痛	23
□ 安五藏	23

1. [卷首]

書名=本草經集注 作者=梁·陶弘景 编 Text Layer:: 神農本草經 + 名醫別錄 + 陶弘景注

-◆-◆-◆ M0012abc.002_Bencaojing jizhu_SNBC.MYBL.TZ.txt_mar...

2. 玉屑

卷第二　玉石三品 上品 玉屑　味甘，平，無毒。主除胃中熱、喘息、煩滿，止渴，屑如麻豆服之。久服輕身長年。生藍田，采無時。惡鹿角。此云玉屑，亦是以玉為屑，非應別一種物也。《仙經》服穀玉，有擣如米粒，乃以苦酒輩，消令如泥，亦有合為漿者。凡服玉，皆不得用已成器物，及塚中玉璞也。好玉出藍田，及南陽徐善亭部界中，日南、盧容水中，外國於闐、疏勒諸處皆善。《仙方》名玉為玄真，潔白如豬膏，叩之鳴者，是真也。其比類甚多相似，宜精別之。所以燕石入笥，卞氏長號也。

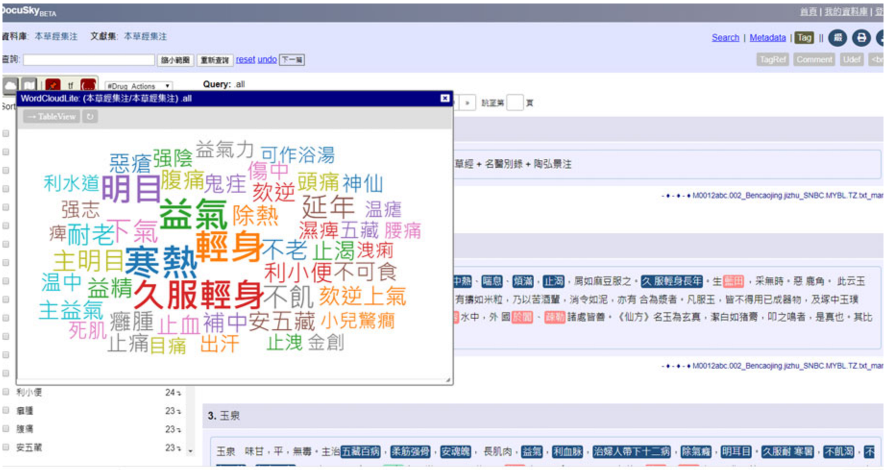
-◆-◆-◆ M0012abc.002_Bencaojing jizhu_SNBC.MYBL.TZ.txt_mar...

3. 玉泉

玉泉　味甘，平，無毒。主治五藏百病，柔筋強骨，安魂魄，長肌肉，益氣，利血脈，治婦人帶下十二病，除氣癃，明耳目。久服耐寒暑，不飢渴，不

Figure 1. The *Bencaojingjizhu* database generated by DocuSky

Fig. 2 - Colour online, B/W in print



DocuSky BETA　首頁 | 我的資料庫 | 登

資料庫: 本草經集注　文獻集: 本草經集注

查詢: 縮小範圍 重新查詢 reset undo 下一篇

Search | Metadata | Tag || TagRef Comment Udef <br

#Drug_Actions

Query: .all

WordCloudLite: (本草經集注/本草經集注) .all

→ TableView

□ 利小便	24
□ 癰腫	23
□ 腹痛	23
□ 安五藏	23

草經 + 名醫別錄 + 陶弘景注

-◆-◆-◆ M0012abc.002_Bencaojing jizhu_SNBC.MYBL.TZ.txt_mar

中熱、喘息、煩滿，止渴，屑如麻豆服之。久服輕身長年。生藍田，采無時。惡鹿角。此云玉
有擣如米粒，乃以苦酒輩，消令如泥，亦有合為漿者。凡服玉，皆不得用已成器物，及塚中玉璞
水中，外國於闐、疏勒諸處皆善。《仙方》名玉為玄真，潔白如豬膏，叩之鳴者，是真也。其比

-◆-◆-◆ M0012abc.002_Bencaojing jizhu_SNBC.MYBL.TZ.txt_mar

3. 玉泉

玉泉　味甘，平，無毒。主治五藏百病，柔筋強骨，安魂魄，長肌肉，益氣，利血脈，治婦人帶下十二病，除氣癃，明耳目。久服耐寒暑，不飢渴，不

Figure 2. The tag cloud feature

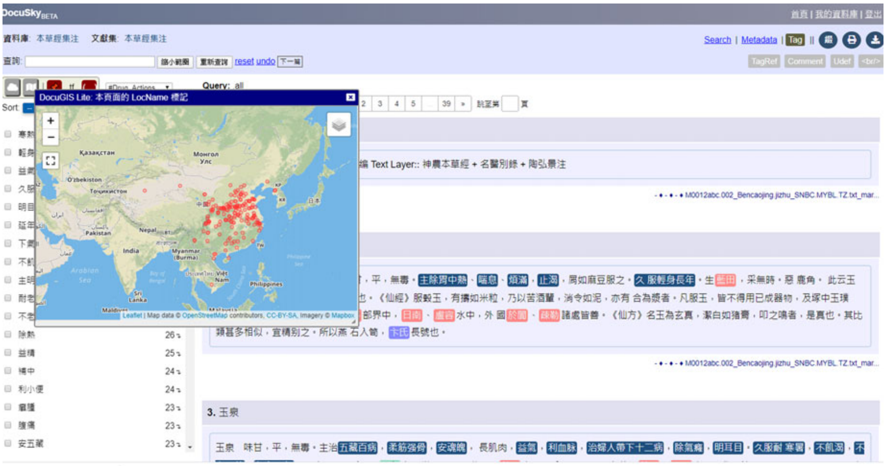
query that asks for any string of up to 20 characters that is between 《 and 》 thus generating a set of all the books (with titles up to 20 characters) cited in the *Bencaojing jizhu* with the frequencies of appearance. Figure 5 is the image of the query, and Figure 6 shows the resulting set.

The complete list of books cited in *Bencaojing jizhu* with frequencies of occurrence is given in Figure 7.

Utilizing External Digital Sources in Docusky

There are many high quality, free Sinological resources on the Web. In addition to creating their own content, scholars can also enrich their data by extracting material from Web

Fig. 3 - Colour online, B/W in print



DocuSky BETA 首頁 | 我的資料庫 | 登出

資料庫: 本草經集注 文獻集: 本草經集注

查詢: 縮小範圍 重新查詢 reset undo 下一篇

Search | Metadata | Tag

TagRef Comment Udef <br/>

#Drug_Actions

Query: all

DocuGIS Lite: 本頁面的 LocName 標記

2 3 4 5 … 39 » 跳至第 頁

…編 Text Layer:: 神農本草經 + 名醫別錄 + 陶弘景注

M0012abc.002_Bencaojing.jizhu_SNBC.MYBL.TZ.txt_mar...

