

# On Building a Full-Text Digital Library of Historical Documents

Szu-Pei Chen<sup>1</sup>, Jieh Hsiang<sup>1†</sup>, Hsieh-Chang Tu<sup>1</sup>, Micha Wu<sup>2</sup>

<sup>1</sup> Department of Computer Science and Information Engineering

<sup>2</sup> Department of History

National Taiwan University,

Taipei, Taiwan

{gail,tu}@turing.csie.ntu.edu.tw, hsiang@csie.ntu.edu.tw, wumc@ntu.edu.tw

**Abstract.** The National Taiwan University Library has built a digital library of historical documents about Taiwan. The content is unique in that it covers about 80% of all primary Chinese historical materials about Taiwan before 1895, and that they are all available in searchable full text, in addition to metadata. To make these materials more accessible to the research community, we have developed, in addition to full-text search and retrieval, a concept of regarding the set of documents retrieved by a query as a sub-collection, and have designed post-query classification methods to help users find the inter-relationships among documents and the collective meaning of a sub-collection. We have also developed techniques for term extraction for old Chinese and a data format for representing governmental structures. We hope that our system will help advance research in Taiwanese history, and will set a model for other similar endeavor.

**Keywords:** historical documents, digital library, Taiwan, classification of query results.

## 1 Introduction

Starting from 2003, the National Taiwan University Library (NTUL) embarked on a major effort to systematically collect and digitize Taiwan related historical documents. The documents, numbered over 80,000, came from a wide range of sources, including imperial court archives, local and central judicial and administrative records, personal records of high-ranking officials, travel journals, diaries of influential people of the time, and land deeds. They were selected by historians, then typed, punctuated, proof-read, and supplemented with metadata records. Currently we have accumulated about 150 million Chinese words (characters), all in full text and searchable. They cover about 80% of all primary Chinese historical materials about Taiwan before 1895.

To our knowledge, there has never been such a collection, in both variety and magnitude, about Taiwan history available in searchable full text before. It should

---

<sup>†</sup> Corresponding author.

provide an exciting playground for anyone interested in pursuing research of Taiwanese history, scholars and laymen alike. To make these materials more widely available and easy to use, we have built a *Taiwan History Digital Library* (THDL). In addition to full-text and metadata search, we have also built referential tools to further facilitate their use. The tools that we have built so far include a Chinese-Gregorian calendar converter, corpus of names of people and locations, and charts of the evolution of local administrative structure with names of the officials and the duration of their terms.

While constructing THDL, we noticed that providing search/retrieval facilities alone is not sufficient for taking full advantage of its rich content. Most retrieval systems, when issued a query, ends at returning a list of relevant items. The question of how to make sense of the query results is usually left for the user to ponder. This way of representing results might be the best one can do if the returned items are independent objects and the meaning of each item is somewhat complete by itself. However, historical documents are often inter-related. For instance, a query may result in a list of reports from officials to the emperor and his responses, about a specific political incident. The documents may span over several years and represent several turns of events during its development. Thus, the retrieval results, if treated as a *collection*, may reveal meaning much more significant than the sum of its parts. Due to this observation, we have developed methods to *post-process* query results to show collection-level information, so that the intricate relationship among documents from the same query and their collective meaning can be investigated.

This paper is structured as follows. We give a (very) brief introduction of Taiwan's history, with emphasis on the political changes, in Section 2. Section 3 describes the content of THDL. Section 4 outlines the technical features and a new concept of treating the documents returned from a query as a sub-collection. A short discussion, including the differences between our work and other related work such as the Million Book Project, is given in Section 5.

Most translations from Chinese to English in this paper are done using *pinyin*. The few exceptions are when the Wade-Giles translation is more commonly used. For instance, we use Taichung (Wade-Giles) instead of Taizhong (pinyin) for 台中. All person names are presented as family name first, followed by given name. For instance, the family name of Zheng Chenggong is Zheng.

## 2 A Brief History of Taiwan

The first trace of human activities in Taiwan, according to archeological findings, dates back to at least 30,000 years. It is not clear if these pre-historic dwellers are the direct ancestors of the indigenous people of Taiwan, part of the Austronesian group, which form about 2% of the current population (numbered about 450,000, the rest are mainly descendants of Han Chinese).

