Discovering land transaction relations from land deeds of Taiwan

Shih-Pei Chen, Yu-Ming Huang, Jieh Hsiang, Hsieh-Chang Tu, Hou-Ieong Ho and Ping-Yen Chen
National Taiwan University, Taiwan

Abstract

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. We have built a full-text digital library that contains almost 40,000 land deeds. The deeds in our collection range over 250 years and are collected from over 100 sources. The unprecedented volume and diversity of the sources provide an exciting source of primary documents for historians. But they also pose an interesting challenge: how to tell if two land deeds are related. In this article, we describe an approach to discover two important relations: successive transactions and allotment agreements involving the same property. Our method enabled us to construct 6,035 such transaction pairs. We also introduce a notion of 'land transitivity graph' to capture the transitivity embedded in these transactions. We discovered 2,436 such graphs, the largest of which includes 104 deeds. Some of these graphs involve land behavior that had never been studied before.

Correspondence:
Jieh Hsiang, Research Center for Digital Humanities, National Taiwan University, Taiwan.
Email:
jieh.hsiang@gmail.com

1 Introduction

Land deeds are among the earliest written social agreements known to men. Among the oldest surviving land deeds is a Hittite tablet that dates back at least 3,000 years (Goetze, 1939). Early land deeds between Native Americans and the settlers provided a glimpse into pre-colonial tribal boundaries as well as early Anglo-Indian relations (Baker, 1989). Individual land deeds, such as the Batman Land Deed of Australia, also played important roles in colonial history (Billot, 1979). Land deeds in the USA, where they have been well maintained, have been used successfully as a primary and crucial source for genealogical research (Hone, 2008).

Land deed research had played a unique role in modern Chinese history. Although there had been an important social contract for at least two millennia—a collection of about 500 dating as far back as 300 AD was discovered in the caves of Dunhuang (Ikeda, 1986)—land deeds were used as a research tool in the 19th century by the foreign powers as a vessel for understanding how Chinese society worked. The first notable example is the work of Hoang (1920), who used the deeds acquired by missionaries through the purchases of land to study Chinese land trading behavior. The British did an extensive survey in order to understand land ownership issues after they leased the New Territory in 1896 (Chun, 1986). A much larger and more elaborate effort related to land deeds was conducted by the Japanese colonial government when they took over Taiwan from the Qing Dynasty in 1895. In order to understand the Chinese traditional laws so as to ensure a smooth transition of power, the Japanese officials spent over a decade collecting and studying land deeds, and eventually published several multi-volume books including the *Supplements to the Investigation of the Grand Leases* (大租取調書附屬參考書) in 1905 (TB, 1963)

Literary and Linguistic Computing, Vol. 28, No. 2, 2013.

doi:10.1093/llc/fqs063 Advance Access published on 22 December 2012

and the seven-volume *Taiwan Private Law* (台灣私法) (TB, 1960–63). In addition to containing over 2,000 deeds (all of which are included in THDL, the Taiwan History Digital Library, the system that we shall describe in this article), the books also attempted to interpret baffling phenomena such as multi-ownership of land or *zhaoxi* (找洗, the custom of requesting additional money a few years after the land was sold), which were common in Taiwan and southern China at the time.

After the 2nd World War, research using land deeds shifted to understanding the traditional Chinese social structure and social movements. Shiga (1967) studied Chinese law using social contracts. Terada (2005, 2006) and Kishimoto (1997) explored various social orders revealed through deeds. Post-war China used deeds as a springboard to investigate the budding of capitalism in modern China (Fu, 1961). Another important research direction is regional studies, for which the local nature of land deeds made them ideal as primary material. Representative works include those on Huizhou (Wang, 2002), Minnan (Yang, 1988; Chen, 2004), and Hong Kong (Chun, 1986).

The most prolific and diverse research activities, however, are conducted in Taiwan. Never having its own 'central government' until after 1949, Taiwan's local social contracts, and land deeds in particular, form a crucial part of its development records and cultural heritage.

Until the turn of the 20th century, hand-written land deeds were the only proof of land activities in Taiwan. A deed may involve a transaction such as selling/buying, lending of land to smaller farmers, dividing the land among children or shareholders, and cultivation permits. The deeds were usually drawn up following, depending on their nature, a typical but not standard format in an *ad hoc* manner. Indeed, even the name of the location may be written in a local convention unfamiliar to the outsiders.