…，平，無毒，主除胃中熱，喘息，煩滿，止渴，屑如麻豆服之，久服輕身長年，生藍田，采無時，惡鹿角，此云玉…也，《仙經》服穀玉，有擣如米粒，乃以苦酒輩，消令如泥，亦有合為漿者，凡服玉，皆不得用已成器物，及塚中玉璞…部界中，日南，盧容水中，外國於闐，疎勒諸處皆善，《仙方》名玉為玄真，潔白如豬膏，叩之鳴者，是真也，其比…類甚多相似，宜精別之，所以燕石入笥，卞氏長號也，

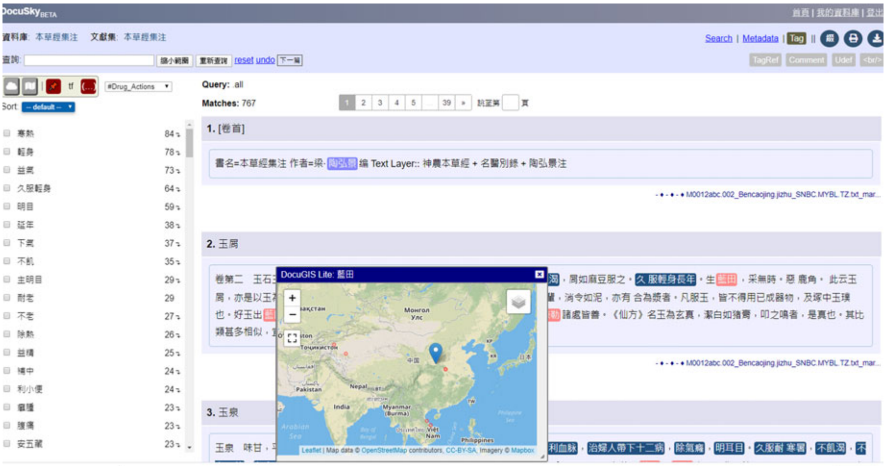
M0012abc.002_Bencaojing.jizhu_SNBC.MYBL.TZ.txt_mar...

3. 玉泉

玉泉 味甘，平，無毒，主治五藏百病，柔筋強骨，安魂魄，長肌肉，益氣，利血脉，治婦人帶下十二病，除氣癃，明耳目，久服耐寒暑，不飢渴，不…

Figure 3. The GIS feature built into the database

Fig. 4 - Colour online, B/W in print



DocuSky BETA 首頁 | 我的資料庫 | 登出

資料庫: 本草經集注 文獻集: 本草經集注

查詢: 縮小範圍 重新查詢 reset undo 下一篇

Search | Metadata | Tag

TagRef Comment Udef <br/>

#Drug_Actions

Sort: -- default --

Query: all

Matches: 767

1 2 3 4 5 … 39 » 跳至第 頁

□ 寒熱	84
□ 輕身	78
□ 益氣	73
□ 久服輕身	64
□ 明目	59
□ 延年	38
□ 下氣	37
□ 不飢	35
□ 主明目	29
□ 耐老	29
□ 不老	27
□ 除熱	26
□ 益精	25
□ 補中	24
□ 利小便	24
□ 癰腫	23
□ 腹痛	23
□ 安五藏	23

1. [卷首]

書名=本草經集注 作者=梁·陶弘景編 Text Layer:: 神農本草經 + 名醫別錄 + 陶弘景注

M0012abc.002_Bencaojing.jizhu_SNBC.MYBL.TZ.txt_mar...

2. 玉屑

卷第二 玉石… …，屑如麻豆服之，久服輕身長年，生藍田，采無時，惡鹿角，此云玉屑，亦是以玉… …輩，消令如泥，亦有合為漿者，凡服玉，皆不得用已成器物，及塚中玉璞也，好玉出藍… …諸處皆善，《仙方》名玉為玄真，潔白如豬膏，叩之鳴者，是真也，其比類甚多相似，…

M0012abc.002_Bencaojing.jizhu_SNBC.MYBL.TZ.txt_mar...

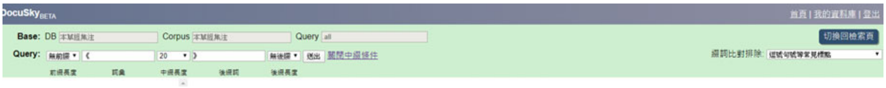
3. 玉泉

玉泉 味甘，… …利血脉，治婦人帶下十二病，除氣癃，明耳目，久服耐寒暑，不飢渴，不…

DocuGIS Lite: 藍田

Figure 4. Showing the location of Lantian

Fig. 5 - Colour online, B/W in print



DocuSky BETA 首頁 | 我的資料庫 | 登出

Base: DB 本草經集注 Corpus 本草經集注 Query all

切換回檢索頁

Query: 無前標 《 20 》 無後標 送出 關閉中繼條件

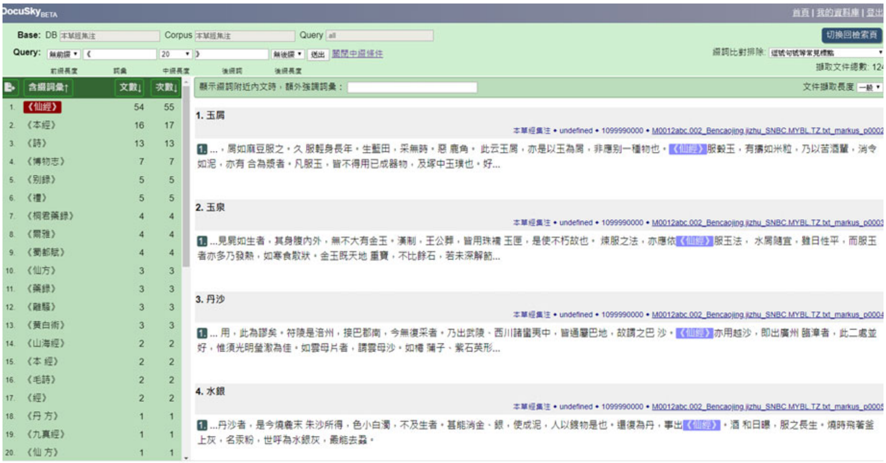
前綴長度 詞彙 中綴長度 後綴詞 後綴長度

綴詞比對排除: 逗號句號等常見標點

Figure 5. The *Simple Concordance* query box

resources through DocuSky. The interaction between an external source and DocuSky is performed through a Web API (Application Programming Interface). For each resource we need to build a separate API. The following are available at the moment:

Fig. 6 - Colour online, B/W in print



Base: DB 本草經集注 Corpus 本草經集注 Query all

Query: 無前綴 《 20 》 無後綴

綴詞比對排除: 逗號句號等常見標點

擷取文件總數: 124

顯示綴詞附近內文時，額外強調詞彙：

文件擷取長度 一般

	含綴詞彙↑	文數↓	次數↓
1.	《仙經》	54	55
2.	《本經》	16	17
3.	《詩》	13	13
4.	《博物志》	7	7
5.	《別錄》	5	5
6.	《禮》	5	5
7.	《桐君藥錄》	4	4
8.	《爾雅》	4	4
9.	《蜀都賦》	4	4
10.	《仙方》	3	3
11.	《藥錄》	3	3
12.	《離騷》	3	3
13.	《黃白術》	3	3
14.	《山海經》	2	2
15.	《本 經》	2	2
16.	《毛詩》	2	2
17.	《經》	2	2
18.	《丹 方》	1	1
19.	《九真經》	1	1
20.	《仙 方》	1	1

1. 玉屑

本草經集注 • undefined • 1099990000 • M0012abc.002_Bencaojing.jizhu_SNBC.MYBL.TZ.txt_markus_p0002

1. …，屑如麻豆服之，久 服輕身長年，生藍田，采無時，惡 鹿角， 此云玉屑，亦是以玉為屑，非應別一種物也，《仙經》服穀玉，有搗如米粒，乃以苦酒輩，消令如泥，亦有 合為漿者，凡服玉，皆不得用已成器物，及塚中玉璞也，好…

2. 玉泉

本草經集注 • undefined • 1099990000 • M0012abc.002_Bencaojing.jizhu_SNBC.MYBL.TZ.txt_markus_p0003

1. …見屍如生者，其身腹內外，無不大有金玉，漢制，王公葬，皆用珠襦 玉匣，是使不朽故也， 煉服之法，亦應依《仙經》服玉法， 水屑隨宜，雖日性平，而服玉者亦多乃發熱，如寒食散狀，金玉既天地 重寶，不比餘石，若未深解節…