The first mentioning of Taiwan in Chinese historical records was during the Three Kingdoms period (230 A.D.), although Han Chinese did not migrate to Taiwan in significant numbers until about 1,000 years ago. In 1624 the Dutch East India Company established a base at southern Taiwan and built the fort of Zeelandia at the

location of present day Tainan. At about the same time (1626), the Spanish also came to northern Taiwan and built three forts. They were driven out by the Dutch in 1642. The Dutch were then driven out by the Ming general Zheng Chenggong (鄭成功, known in the west as Koxinga) in 1662. Koxinga and his descendants established the first Han Chinese government in Taiwan and used it as a base for their attempt to recover China mainland (then occupied by the Manchurian-lead Qing Dynasty, established in 1644 by overthrowing the Han-Chinese Ming Dynasty) for the Ming Dynasty. In 1683 Zhang Keshuang (鄭克塽), a grandson of Koxinga, surrendered to Qing, which established prefectural level governments in Taiwan and upgraded it to the Province of Taiwan in 1885. After loosing the Sino-Japanese, Qing ceded Taiwan to Japan in 1895. After being defeated in the Second World War, Japan returned Taiwan to China in 1945. The Chinese Nationalist government, after loosing the mainland to the communists, moved their seat to Taiwan in 1949. The island moved toward full democracy when the Marshall Law was lifted in 1987. (The State of War between the two governments did not officially end until 1991.) In the year 2000, Mr. Chen Shui-Bien of the Democratic Progressive Party was elected President, thus ending the monopoly of the Nationalist Party to the Presidency. Those who wish to learn more may consult the “History of Taiwan” entry of Wikipedia [1].

### 3 The Content of THDL

The content in THDL can be roughly divided into three categories: imperial court documents of Ming and Qing Dynasties, documents of local governments - in particular the Danxin Archives, and local land deeds. Together they yield a rather extensive picture of the political, sociological, and economic landscape of pre-1895 Taiwan, from the perspective of the central government to everyday people.

Our content has at least two other distinctive features. One is that most of our content are primary documents, and they cover at least 80% of all such Chinese materials. Other than diaries of important officials and travel journals, we did not include any “secondary” material such as memoir, biography, or scholarly research work. The reason is that we wish to present historical documents in their original form, with as little later interpretation as possible. (We remark that we are building a database of research work on Taiwanese history. This database, however, will serve a purpose different from that of THDL.)

The second distinctive feature is that all of the documents in the digital library are keyed-in as full text, with punctuation added, in addition to metadata. The availability of full text and full-text search sets THDL apart from any other database of Chinese historical documents that we know of. It also makes it an exciting research environment for finding associations among documents and among collections of documents that cannot be done with metadata alone. From the technological side, it serves as a good source for experimenting in text mining and other information techniques. Indeed, we have already developed tools for extracting terms, names, and dates [2].

In the following we describe the three categories of contents in THDL.

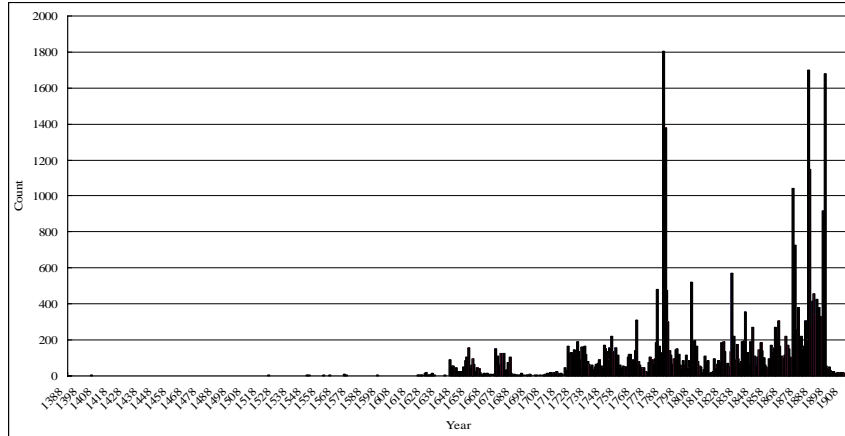
### 3.1 Selected Collections of Taiwan-related Documents from the Imperial Governments of the Ming and Qing Dynasties