While each land deed may have significance only to its owner, a large collection of them provides a fascinating glimpse into the pre-modern Taiwanese grassroots society. A recent survey paper (Li, 2010) pointed out that land deeds research toppled the conventional perception of Qing Taiwan as a Han-centered society (Ka, 2001). Such research also catalyzed the emergence of a regional social theory (Shih, 1995) that challenged the interpretation of social contracts based on modern law as done in *Taiwan Private Laws*. Other research topics include regional histories, the transition of land rights, Han-indigenous relations, family histories, commerce, law, and other social issues (Chen, 1997; Ka, 2001; Shih, 2001; Hong, 2005).

One challenge facing land deed research is that the original material is hard to come by. Many of the old deeds were either discarded or sold to individual collectors or museums. In addition to the ones transcribed in books such as *Taiwan Private Law*, research relies heavily on a small number of sizable (usually several hundred deeds) collections that are lucky enough to be kept intact. Indeed, whenever a new collection is discovered, it always causes excitement because past experience shows that new research discoveries can usually be made. It was estimated (Li, 2004) that there are only 35,000 land deeds in existence.

In the past few years, we have built a full-text digital library of primary historical documents of Taiwan called THDL. Among its corpuses is a collection of almost 40,000 land deeds, spanning from 1666 to the first decade of the 20th century, and collected from over 100 sources of origin (Hsiang *et al.*, 2009). This collection is unprecedented in terms of volume, time span, geographical distribution, and variety. (However, our effort also showed that Li's estimation of 35,000 surviving deeds is inaccurate.) While THDL presents an exciting source of primary materials for historians, it also poses a challenge: how to find the relationship between two land deeds or how to find all the land deeds involving the same piece of land. Although it was customary to hand down earlier deeds to the new owner during the transaction of property, most of these links were broken when the Japanese, during their colonial rule of Taiwan between 1895 and 1945, modernized the land management system (Li, 2004). That is because the officials only recorded the last deed as the proof of ownership but ignored the previous ones. Consequently, many of the older deeds were either destroyed or (later) sold as collector's items because they had lost their original significance.

In this article, we present a semi-automated method to discover the transaction relations among land deeds. Our method makes use of many 'features' that are implicitly embedded in the full text of a land deed. These features are often named entities such as the transaction date, the names and roles of the people involved in the transaction, the general location of the land, and some others that we will describe in more detail later. Due to the lack of corpus training data, we choose not to use general-purpose entity recognition methods (Sun *et al.*, 2003; Fu and Luke, 2005; Nadeau and Sekine, 2007). Instead, we use carefully designed regular expressions to extract required features from high-quality metadata and full text (Bradley, 2007; Nguyen and Shimazu, 2007). Although land deeds do not follow rigorous standard formats, their similarity in nature made it possible to achieve satisfactory results with regular expressions alone. We shall focus on two important relations: 'successive transaction pairs' and 'allotment agreements'. Two less important relations, 'red deeds' and 'duplications of deeds', will also be presented. We further connect the transitive activities on the same property into a concept called 'land transitivity graph', which captures the history of the land over time. The largest such graph that we found has led to a discovery of a new type of land use that had never been observed before (Tu, 2010).

2 Discovering Land Transaction Relations

We start by describing the four relations among land deeds that our approach tries to capture.

> Successive transaction pairs: A piece of land could be sold from A to B, then from B to C. In this case, there should be two land deeds recording the two transactions. We call them a 'successive transaction pair'. Note that the situation could be rather complicated. For instance, it could have been B's son who sold it to C. If B divided the land among his descendents, the first selling transaction and the ensuing allotment agreement (see below) also form a successive transaction pair. Thus, a successive transaction pair can be loosely defined as a pair of deeds that record successive transactional activities involving the same land.
>
> Allotment agreements: An allotment agreement is a deed that records how a land is divided among the owner's descendants or among the shareholders. In both cases, the usual practice was to first parcel the land, then to have each participant drawing from the lot. Once the decision was agreed upon, an agreement was written and copies were made and given to each person involved. In the case of division among shareholders, the allotment agreements were usually preceded by a 'cultivation permit', a permission from the government to allow a group of people to cultivate the land. (In this case, the cultivation permit and the ensuing allotment agreement also form a successive transaction pair.) Allotment deeds from the same transaction are usually almost identical except for the name of the recipient, which is different in every deed.
>
> Red deeds (transaction deed and the sales tax receipt): When the sale of a property was conducted, the buyer was required to report to the local government about such a sale, pay land taxes, and receive a tax receipt which should be attached to the deed. This tax receipt is called *qiwei* (契尾). The deed/receipt pair is called a 'red deed' because of the red seal on the *qiwei*. (A land transaction deed without a *qiwei* is called a 'white deed'.) However, not only the majority of the transaction deeds that we know of do not have a *qiwei*, most of the *qiwei* that are included in THDL (and in other sources) are detached from the land deed with which it is supposed to be associated. Thus, the 3rd type of relationship that we want to find is to reconnect a land deed and its associated *qiwei*.
>
> Duplications of deeds: Since the land deeds in THDL came from more than 100 different sources, some of which books consisting of deeds selected by scholars, it is not unreasonable to expect the same deeds to appear in more than one source. Thus, we also try to discover duplications of a deed. Although