3. 丹沙

本草經集注 • undefined • 1099990000 • M0012abc.002_Bencaojing.jizhu_SNBC.MYBL.TZ.txt_markus_p0004

1. … 用，此為謬矣，符陵是涪州，接巴郡南，今無復采者，乃出武陵、西川諸蠻夷中，皆通屬巴地，故謂之巴 沙，《仙經》亦用越沙，即出廣州 臨漳者，此二處並好，惟須光明瑩澈為佳，如雲母片者，謂雲母沙，如樗 蒲子、紫石英形…

4. 水銀

本草經集注 • undefined • 1099990000 • M0012abc.002_Bencaojing.jizhu_SNBC.MYBL.TZ.txt_markus_p0005

1. …丹沙者，是今燒麤末 朱沙所得，色小白濁，不及生者，甚能消金、銀，使成泥，人以鍍物是也，還復為丹，事出《仙經》，酒 和日曝，服之長生，燒時飛著釜上灰，名汞粉，世呼為水銀灰，最能去蝨，

Figure 6. Outcome of the *Simple Concordance* query

1. Ctext.org (https://www.ctext.org): Download plain text through API.
2. Kanripo (https://www.kanripo.org): Download plain text with simple metadata through API
3. CBETA (https://www.cbeta.org): Download files of texts from CBETA through API. The files will automatically be converted into DocuXML format[16].
4. THDL (http://thdl.ntu.edu.tw): Files in THDL can be exported directly into DocuSky in DocuXML format through API.

We are also preparing an API to extract files directly from Wikisource while preserving their original structure.

Once data are prepared and the database built, a number of tools can be utilized to make further analysis and visualization. Below we describe several useful tools.

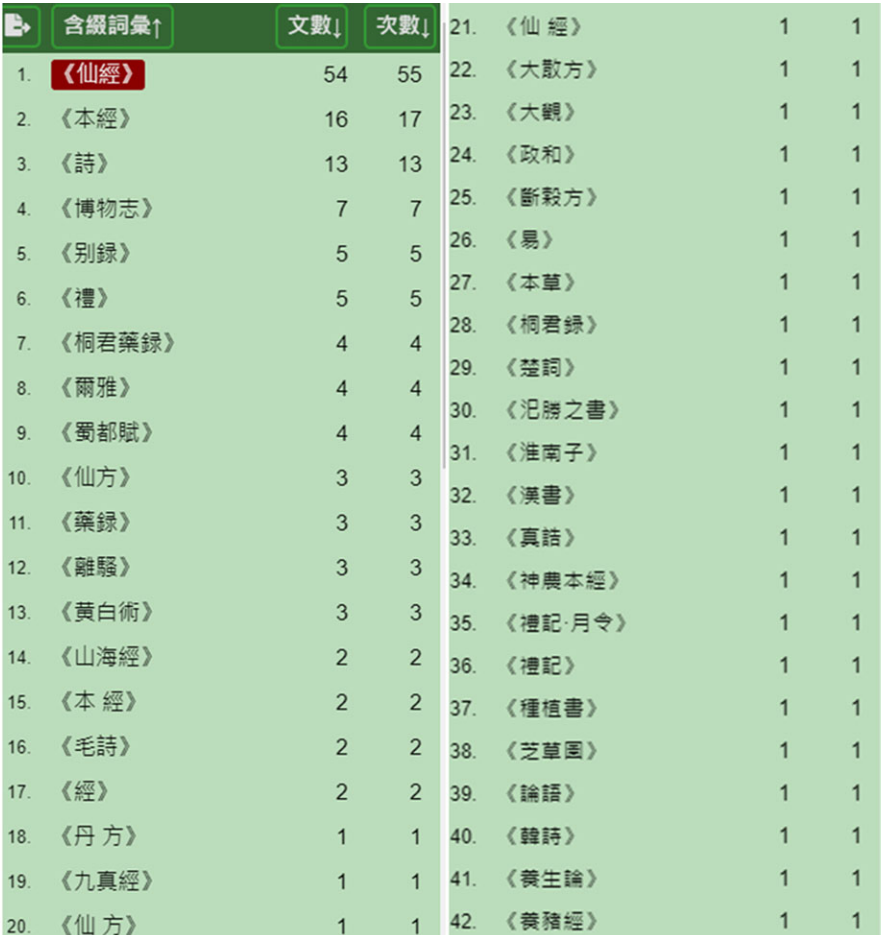
Docugis, the GIS Tool in Docusky

Map visualizations using a GIS (geographic information system) is useful but hard to master for many. We have developed a GIS tool in DocuSky called DocuGIS.[17] DocuGIS has several unique features. First, given a spreadsheet of locations with coordinates, it can plot the points on the map with just one-click. In other words, anyone who knows how to prepare a spreadsheet can draw a GIS map in a few minutes. Second, it is *interactive*. If the spreadsheet contains a column of text, then DocuGIS allows full-text search and highlighting the results on the map. Third, DocuGIS has also collected a number of historical maps so that results can be shown on historical maps.

In the following we use the example of Xuanzang's (玄奘) seventh-century *Pilgrimage to the West during Tang Dynasty* (大唐西域記). We first obtain the full-text

[16] The API (with the conversion to DocuXML) was done by JenJou Hong of DILA (Dharma Dram Institute of Liberal Arts) and CBETA. His help is gratefully acknowledged.

[17] Designed and implemented by Nungyao Lin. See Nung Yao Lin, *Textual Geographic Information System and Applications*, Graduate Institute of Networking and Multimedia, National Taiwan University, 2018. (In Chinese).



	含綴詞彙↑	文數↓	次數↓
1.	《仙經》	54	55
2.	《本經》	16	17
3.	《詩》	13	13
4.	《博物志》	7	7
5.	《别録》	5	5
6.	《禮》	5	5
7.	《桐君藥録》	4	4
8.	《爾雅》	4	4
9.	《蜀都賦》	4	4
10.	《仙方》	3	3
11.	《藥録》	3	3
12.	《離騷》	3	3
13.	《黄白術》	3	3
14.	《山海經》	2	2
15.	《本 經》	2	2
16.	《毛詩》	2	2
17.	《經》	2	2
18.	《丹 方》	1	1
19.	《九真經》	1	1
20.	《仙 方》	1	1
21.	《仙 經》	1	1
22.	《大散方》	1	1
23.	《大觀》	1	1
24.	《政和》	1	1
25.	《斷穀方》	1	1
26.	《易》	1	1
27.	《本草》	1	1
28.	《桐君録》	1	1
29.	《楚詞》	1	1
30.	《氾勝之書》	1	1
31.	《淮南子》	1	1
32.	《漢書》	1	1
33.	《真誥》	1	1
34.	《神農本經》	1	1
35.	《禮記·月令》	1	1
36.	《禮記》	1	1
37.	《種植書》	1	1
38.	《芝草圖》	1	1
39.	《論語》	1	1
40.	《韓詩》	1	1
41.	《養生論》	1	1
42.	《養豬經》	1	1