During the Ming and Qing dynasties (1368-1644 and 1644-1911), China extended its border to Taiwan and neighboring areas. Hence, the imperial governments of China produced a significant number of court documents involving Taiwan during the two dynasties, especially Qing. These documents are hidden in various imperial government archives that are kept in different institutions, some in China and some in Taiwan. Furthermore, most of them are never published and, if available for viewing, are only as microfilms. It is obvious that they are of fundamental importance for anyone who wishes to study Taiwan history during the Ming and Qing dynasties, but it is almost impossible to access them in their entirety. Thus, the Council for Cultural Affairs (CCA) of Taiwan commissioned the National Taiwan University Library (NTUL) in 2003 to make a comprehensive digitization of imperial court documents related to Taiwan. In this ambitious project, NTUL first collected copies of the imperial government archives of Ming and Qing dynasties from different libraries and archives, then collaborated with a team of historians (first lead by Professor Li Wen-Liang, then by Professor Wu Micha, both of National Taiwan University) to carefully select the documents within those archives that are related to Taiwan. In addition to creating metadata for those selected documents, the full text contents were also keyed-in. The full texts were then proof-read with punctuation added. In the two years that followed, more than 40,000 such documents were selected and over 35,000,000 characters of punctuated full text were produced [3].

Although the funding from CCA stopped after two years, NTUL continued looking for new sources of historical documents and has added at least another 4,000,000 characters since then.

The contents of this collection are selected from the Ming Reign Chronicles and the Qing Reign Chronicles (明實錄,清實錄), Palace Memorials (奏摺), Archives of the Grand Council (the Administration of Military Affairs) (軍機處檔案), the Imperial Decrees Archives (上諭檔), the Grand Secretariat Archives (內閣大庫), Archives of the Diary-Keepers (月摺檔, 起居注), Archives of the Imperial Palace (宮中檔), Diplomatic Documents (照會), officially edited local gazetteers (地方志), Court Edicts concerning Revolts (剿捕廷寄檔), and others. Altogether there are at least 235 different archives and collections that we have examined. The earliest document in this collection was written in 1388 and selected from the Ming Reign Chronicles, and the latest was written in 1911, just before the Qing dynasty ended.

This collection represents the history of Taiwan from the perspective of the Chinese imperial government. As an outlying island of a vast empire, Taiwan did not get mentioned often in the imperial court unless something bad, such as a rebellion or famine, had occurred. Fig. 1 gives a breakdown of documents according to the years that they were written. Indeed, each peak in the chart corresponds to an event that was crucial to Imperial China. For instance, the peaks of 1786 to 1788 reflects the most serious revolt ever occurred in Taiwan (the rebellion of Lin Shuangwen 林爽文). The peak of 1884 corresponds to the Sino-Franco War (1883-1885), during which the French invaded, unsuccessfully, northern Taiwan. The peak of 1895 indicates the Sino-Japanese War (1894-1895), which ended with Taiwan's secession to Japan.



**Fig. 1.** Yearly count of documents in the Selected Collections of Taiwan-related Documents from the Imperial Governments of the Ming and Qing Dynasties.

### 3.2 The Danxin Archives: Official Documents of the Danshui Sub-prefecture and the Xinzhu County (1789-1895)

The Danxin Archives is a collection of administrative and judicial records of the Danshui sub-prefecture and the Xinzhu County. The area involved covers the entire northern Taiwan, with the present day Miaoli at the southern end. The dates spanned from 1789 (the 54th year of the Qianlong's 乾隆 reign) to 1895 (the 21st year of the Guangxu's 光緒 reign). There are 19,281 documents, grouped into 1,143 cases. More than half of the cases were produced during Guangxu's reign, from 1875 to 1895 [4].

The Danxin Archives was organized and classified by the legal scholar Dai Yanhui (戴炎輝) into three categories: administrative, civil, and criminal [5]. Documents in each case are further arranged into a series in chronological order. The Danxin Archives is one of the only three pre-1911 Chinese local government archives known to exist. Different from official local gazetteers, the Danxin Archives provides a first-hand detailed account of the social life of citizens in the Qing dynasty. It is invaluable for anyone who wishes to study the political, economical, judicial, or administrative development of late 19th century Taiwan and China.

The original Danxin Archives is in the care of the National Taiwan University Library. There have been quite a few research articles and books based on Danxin Archives (see, for instance, [6]), mainly via studying a microfilm version produced by the University of Washington. In order to make this important material more available to the research community, NTUL embarked on a project to publish the full text, on the average of releasing 4 volumes a year. So far we have published 20 volumes, and the total is estimated to be about 36. We have also scanned all images (27,017), which will be incorporated into THDL. Currently we have proof-read and punctuated 11,611 full-text documents, all of which are available in THDL. There are also 11,242 associated metadata records.