this seems to be a trivial task, it actually allows us to observe a peculiar case to be described later.

2.1 Algorithms for discovering the relations

Each of the relations mentioned above requires a different algorithm to discover. In the following, we give an outline of the methods we designed. To tackle the problem of finding successive transaction pairs, we developed a three-step semi-automatic process (Fig. 1). We first used text processing techniques to extract features of each land deed from its metadata and full text. Such features include the transaction type, the general location of the land and the 'four reaches' (boundaries identifying the land via some obscure way such as 'bordering Lee's house on the south' and 'a large camphor tree on the west'), the names of the people involved in the transaction and their roles (seller, buyer, and scrivener), description of the source of the land (how and when the current owner obtained it), the size, the price, and the amount of taxes paid (Lu, 2008; Huang, 2009). Figure 2 is an example of a typical land deed. We designed an XML format to hold the original metadata and the information extracted (Fig. 3). We call it the 'expanded metadata of the land deed'. Some features of the expanded metadata, such as the transaction type and the seller/buyer information of a land deed, can often be obtained directly from the original metadata. Others can be extracted from the full text using carefully designed regular expressions. For instance, let *$numchars* contain all possible Chinese characters that denote numbers. Given the full text of a land deed, we use the following regular expression to extract its transaction date:

(康熙|雍正|乾隆|嘉慶|道光|咸豐|同治|光緒|明治|大正)[$numchars]+年+月+日.

We then convert the date from Chinese calendar to the corresponding Gregorian one (for example, 光緒十三年十二月五日 becomes 17 January 1888). As another example, we use the following four regular expressions:

東至(.{2}) 西至(.{2}) 南至(.{2}) 北至(.{2})

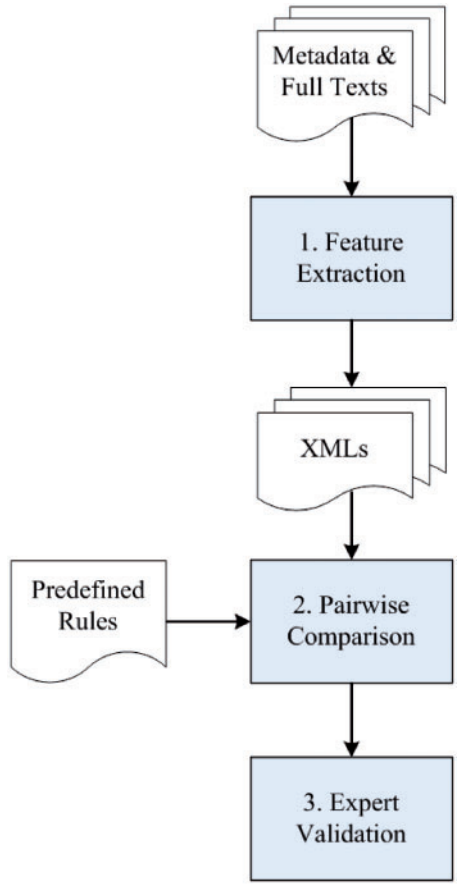

Fig. 1 The process for discovering land transaction relations

to identify the locations of the four reaches of a deed. Since the same four reaches may be written slightly differently in different deeds (e.g. the same west reach was written as 西至社寮庄背小崑透上大崗倒水爲界 and西至社寮庄背小崑透上倒水爲界 in two deeds), we extract only the first two characters of the locations of the each of the four reaches for the feature matching algorithm that we shall discuss later. Other regular expressions to extract the rest of the features can be found in Huang (2009).