Figure 7. Complete list of books cited in *Bencaojingjizhu*

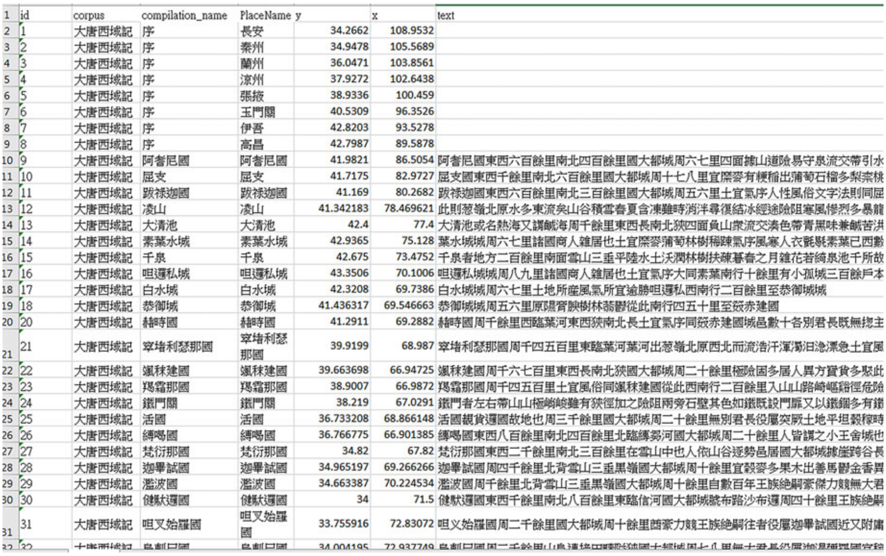
Fig. 7 - Colour online, B/W in print

of the book from ctext.org. Since the book was written in the order of cities (countries) that Xuanzang visited, we divide the text into segments accordingly. We then create an Excel spreadsheet, with each of the row containing a numerical *id* (in the order of appearance in the book), *CompilationName* (section header in the book), *PlaceName* (name of place in the book), *Coordinates* (y, x), *Text* (the text of the segment), as shown in Figure 8.

The data in the Excel sheet can be easily converted into csv format,[18] which is then fed into DocuGIS (Figure 9) though the "upload csv" option.

This action immediately creates a map with all locations in the book highlighted. Clicking on any of the points shows the corresponding text in that entry (Figure 10).

[18] This can be done by simply click on the top-left corner of the Excel sheet. CSV, or Comma-Separated Values, is a plain text file in which each record (of the original Excel file) occupies a line, with the records of different fields separated by commas.



1	id	corpus	compilation_name	PlaceName	y	x	text
2	1	大唐西域記	序	長安	34.2662	108.9532	
3	2	大唐西域記	序	秦州	34.9478	105.5689	
4	3	大唐西域記	序	蘭州	36.0471	103.8561	
5	4	大唐西域記	序	涼州	37.9272	102.6438	
6	5	大唐西域記	序	張掖	38.9336	100.459	
7	6	大唐西域記	序	玉門關	40.5309	96.3526	
8	7	大唐西域記	序	伊吾	42.8203	93.5278	
9	8	大唐西域記	序	高昌	42.7987	89.5878	
10	9	大唐西域記	阿耆尼國	阿耆尼國	41.9821	86.5054	阿耆尼國東西六百餘里南北四百餘里國大都城周六七里四面據山道險易守泉流交帶引水
11	10	大唐西域記	屈支	屈支	41.7175	82.9727	屈支國東西千餘里南北六百餘里國大都城周十七八里宜穈麥有粳稻出蒲萄石榴多梨柰桃
12	11	大唐西域記	跋祿迦國	跋祿迦國	41.169	80.2682	跋祿迦國東西六百餘里南北三百餘里國大都城周五六里土宜氣序人性風俗文字法則同屈
13	12	大唐西域記	凌山	凌山	41.342183	78.469621	此則葱嶺北原水多東流矣山谷積雪春夏含凍雖時消泮尋復結冰經途險阻寒風慘烈多暴龍
14	13	大唐西域記	大清池	大清池	42.4	77.4	大清池或名熱海又謂鹹海周千餘里東西長南北狹四面負山衆流交湊色帶青黑味兼鹹苦洪
15	14	大唐西域記	素葉水城	素葉水城	42.9365	75.128	葉水城城周六七里諸國商人雜居也土宜穈麥蒲萄林樹稀疏氣序風寒人衣氈褐素葉已西數
16	15	大唐西域記	千泉	千泉	42.675	73.4752	千泉者地方二百餘里南面雪山三垂平陸水土沃潤林樹扶疎暮春之月雜花若綺泉池千所故
17	16	大唐西域記	呾邏私城	呾邏私城	43.3506	70.1006	呾邏私城城周八九里諸國商人雜居也土宜氣序大同素葉南行十餘里有小孤城三百餘戶本
18	17	大唐西域記	白水城	白水城	42.3208	69.7386	白水城城周六七里土地所產風氣所宜逾勝呾邏私西南行二百餘里至恭御城城
19	18	大唐西域記	恭御城	恭御城	41.436317	69.546663	恭御城城周五六里原隰膏腴樹林蓊鬱從此南行四五十里至笯赤建國
20	20	大唐西域記	赭時國	赭時國	41.2911	69.2882	赭時國周千餘里西臨葉河東西狹南北長土宜氣序同笯赤建國城邑數十各別君長既無總主
21	21	大唐西域記	窣堵利瑟那國	窣堵利瑟那國	39.9199	68.987	窣堵利瑟那國周千四五百里東臨葉河葉河出葱嶺北原西北而流浩汗渾濁汩淴漂急土宜風
22	22	大唐西域記	颯秣建國	颯秣建國	39.663698	66.94725	颯秣建國周千六七百里東西長南北狹國大都城周二十餘里極險固多居人異方寶貨多聚此
23	23	大唐西域記	羯霜那國	羯霜那國	38.9007	66.9872	羯霜那國周千四五百里土宜風俗同颯秣建國從此西南行二百餘里入山山路崎嶇谿徑危險
24	24	大唐西域記	鐵門關	鐵門關	38.219	67.0291	鐵門者左右帶山山極峭峻雖有狹徑加之險阻兩旁石壁其色如鐵既設門扉又以鐵錮多有鐵
25	25	大唐西域記	活國	活國	36.733208	68.866148	活國覩貨邏國故地也周三千餘里國大都城周二十餘里無別君長役屬突厥土地平坦穀稼時
26	26	大唐西域記	縛喝國	縛喝國	36.766775	66.901385	縛喝國東西八百餘里南北四百餘里北臨縛芻河國大都城周二十餘里人皆謂之小王舍城也
27	27	大唐西域記	梵衍那國	梵衍那國	34.82	67.82	梵衍那國東西二千餘里南北三百餘里在雪山中也人依山谷逐勢邑居國大都城據崖跨谷長
28	28	大唐西域記	迦畢試國	迦畢試國	34.965197	69.266266	迦畢試國周四千餘里北背雪山三垂黑嶺國大都城周十餘里宜穀麥多果木出善馬鬱金香異
29	29	大唐西域記	濫波國	濫波國	34.663387	70.224534	濫波國周千餘里北背雪山三垂黑嶺國大都城周十餘里自數百年王族絕嗣豪傑力競無大君
30	30	大唐西域記	健馱邏國	健馱邏國	34	71.5	健馱邏國東西千餘里南北八百餘里東臨信度河國大都城號布路沙布邏周四十餘里王族絕嗣
31	31	大唐西域記	呾叉始羅國	呾叉始羅國	33.755916	72.83072	呾义始羅國周二千餘里國大都城周十餘里酋豪力競王族絕嗣往者役屬迦畢試國近又附庸
32	32	大唐西域記	烏剌尸國	烏剌尸國	34.004195	72.937749	

Figure 8. spreadsheet with content from *Pilgrimage to the West during Tang Dynasty*

Fig. 8 - Colour online, B/W in print

Figure 11 showcases three additional features. First, we added a map of Tang administrative divisions.[19] Second, by using the *id* to specify the sequence of points, the locations become a sequence which indicate the route of Xuanzang. (This can be easily done with a couple of clicks in DocuGIS.) Third, the balloons in the same figure also show the query result with keyword Brahman (婆羅門).

The ease with which DocuGIS creates a map from full-text, and the ability to query in a GIS environment, make it unique among GIS tools. DocuGIS can be used in conjunction with DocuSky or independently, without first building a database.