### 3.3 Collections of Land Deeds of Taiwan

Until the beginning of the 20th century, land deeds were the only proof of ownership and transition of land in Taiwan. They were hand written and were usually prepared by a scrivener. In the early phase of the Japanese occupation of Taiwan, the Japanese government brought in western style land measurement mechanism and established a modern system to manage land and ownership, thus replaced the ad hoc land deed system. During the transformation of the land management system, the Japanese Governor-General made a concerned effort to record land deeds so that they can be ported to the new system. Thus, about 15,000 land deeds are included in the Archives of the Japanese Taiwan Governor-Generals (臺灣總督府檔案). It was estimated [7] that there are an additional 20,000 land deeds in the hands of libraries, museums, private collectors, and individual families which were not accounted for in the aforementioned archives. In 2003 and 2004, CCA commissioned the National Taichung Library (NTL) to collect and digitize (in full text) the hand-written copies of land deeds from the Archives of the Japanese Taiwan Governor-Generals. In this project, NTL keyed-in the full text of 15,901 land deeds from the Archives of the Japanese Taiwan Governor-Generals, 1,674 from published literature, and 157 from private collections. In addition, NTUL has digitized its own collections of land deeds which totaled at 3,667. Together, NTL and NTUL have collected more than 20,000 land deeds in Taiwan, all incorporated into THDL. We are adding another 3,000, which should be done in a few months.

While each land deed may have significance only to its owner, the collection as a whole provides a fascinating glimpse into the pre-1895 Taiwanese grassroots society. For example, the fluctuation of value of land reveals a great deal about the economic development of each region. Since many of the deeds were between indigenous people and Han immigrants, they also provide clues to the intricate relationship among the various peoples of Taiwan [8], the evolution of rights to land, and the gradual assimilation of the indigenous people (in particular the Pinpu 平埔族群) into the Han society.

Land deeds usually have a fixed format, many of which differ only in the names of the parties involved, the names of the witnesses and scrivener, location and boundary of the land, and the date. They are thus ideal for experiments in term extraction and text mining. Indeed, we have already mined over 90,000 names of people and locations. Just people's names alone should be a valuable source for research in history.

## 4 Post-query Analysis and Referential Tools

The availability of full text opens an exciting new door for using primary important historical documents in research and teaching. While full-text search is a must-have, we try to explore other ways with the aim of building a research environment around these resources for historians and researchers of other disciplines.

#### 4.1 Query Returns as a Sub-Collection

Most document retrieval systems regard query results as a set of more or less independent documents. It is the user's job to go through the returns to see if any is relevant to her request. The facility of full-text retrieval sometimes makes the resulting set too large to manage. Thus some systems also provide query refinement mechanisms to further restrict the query results.

Historical documents have the feature that they are often *inter-related*. For example, while a single land deed may yield little meaning by itself, a series of land deeds of the same piece of property may reveal significance far greater than the sum of its parts. Indeed, historians rarely look at a single document alone (unless they are looking for something to refute a conjecture or to add support to an observation). They gather documents from different sources, try to figure out their *collective meaning* (relationships among the documents, and the meaning of the collection as a whole), and draw conclusions accordingly. Since THDL has already accumulated an unprecedented amount of full-text documents on Taiwan history under one roof, our next step is to represent the results from a query as a collection by itself and try to show the various possible relationships among the query result documents. This is done mainly through *post-processing* a query's returns as a sub-collection.

THDL features two facilities to help user find collective meanings of query returns, *multi-dimensional post-query classification* and *term frequency analysis*.

**Post-Query Classification.** The post-query classification mechanism classifies documents of the resulting set of a query according to several predefined dimensions, which are metadata fields describing important background knowledge of documents such as dates, authors, and sources. After the resulting set of a query is returned (we call it a *sub-collection*), THDL classifies its contents according to year, author, and source on the left of the web page, and presents summaries of the documents themselves on the right. Fig. 2 shows the returns of the query *Lin Wencha* (林文察), a well-known Taiwanese native general during the Qing dynasty, with the timeline on the left. Among the 375 returned documents (all *zouje* 奏摺 - reports from officials to the emperor), 331 appeared between 1861 and 1864, the year when Lin was killed in battle. What is interesting is that 31 additional *zouje* mentioning Lin appeared after his death, until as late as 1906. An examination of these documents reveals a series of family tragedies involving his brother's (Lin Wenmin, also a Qing general) being wrongly accused of crimes and executed, his son's (Lin Chaodong) death in battle, and so forth. Thus a family's story unfolds in the sub-collection of a single query.