Second, we defined rules that use the information in the expanded metadata to identify deeds that may be related. Figure 4 shows the rules we used for identifying the 'successive transaction pairs'. The basic idea is that if deeds A and B are a transaction

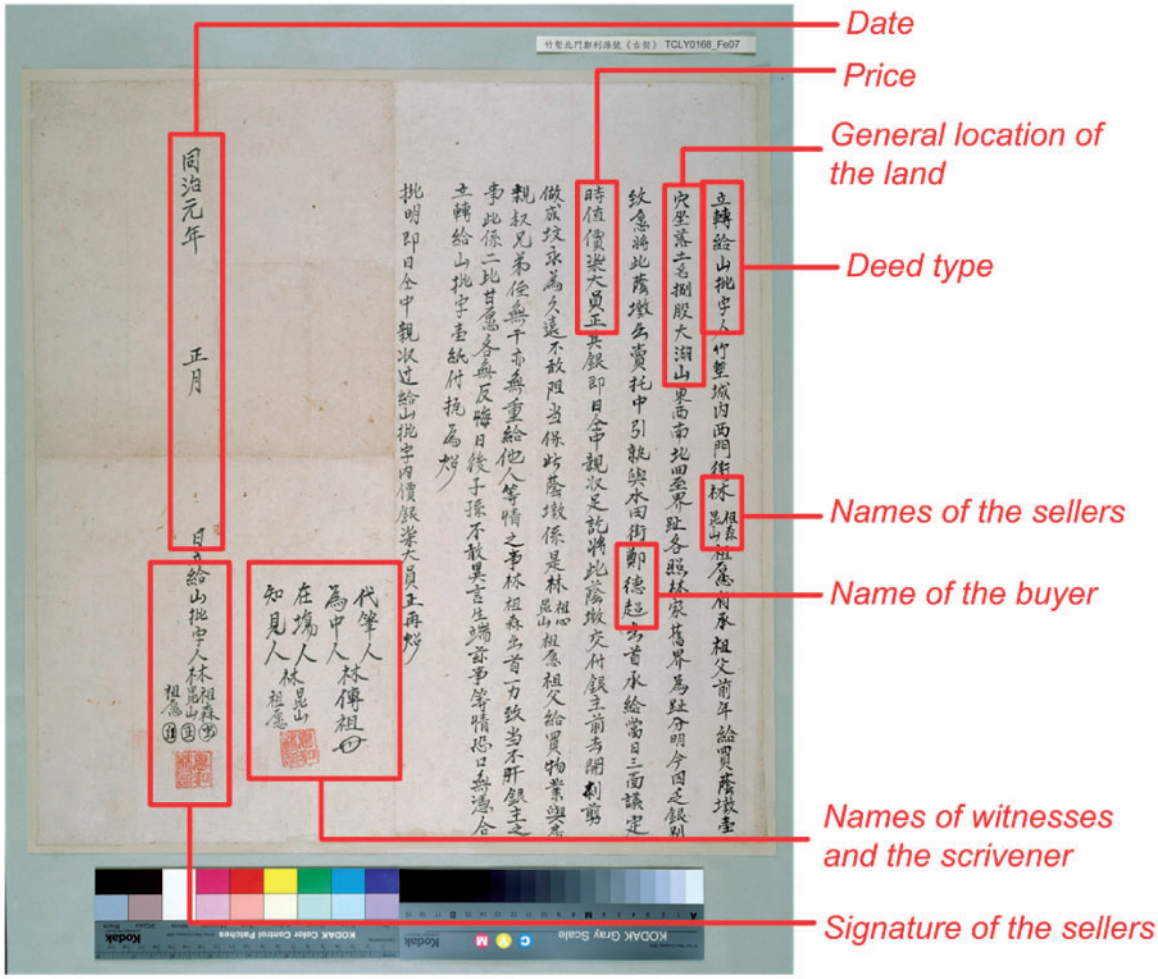

Fig. 2 An example of a typical land deed of Taiwan

pair where A precedes B, then the transaction time of A must precede that of B, the general location of the two deeds should be the same, and there should be at least one person who had been mentioned in both deeds. One problem is that the writing of a deed (usually done by a scrivener) can be imprecise. For instance, the names of the same person can be written differently in different deeds. A person named Chen Yi (陳義) in one deed was written as Chen Tongyi (陳同義) in another. (The two deeds in fact form a transaction pair.) As another example, the location *Dapingding* was written as大坪頂 in one deed and大平頂 in another. This phenomenon is quite common because the people involved, except the scriveners, were often illiterate. To circumvent this problem, we relaxed the requirements of the feature-matching algorithm to allow some degree of fuzziness, such as matching two characters out of three in a name. However, such a relaxation will produce too many false alarms. We therefore designed eight additional conditions (such as the locations of the four reaches), as described in the algorithm in Fig. 4 and required that at least one of them be met. We wrote a program to compare every pair of land deeds in THDL to see if any pair satisfied the rules (Huang, 2009). Finally, we give all the pairs produced to human experts to verify.

```
<document>
  <filename>cca100003-od-ta_05716_000115-0001-u.txt</filename>
  <collection>總督府檔案-開墾地業主權認定及池沼山林原野ヲ開墾地トシテ整理方認可（臺北廳）</collection>
  <transaction_type>杜賣契</transaction_type>
  <location>一;雙;溪;内;鷄;尾;山</location>
  <boundary_E>聖人</boundary_E>
  <boundary_W>崙脊</boundary_W>
  <boundary_S />
  <boundary_N>余家</boundary_N>
  <seller>何;長;來</seller>
  <buyer>李;崑;岡</buyer>
  <time day="18980101" month="189801" year="1898" dynasty="1868" timelevel="year">明治三十一年</time>
  <source_description>水田山園茶穫果子;先祖父遺下應得于;過何景奇等山業厝;先問房親人等不欲;受外托中引就與李;買同
    堂議定時值盡;價銀捌百貳拾大員;主前去掌管收租納;終休寸土無留來及;先祖父自置遺下應;掛他人財物與及來;主之事此
    乃明買明;二比甘愿並非迫勒;買印契連司單壹紙;茶穫果子竹木屋宇;因乏銀別用愿將此</source_description>
  <source_time>1869</source_time>
  <land_size>0.9492</land_size>
  <transaction_price>820</transaction_price>
  <tax>0.483</tax>
  <land_number>四六〇之一</land_number>
</document>
```

Fig. 3 The expanded metadata of a 'selling' type of land deed, stored in XML

A pair of land deeds (A, B) is a *successive transaction pair* if A and B satisfy the rules #1 - #3, and at least one of #4.1 - #4.8:

1. The transaction time of A < the transaction time of B
2. The general location of A = the general location of B
3. At least one person who is involved in A is also involved in B