Tools for Term Analysis

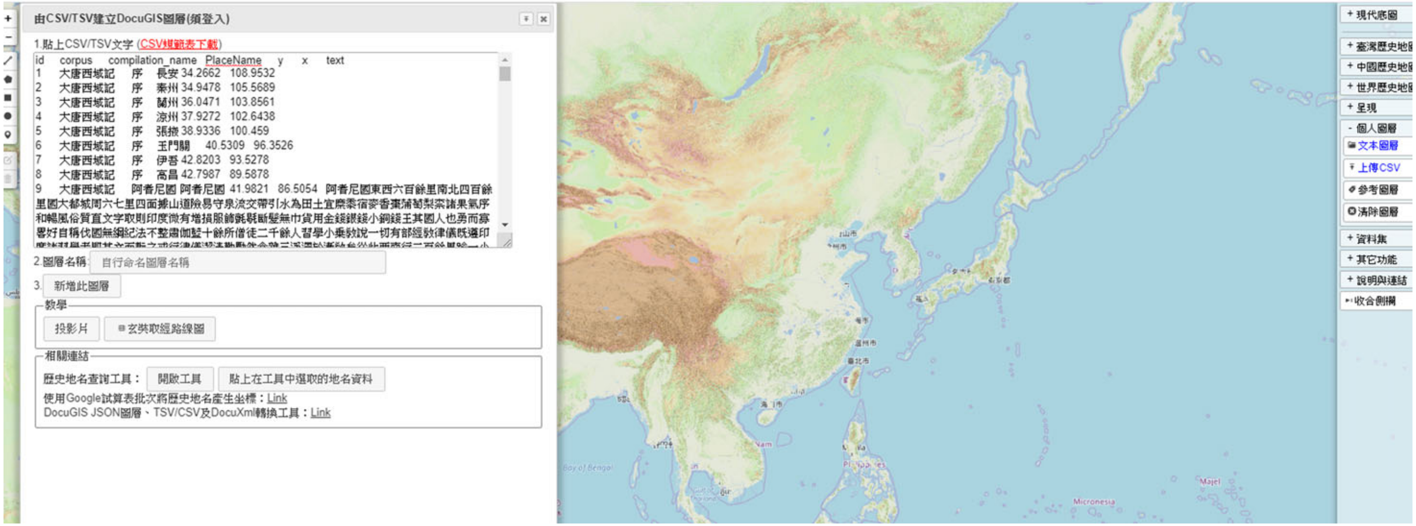
DocuSky also provides a number of text analysis tools, including word clipping for discovering new terms,[20] n-gram analysis, term frequency analysis, text style analysis,[21] and relevant document discovery.[22] In the following we show an example using Hsieh's *TermStat analysis* tool. Given a group of texts and a list of terms, *TermStat* returns the frequencies of terms appearing in each text. Our example wants to see what types of disease were mentioned in the Standard Histories of the Southern and Northern Dynasties (420CE~589CE). There are nine Standard Histories of that era, four for each of the dynasties in the south, three for dynasties in the north, and the

[19] Courtesy of the GIS Center of Academia Sinica.

[20] H. C. Tu, "Semi-Automatic Term Extraction with Simplified Term-Clips Method." *Digital Humanities: Between Past, Present, and Future*, edited by J. Hsiang (Taipei: NTU Press, 2016), 171–206.

[21] All three were developed by Po-Yu Hsieh. See P.Y. Hsieh, "Development and Deployment of Tools Based on Docusky Platform." *The 7th international Conference of Digital Archives and Digital Humanities*, The Research Center for Digital Humanities of NTU, 2016.

[22] Developed by Hsing Hsuan Song. See Hsin-Hsuan Sung, et al, "Finding Documents Related to Taiwan in the Veritable Records of Qing Using Relevance Feedback," *TPDL*, 2019, 280–87.



由CSV/TSV建立DocuGIS圖層(須登入)

1.貼上CSV/TSV文字 (CSV規範表下載)

```
id	corpus	compilation_name	PlaceName	y	x	text
1	大唐西域記	序	長安 34.2662	108.9532
2	大唐西域記	序	秦州 34.9478	105.5689
3	大唐西域記	序	蘭州 36.0471	103.8561
4	大唐西域記	序	涼州 37.9272	102.6438
5	大唐西域記	序	張掖 38.9336	100.459
6	大唐西域記	序	玉門關	40.5309	96.3526
7	大唐西域記	序	伊吾 42.8203	93.5278
8	大唐西域記	序	高昌 42.7987	89.5878
9	大唐西域記	阿耆尼國 阿耆尼國 41.9821	86.5054	阿耆尼國東西六百餘里南北四百餘
里國大都城周六七里四面據山道險易守泉流交帶引水為田土宜穈黍宿麥香棗蒲萄梨柰諸果氣序
和暢風俗質直文字取則印度微有增損服飾氎毼斷髮無巾貨用金錢銀錢小銅錢王其國人也勇而寡
略好自稱伐國無綱紀法不整肅伽藍十餘所僧徒二千餘人習學小乘教說一切有部經教律儀既遵印
```