Note that the predefined dimensions are dependent on the characteristics and metadata of each collection, so different collections may need different dimensions. Our system also provides reordering facility, so that the user can examine the returned list of documents in any of the predefined dimensions. By presenting classifications of multiple dimensions, the user can switch from dimension to dimension and observe the distribution and behavior of each dimension. We also allow users to bookmark documents and create their own sub-collections.

**THDL 台灣歷史數位圖書館**

文件搜尋:  [Default Query]

最近幾次查詢: 林文察 375

林文察 1554  
臺灣 1160  
ntu-0105539-000060... 1

最近幾次查詢: 年代 (TM) 375  
咸豐 59

咸豐八年 (1858) 6  
咸豐九年 (1859) 2  
咸豐十年 (1860) 5  
咸豐十一年 (1861) 46

同治 300  
同治 (1862) 11  
同治初 (1862) 5  
同治一年 (1862) 72  
同治二年 (1863) 70  
同治三年 (1864) 127

光緒 16  
光緒五年 (1879) 1  
光緒八年 (1882) 2  
光緒十年 (1885) 2  
光緒十一年 (1885) 1  
光緒十四年 (1888) 2  
光緒十五年 (1889) 1  
光緒十六年 (1890) 5  
光緒十八年 (1892) 1  
光緒三十二年 (1906) 1

找到筆數 375 頁次 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >>

1. 為閩省兵勇克復江山縣城現又協同收復常山謹將錄報疊次接仗情形恭摺由驛馳奏仰祈聖鑒事

閩浙總督奴才慶端跪奏，為閩省兵勇克復江山縣城，現又協同收復常山，謹將錄報疊次接仗情形，恭摺由驛馳奏，仰祈聖鑒事。竊浙逆運陷常山、江山各城，業經奴才於奏報汀州軍務摺內先後陳明在案。蓋常山緊接江西之玉山，而江山則與浦城毗連。該逆明、升任浙江總督便署金龍鳳運往後文、督衢州鎮總兵崇志會派參將林文察統帶臺灣兵勇馳往江山攻剿，再經浙江撫臣王有齡、前廣西提督臣張...

清政府鎮壓太平天國檔案史料(v.23) • 閩浙總督慶端 • 咸豐十一年 • imh-1654743-0003400036-000026.txt

2. 王有齡奏報兵力不敷添募壯勇並請准江西欠餉片

再，浙省賊勢有增無減，而兵力有減無增，前准江蘇撫臣薛煥咨函，逆匪逼近上海，懇圖於運，隨路帶兵大員向望雙目失明，不能辦事，僅令總兵馬得昭迅速前往。其時蘇興之賊竄近王店，預伺海甯，正在十分危急之際，臣等知上海為江北餉源所繫，不敢視為緩圖，撥撥不暇水士者一千名，交副將惠壽帶同歸伍。其總兵曾玉明、參將林文察、都司劉紹基所部臺灣兵勇三千名，因武平、汀州先後失守，經督臣...

清政府鎮壓太平天國檔案史料(v.23) • 浙江巡撫王有齡 • 咸豐十一年 • imh-1654743-0004100042-000003.txt

3. 議政王、軍機大臣字寄閩浙總督慶、督辦軍務、浙江巡撫左。同治元年正月十四日奉上諭：前據曾國藩奏...

議政王、軍機大臣字寄閩浙總督慶、督辦軍務、浙江巡撫左。同治元年正月十四日奉上諭：前據曾國藩奏，浙省賊勢驟張，恐由遂安、龍泉一路取道浦城竄至江西之鉛山、福建之光澤，當經諭令慶端運兵浦城，杜賊旁竄鉛山而馳左宗棠後路，並因上海危急，嚴抗，但能攻拔一城，則全局自振，所籌尚合機宜。著即飭李元度、林文察、李如先等督統兵勇實力進剿。左宗棠應已由廣信進駐衢州，慶端由...

清政府鎮壓太平天國檔案史料(v.24) • • 同治一年 • imh-1654750-0004200043-0000032.txt

4. 議政王、軍機大臣字寄閩浙總督慶。同治元年二月二十七日奉上諭：前因浙省賊匪有侵犯江西等處之勢...