2.圖層名稱：自行命名圖層名稱

3. 新增此圖層

教學

投影片　玄奘取經路線圖

相關連結

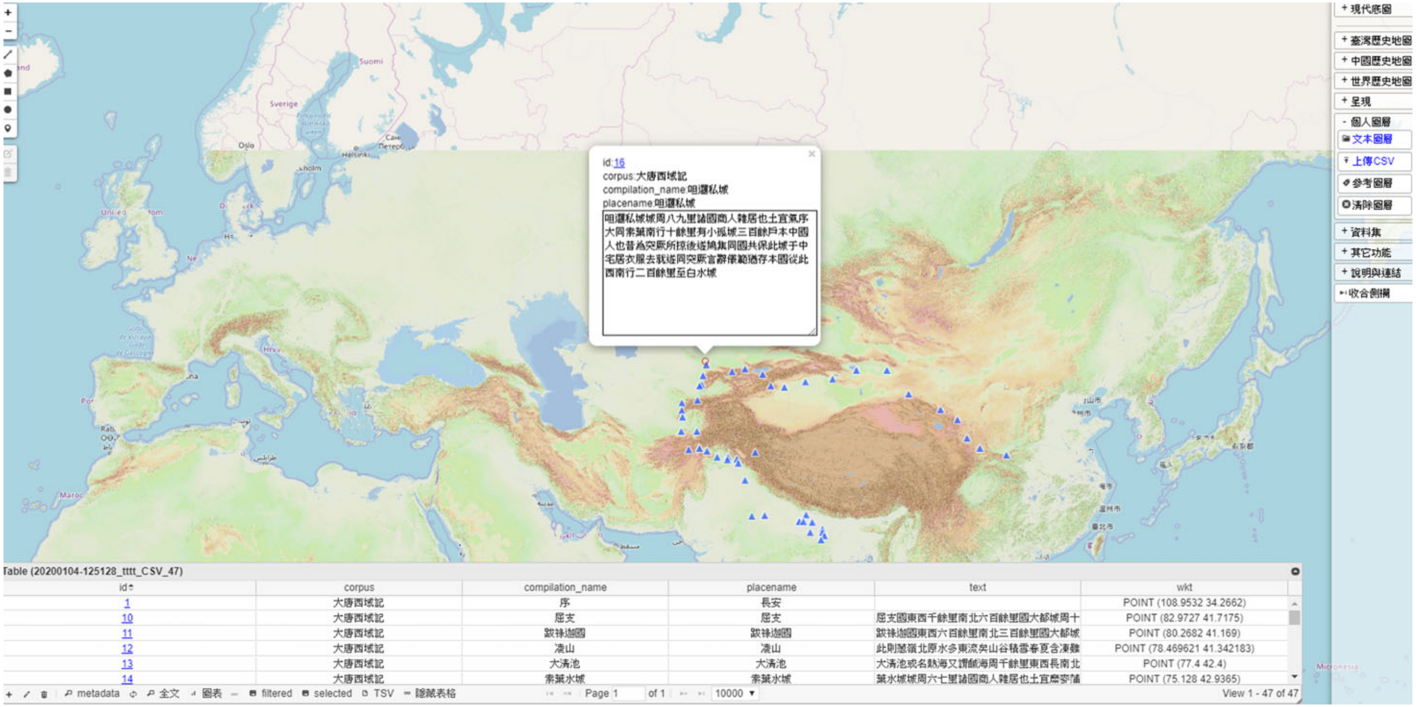
歷史地名查詢工具：開啟工具　貼上在工具中選取的地名資料

使用Google試算表批次將歷史地名產生坐標：Link

DocuGIS JSON圖層、TSV/CSV及DocuXml轉換工具：Link

Figure 9. Uploading csv to DocuGIS

Fig. 9 - Colour online, B/W in print



id:16
corpus:大唐西域記
compilation_name:呾邏私城
placename:呾邏私城

呾邏私城城周八九里諸國商人雜居也土宜氣序大同素葉南行十餘里有小孤城三百餘戶本中國人也昔為突厥所掠後遂鳩集同國共保此城于中宅居衣服去就遂同突厥言辭儀範猶存本國從此西南行二百餘里至白水城

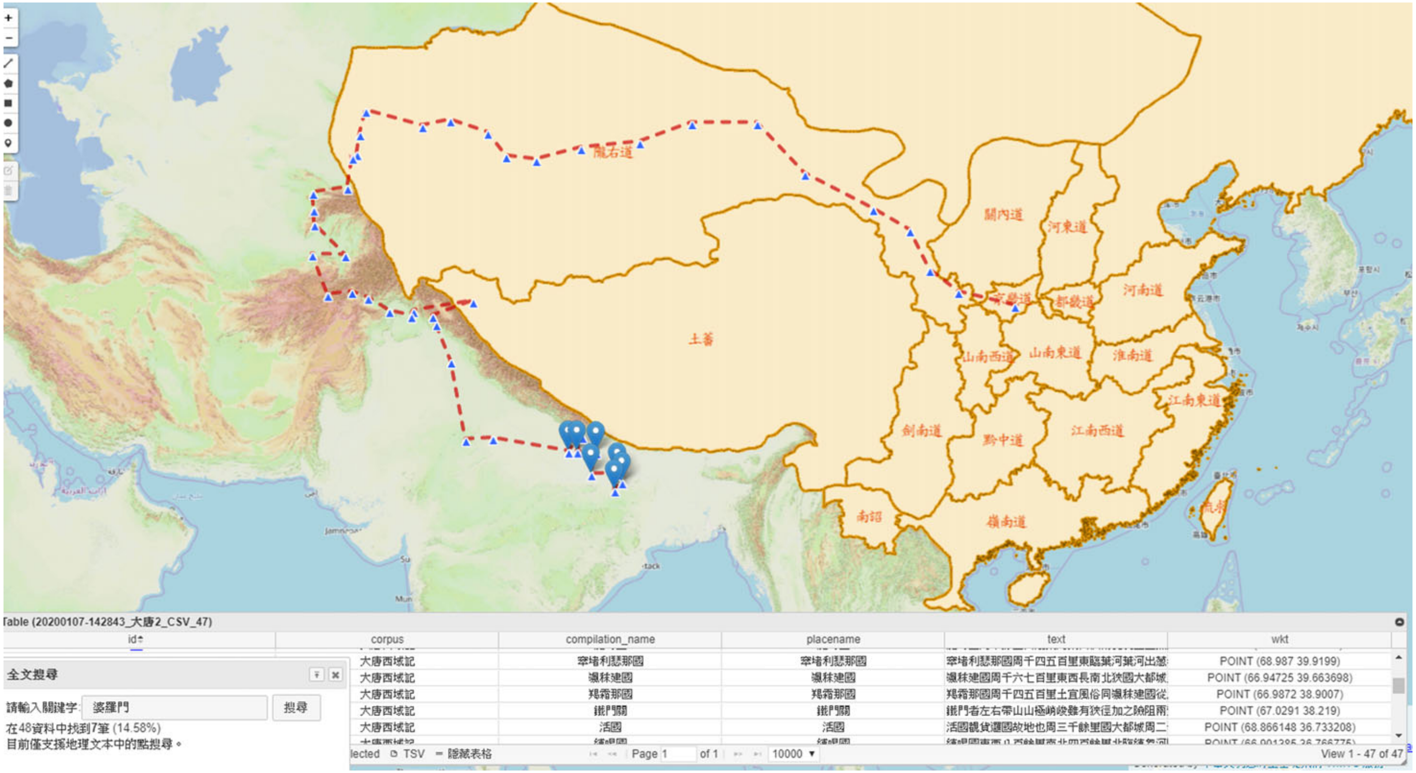
Table (20200104-125128_tttt_CSV_47)

id	corpus	compilation_name	placename	text	wkt
1	大唐西域記	序	長安		POINT (108.9532 34.2662)
10	大唐西域記	屈支	屈支	屈支國東西千餘里南北六百餘里國大都城周十	POINT (82.9727 41.7175)
11	大唐西域記	跋祿迦國	跋祿迦國	跋祿迦國東西六百餘里南北三百餘里國大都城	POINT (80.2682 41.169)
12	大唐西域記	凌山	凌山	此則葱嶺北原水多東流矣山谷積雪春夏含凍雖	POINT (78.469621 41.342183)
13	大唐西域記	大清池	大清池	大清池或名熱海又謂鹹海周千餘里東西長南北	POINT (77.4 42.4)
14	大唐西域記	素葉水城	素葉水城	葉水城城周六七里諸國商人雜居也土宜糜麥蒲	POINT (75.128 42.9365)

+ metadata 全文 圖表 filtered selected TSV 隱藏表格 Page 1 of 1 10000 View 1 - 47 of 47

Figure 10. Resulting interactive GIS system

Fig. 10 - Colour online, B/W in print



全文搜尋

請輸入關鍵字: 婆羅門 搜尋

在48資料中找到7筆 (14.58%)
目前僅支援地理文本中的點搜尋。

id	corpus	compilation_name	placename	text	wkt
	大唐西域記	窣堵利瑟那國	窣堵利瑟那國	窣堵利瑟那國周千四五百里東臨葉河葉河出蔥	POINT (68.987 39.9199)
	大唐西域記	颯秣建國	颯秣建國	颯秣建國周千六七百里東西長南北狹國大都城	POINT (66.94725 39.663698)
	大唐西域記	羯霜那國	羯霜那國	羯霜那國周千四五百里土宜風俗同颯秣建國從	POINT (66.9872 38.9007)
	大唐西域記	鐵門關	鐵門關	鐵門者左右帶山山極峭峻雖有狹徑加之險阻兩	POINT (67.0291 38.219)
	大唐西域記	活國	活國	活國覩貨邏國故地也周三千餘里國大都城周二	POINT (68.866148 36.733208)
	大唐西域記	縛喝國	縛喝國	縛喝國東西八百餘里南北四百餘里北臨縛芻河	POINT (66.901385 36.766775)

Figure 11. GIS with Tang Map, route plotted, and query issued

Fig. 11 - Colour online, B/W in print

two books Southern History and Northern History. We first obtain the texts of the nine histories from ctext.org through its ctext API. After producing, within minutes, the full-text database with nine documents, one for each Standard History, we export the corpus to the *TermStat* tool together with a list of diseases.[23] The output from *TermStat* (nine lists of disease names with their frequencies) is then exported in csv form and uploaded to Palladio,[24] a visualization tool developed at Stanford University. The result is shown in Figure 12. An edge in Figure 12 connects a disease name and the history in which it appears. It is therefore very easy to see which diseases are shared by which histories. Observe that the five books of the Southern Dynasties (left hand side) and the four of the Northern Dynasties are clearly clustered separately.