議政王、軍機大臣字寄閩浙總督慶。同治元年二月二十七日奉上諭：前因浙省賊匪有侵犯江西等處之勢，諭令慶端督軍進軍浦城，以杜逆賊旁竄廣信、鉛山，繞出左宗棠後路。茲據慶端奏，龍巖等處賊匪分路竄擾衢州，及閩省邊防同時喫緊，分別布置嚴防。駐衢州。李定太一軍，前據該督奏稱該令撤隊分堵龍巖、壽昌，並令林文察一軍進攻壽昌，此次賊即由龍巖等處上竄，該總兵何以不能堵禦。其...

清政府鎮壓太平天國檔案史料(v.24) • • 同治一年 • imh-1654750-0016100162-0000116.txt

5. 議政王、軍機大臣字寄閩浙總督慶、浙江巡撫左。同治元年三月初六日奉上諭：慶端奏，大股逆匪分竄...

議政王、軍機大臣字寄閩浙總督慶、浙江巡撫左。同治元年三月初六日奉上諭：慶端奏，大股逆匪分竄浙省，著即飭李元度、林文察、李如先等督統兵勇實力進剿。左宗棠應已由廣信進駐衢州，慶端由...

Fig. 2. Query results of “Lin Wencha”

**Term Frequency Analysis.** We have also built tools in THDL for term extraction, and have used the tools to build a corpus of over 90,000 names of people and locations. The names are used to provide term frequency analysis which further helps the user to explore and expand the sub-collections. Term frequency analysis tabulates the numbers of times terms appear in the sub-collection and presents them to the user (see Fig. 3). The user can use them to decide how relevant a person or a location is to the present query (and the associated sub-collection) and explore further.

Through post-query classification and term analysis, we hope to provide users better ways to analyze the sub-collection retrieved from a query as a whole, rather than as individual documents.

#### 4.2 Referential Tools

A historian always has referential resources within reach when she conducts research. To make THDL more useful, we have also built several referential tools and are in the process of building more. We describe some of them here.

**Chinese-Gregorian Calendar Converter.** Date is obviously among the most important information for history. The traditional Chinese calendar is lunar and does not correspond directly to the Gregorian calendar. During the dynastic era, the years were indicated by the reign title of an emperor, who may change the title from time to

time. To complicate things further, Koxinga continued to use Ming reign even if it no longer existed (after being terminated by Qing). To maintain consistency, we use Gregorian calendar as the standard metadata format for dates and wrote a converter that made a daily correspondence between Gregorian and the Chinese calendars starting from the first day of Ming (1368/01/25) to the last day of Qing (1912/02/17). We also included part of the Japanese calendar for documents produced during the period of Japanese rule (1895-1945). The total number of dates in our database is about 250,000.

We also wrote a program to automatically recognize dates appeared in documents and convert them to Gregorian. A query interface for users to directly access the conversion table is available.

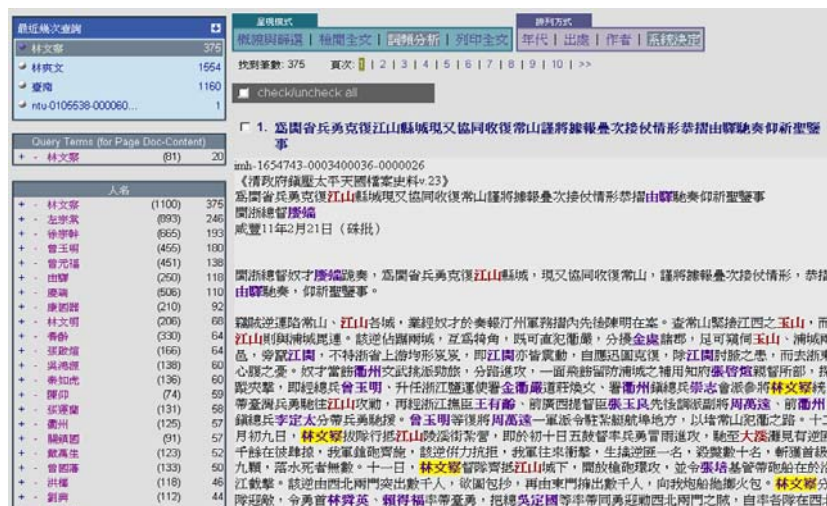


Fig. 3. Term frequency analysis of the sub-collection from the query “Lin Wencha”

**Corpus of Names of People and Locations.** *Who* and *where* are no less important to historical events than *when*. Thus we need ways to extract names of people and locations from our documents. While the names mentioned in the court documents are relatively easy to collect, since they usually only involve important officials and locations, those that appeared in local documents and land deeds pose a problem.