With this *DocuSky-TermStat-Palladio* scenario one can conduct term analysis and visualize the outcome. Stanley-Baker's work on ancient Chinese medical texts is a particularly good example.[25]

Connecting Humanities Scholars and it Researchers Via Docuxml

An important goal of DocuSky is to enable digital humanities research without the direct help of an IT expert. On the other hand IT developers need to know what scholars are trying to do if they are to build suitable tools. We need a way to connect scholars and IT researchers so that the technical needs of the scholar, present and future, can be advertised and interested IT developers can heed the call. Tool building should not be limited to the DocuSky development team. Tools may be applied to different contents or the same content may make use of different tools. One of our principles is also to utilize existing tools as much as possible. Markus and Palladio are two such examples. They were developed at Leiden and Stanford Universities respectively, but they have been incorporated into the process flow of DocuSky and play important roles.

Thus we need to establish a standard for communicating between content and tools. In XML format we call this standard DocuXML.[26] DocuXML defines two classes of objects, *corpus* and *document*, where a corpus may contain several documents. DocuXML defines attributes of Corpus Metadata and tags for annotating documents. Users can also define their own tags. Any file in DocuXML format can be easily turned into a database through one click using *db_builder* in DocuSky. In addition to full-text search, metadata contexts and tagging contexts are realized through the built-in post-query classification features. In order to utilize an existing annotation tool that uses another format, a converter may be necessary to convert that format into DocuXML.[27] Currently we have written converters for plain text, Excel, Markus output in html, and csv. The same is true when a file in DocuXML format uses other tools for analysis and visualization. For example, in order to use Palladio to observe the co-occurrence relation of terms in several documents, we implemented the *TermStat* tool to generate the document/term tuples needed in Palladio and to convert the result into csv, the input format that Palladio accepts. Similarly, to utilize Web resources such

[23] Courtesy of Michael Stanley-Baker.

[24] Humanities + Design, Stanford University, "Palladio," Stanford University, 2013, https://hdlab.stanford.edu/palladio/.

[25] Michael Stanley-Baker, "Health and Philosophy in Pre- and Early Imperial China," *Health: A History*, edited by Peter Adamson (Oxford: Oxford University Press, 2019), 7–42.

[26] Research Center for Digital Humanities of NTU, "DocuXML 1.2 Scheme," http://docusky.org.tw/DocuSky/documentation/docs/DocuXml-1.2-Scheme.html.

[27] Implementing a converter is usually significantly easier than creating a tool.

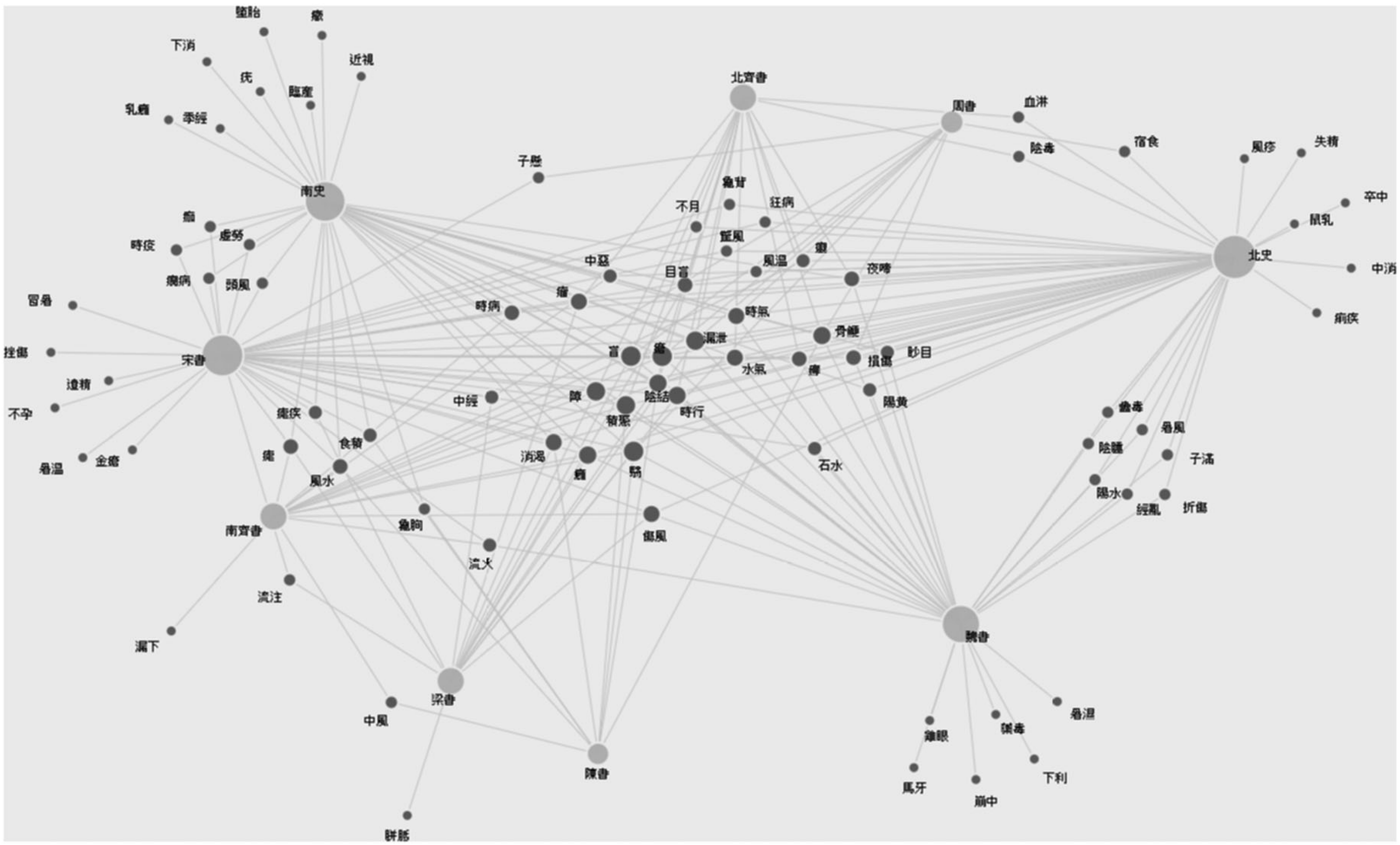


Figure 12. Graph shows relationships between diseases and Standard Histories

Fig. 12 - B/W online, B/W in print

as ctext.org or CBETA, we need to create an API for each of the resources so that their content can be downloaded and converted into DocuXML format. The successful implementation of CBETA API and the ensuing conversion of the downloaded files from TEI (the standard used by CBETA) to DocuXML, done by the DILA team led by Jenjou Hong, testifies to the simplicity and functionality of DocuXML.

If an IT researcher wishes to implement a tool that uses content through DocuSky, she needs to use the DocuSky API to access content. To simplify this work, we have developed a suite of widgets (DocuWidgets). Instead of understanding the communication details between DocuSky API and DocuSky, the implementer only needs to know how to use DocuWidgets to access data in a personal database generated by DocuSky. This way the implementer can spend more time on developing the tool, instead of understanding the inner workings of DocuSky.[28]

While DocuXML is still evolving, the experiment with using DocuXML as the link to turn a scholar's needs into working tools has been satisfactory so far. Except for the overall architectural design and the implementation of the core such as *db_builder* and *DocuWidgets* and some converters which are built by full-time staff, most of the tools, including the entire DocuGIS, were built by students and external IT researchers.