The challenge is further compounded by the fact that the Chinese language has no spacing between characters and the old scripts have no punctuation. Although there have been research of term segmentation for Chinese [11, 12], they are not directly applicable because the grammatical structure of old Chinese is quite different from that of the modern Chinese. Term frequency count won't do the trick either since many individuals only appear once (such as those appearing in only one land deed). The irregularity of names of the indigenous people, as opposed to the simple family-given name format of Han Chinese, adds more challenge.

To solve this problem we developed a *word-clip* algorithm to recognize proper nouns in Chinese documents [2]. The idea is to recognize existing relations between

proper nouns and their context in a collection in which the documents have similar styles of writing. In such a collection, a specific type of proper nouns (such as locations) is often surrounded by similar leading phrases and ending phrases. We call such a pair of leading and ending phrases a *clip*. The characters within a clip usually form a proper noun.

Our algorithm starts with a set of known names to find useful clips. The clips are then used to catch more names. This process is iterated until it reaches saturation (very few new clips are generated). We have applied the word-clip method to two of the collections (except the Danxin Archives). Experimental results show that the average precision rate of identifying person names is about 50% and the estimated recall rate is about 75%. For location names the precision rate is 82% and the estimated recall rate is 84%. So far we have collected 90,948 person names and 3,496 location names in the two collections.

The corpus is used in the term frequency analysis utility mentioned in the previous section.

**The Local Officials Chart of Taiwan during the Qing Dynasty.** Knowing who was in charge of what at what time is very useful when studying history. We took an important source book on the local officials of the Qing Dynasty, Listing of Administrative Offices and Officials [9], and built a tool to help users find information about officials. The book lists all the administrative offices of Taiwan during the Qing dynasty, together with the names of the officials who occupied the offices, with their starting and ending dates.

Instead of just making a table, we designed a data format that allows one to fully utilize the information provided by the book. The design principle of the data format is to make each tuple as small as possible so that local changes can be made easily. This is necessary because not all records in the book are complete or accurate. However, the tuples have to contain enough fields so that they can be connected to provide answers to new types of queries. We have proved that the data format we designed is both minimal and sufficient [10].

Our system provides three types of queries at the moment: *query by office*, *query by person*, and *query by year*. A *query by office* gives the list of names of people who ever held that office, in chronological order. This is also the way data were organized in the book. The other two types of queries, however, show how digitized data can provide a lot more than the source book was originally intended for. A *query by name* returns a chronology of all the offices of Taiwan that the named person ever held. A *query by year* returns a chart that shows the entire governmental structure of Taiwan of that year with the names of the officials for each office during that year (see Fig. 4). A click on the name of an office shows another chart that represents the internal structure of that office, also with names of officials holding positions within that office during that year. A click on any person appeared in the tree sends a *query by person* command to the system and generates the respective chronological chart accordingly.

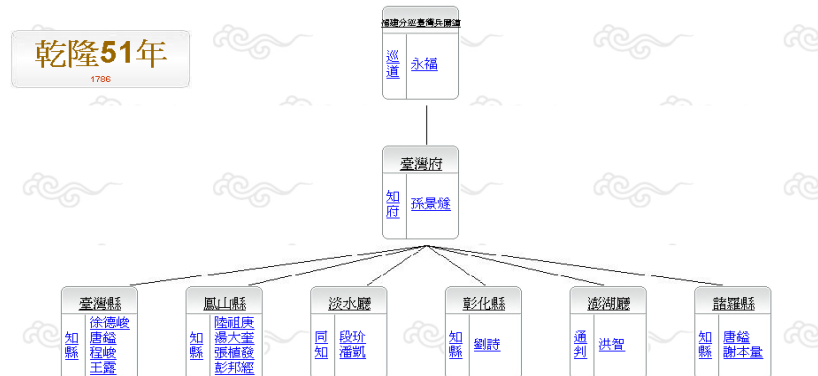


Fig. 4. Taiwan government structure during the year 1786.

## 5 Discussion

This paper describes THDL, the Taiwan History Digital Library of historical documents. In addition to incorporating the full texts of over 80% of primary Chinese documents about Taiwan before 1895, we have also developed different ways to facilitate THDL for research in history. They include full-text search, techniques and interfaces for classifying and exploring a query result as a sub-collection, term frequency analysis, and referential tools.