Final Remarks and Future Work

Figure 13 illustrates a simple research cycle of historical research. Given a research objective, the scholar collects relevant research data, processes them, organizes them, analyzes and sometimes visualizes the outcome, then interprets the findings. Interpretation sometimes leads to new research objectives, in which case the research cycle restarts. Eventually the cycle closes and hopefully results in publication of the findings.

In the age of digital archives and digital humanities, the data development process has emphasized building large-scale digital archive systems equipped with tools for analysis and visualization. Such systems contribute to the data collection, data process, data access (by building retrieval database systems), data analysis, and visualization parts of the research process (see Figure 14). Although some of these systems can be very sophisticated, they are usually built around specific data collections and cannot deal with the varying needs of individual scholars.

DocuSky advocates the separation of content and tools. In this sense, it is designed as a *platform*, not a system. That is, it allows content and tools both to be freely added by anyone. DocuSky maintains the standard format DocuXML through which content and tools can interact. It also provides the core database building facility which transforms any file in DocuXML form into a full-text searchable database. Another important aspect of DocuSky is that it does not build its own tools and contents unless necessary. In other words, it facilitates open contents over the Web and open source tools. Figure 15 shows how we envision the way DocuSky fits into the research cycle.

Note that we see a division of work between scholars and DocuSky. Scholars define their research objectives, collect research material that is unique to their work, and interpret what emerged from their analysis. DocuSky, on the other hand, provides tools for both such nitty-gritty work as downloading related material from the Web and processing and enriching the data, and for such technically demanding work as

[28] H.C. Tu, "A Platform for Constructing and Analyzing Personal Databases," *Journal of Digital Archives and Digital Humanities*, 2.1 (2018), 71–90, doi:10.6853/DADH.201810_2.0004.

Fig. 13 - Colour online, B/W in print

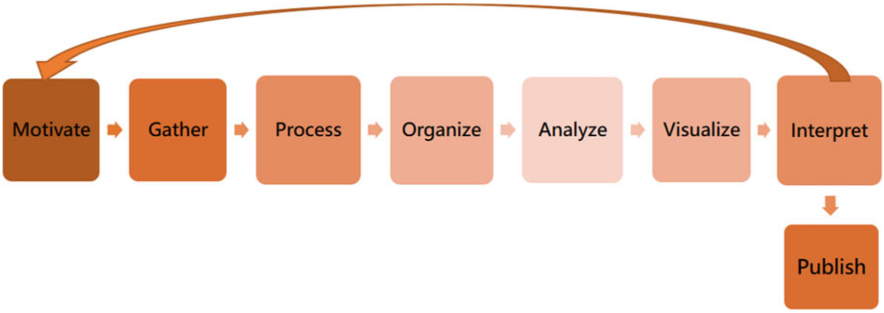


Figure 13. Naïve research cycle

Fig. 14 - Colour online, B/W in print

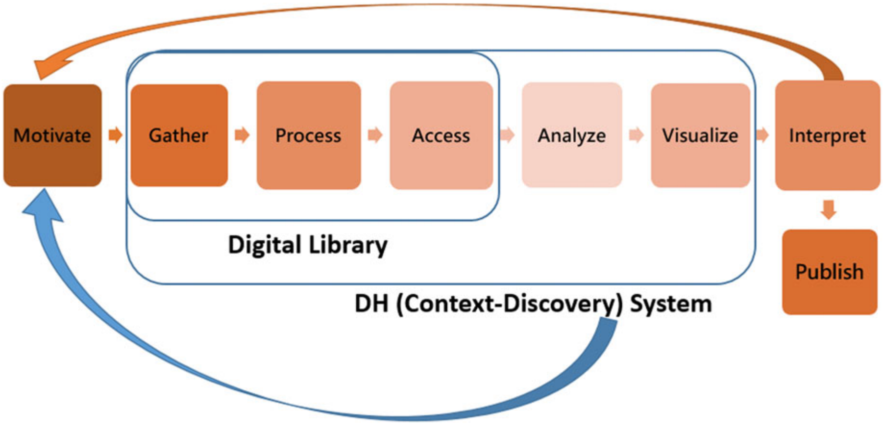


Figure 14. research cycle with DH archives

Fig. 15 - Colour online, B/W in print

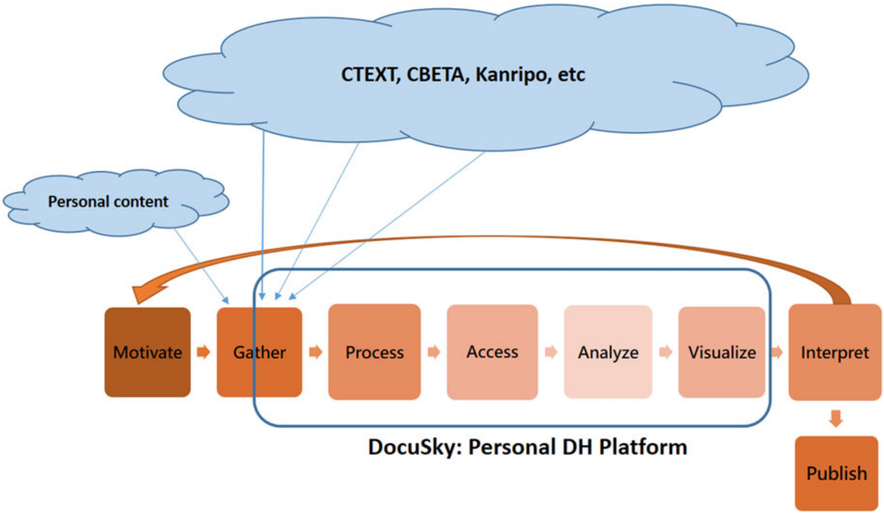


Figure 15. The research cycle with DocuSky

building a searchable database from processed data and creating analysis and visualization of the discoveries.

Since DocuSky is designed as an open platform, it is ever-evolving in scale and functionality. Although there are already more than 1,000 users of DocuSky, and it has been used in scores of projects, we welcome more. Moreover, the interface to enable a smooth exchange between scholar and IT developer—allowing the former to articulate her needs and the latter to respond to the request—has not been developed fully. Both human and technical aspects need to be carefully considered in order to tackle this problem properly. Better ways to receive and implement user feedback are also needed.

Although DocuSky currently works mainly on Chinese-language material, the principle of separating content and tools and the need to have a personal DH platform to empower scholars are certainly universal. We plan to put more language supports into DocuSky.

The contents that people have tried with DocuSky range from Shang oracle inscriptions to Taiwanese primary school textbooks. As new genres are tried and new requests for tools emerge constantly, we need more effective ways to engage IT researchers and to incorporate existing tools.

Acknowledgement and Credit. DocuSky was conceived by Hsieh-Chang Tu around 2015. It has been developed by the Research Center for Digital Humanities and the Laboratory of Automated Deduction and Digital Archives of the Department of Computer Science and Information Engineering of the National Taiwan University since then. This effort is led by Jieh Hsiang as Principle Investigator, Hsieh-Chang Tu as Principle Architect, with Yi-Mei Hung and Chijui Tu as Project Managers. But DocuSky is a genuine collaborative effort. Many students of the Laboratory, staff of RCDH, and external users have helped developing tools or testing their effectiveness. Many scholars have also used DocuSky and generously provided feedback for improvement. We are particularly grateful for the support from Dagmar Schaefer and Shih-Pei Chen of MPI Berlin, Hou-Ieong Ho of the Berlin State Library, Hilde De Weerdt of Leiden University, Peter Bol of Harvard University, Michael Stanley-Baker of Nanyang Technological University, and Jenjou Hung of the Dharma Dram Institute of Liberal Arts.

Cite this article: Tu H-C, Hsiang J, Hung I-M, Hu C (2020). DocuSky, A Personal Digital Humanities Platform for Scholars. *Journal of Chinese History* 1–17. https://doi.org/10.1017/jch.2020.28