There are many ambitious digitization projects currently in progress, such as Google Books (see <http://books.google.com/googlebooks/about.html>) and the Million Book Project [13]. THDL is different from these projects in many ways. First, THDL is about a single domain – Taiwan history, and thus can utilize domain-specific knowledge and metadata to provide important features such as post-query classification. Second, we only select the contents in the archives that are related to Taiwan. In other words, we do not cover entire archives or books, only the parts that are relevant to our purpose. Third, we did not (could not) use OCR to obtain full texts because all the documents are hand-written, which renders OCR impossible. Thus all our texts were keyed-in. Lastly, there is no copyright concern since we only deal with ancient materials. Like the other two projects, we emphasize on the ease of use. In addition to providing full-text search, we want our interface to be at least as friendly as Google's, with additional emphasis on providing help *after* the query results are returned.

Another project that we should mention is Taiwan's National Digital Archives Program (NDAP) [14]. Although NDAP also deals with historical subjects about Taiwan, it does not emphasize on full texts. Thus we did not request/receive funding from NDAP for constructing THDL, except for building the full text of the Danxin Archives.

We are still building more contents, both primary materials and research work, about Taiwan history. We are also developing text mining techniques to provide

better analytical tools of historical documents. We hope THDL can help push research on Taiwan history to a new horizon.

**Acknowledgments.** Fundings for THDL were provided by the National Taiwan University and the Council for Cultural Affairs for building the full text of the Imperial Court documents, National Science Council under grant NSC94-2422-H-002-001 for building the full text of the Danxin Archives, and NSC95-2221-E-002-277 for building the THDL system. Their generous supports are greatly appreciated.

Many people helped in creating the content of THDL. We thank Chiu Wan-Jung and her staff of the Special Collections Department of the NTUL, the National Taichung Library, Professor Lee Wen-Liang and his assistants of the Department of History of NTU. We also thank Lin Hsin-Yi who provided many comments, and the graduate students of the Laboratory of Digital Archives of the Department of CSIE of the National Taiwan University for building some of the tools.

## References

1. <http://en.wikipedia.org/wiki/Taiwan/History>
2. Chang, S.P.: A Word-Clip Algorithm for Named Entity Recognition: by Example of Historical Documents. Master Thesis, National Taiwan University, Taiwan (2006) [in Chinese]
3. Chiu, W.J.: The Digital Project of Taiwan-Related Archives in Ming and Qing Dynasty. The Library Yearbook of ROC 2006. National Central Library, Taiwan (2006) 128-129 [in Chinese]
4. NTU Library, [http://140.112.113.4/project/database1/database1\\_1.htm](http://140.112.113.4/project/database1/database1_1.htm) [in Chinese]
5. Dai, Y.H.: Preliminary Remarks on Putting in Order the Qing Danxin Archives (清代淡新檔案整理序說). Taipei Cultural Relics (台北文物), Vol. 2, No. 3 (1953) [in Chinese]
6. Allee, M.: Law and Local Society in Late Imperial China: Northern Taiwan in the Nineteenth Century. Stanford University Press (1994)
7. Wu, M.C., Ang K.I., Lee, W.L., Lin, H.Y.: A Brief Introduction to the Integrated Collections of Taiwan-related Historical Records. CCA and Yuan-Liou Publishing, Taiwan (2005) [in Chinese]
8. Hong, L.W.: A Study of Aboriginal Contractual Behavior and the Relationship between Aborigines and Han Immigrants in West-Central Taiwan, Vol. 1. Taichung County Cultural Center, Taiwan (2002) 5 [in Chinese]
9. Pan, C.W. (ed.): Taiwan Geography and History, Vol. 9, No. 1 (臺灣地理及歷史卷九官師志第一冊文職表). Taiwan Provincial Literature Committee, Taiwan (1980) [in Chinese]
10. Chang, J.T.: Model and Implementation for Representing Governmental Structures and Officials. Master Thesis, National Taiwan University, Taiwan (2007) [in Chinese]
11. Chien, L.F.: PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval. Proceedings of 1997 ACM SIGIR Conference (SIGIR'97), Philadelphia, USA (1997) 50-58
12. Chen, H.H., Lee, J.C.: Identification and Classification of Proper Nouns in Chinese Texts. Proceedings of 16th International Conference on Computational Linguistics, Copenhagen, Denmark (1996) 222-229
13. Reddy, R., StClair, G.: The Million Book Digital Library Project. <http://www.rr.cs.cmu.edu/mbdl.htm> (2001)
14. National Digital Archives Program. [http://www.ndap.org.tw/index\\_en.php](http://www.ndap.org.tw/index_en.php) (2007)