4.1 At least one of the lot numbers of A is a lot number of B
4.2 At least one of the prices in A occurs in B
4.3 At least one of the taxes in A occurs in B
4.4 The four reaches of A match the four reaches of B
4.5 At least one of the sizes in A occurs in B
4.6 The transaction time of A matches the time mentioned in the source description
4.7 One of the buyers of A is mentioned in the source description, or one of the sellers of A is mentioned in the source description
4.8 A and B are from the same collection, and they are adjacent in the collection.

* Note that since a transaction recorded in a land deed may involve more than one piece of land, our rules require that A and B involve at least one identical piece of land.

Fig. 4 The rules for identifying successive transaction pairs

Groups of the same allotment deeds are easier to identify. If deeds A and B are the same involve the same allotment, they must both have been classified as allotment deeds, have the same set of people involved and at the same location.

If deed B is a *qiwei* of A (thus A and B form a pair of red deeds), then B must have been classified as a *qiwei*, has a date later than A, has the same location as A, and the groups of people involved in A and B must have a non-empty intersection (although need not be identical).

Checking for duplications of deeds is quite similar to allotment deeds, except that the classification of the deeds involved can be of any type.

We also remark that while the above algorithms all utilize the expanded metadata of the land deeds, we have also designed another algorithm that is based on matching the 'longest common subsequences' (Cormen *et al.*, 2003) in the full texts of two deeds to find allotment deeds and identical deeds. This method was implemented before we discovered the expanded metadata approach. It can also be applied to other corpuses to find duplications and documents with a similar pattern (Hsiang *et al.*, 2012).

2.2 Experimental results and discussions

The relations that we discovered using our methods are summarized in Table 1. The first column contains the potential candidates of relations that we found using the algorithms based on expanded metadata. They were then checked manually by experts, and the results are given in Column 2. Column 3 is the percentage of correctness. Column 4, the pairwise completion column, needs some explanation. Suppose we have found a successive transaction pair, which we call A and B. If there are two copies of deed A and three copies of deed B (discovered through the 'identical deed' operation), then there are in fact six successive transaction pairs instead of 1. We call this process of pairwise matching 'Pairwise Completion', and (B) indicates the numbers of new pairs that we have found through this operation.

Among the 32,074 land deeds in THDL that we used in this experiment (more than 5,000 had been added to THDL later), we have found 6,035 successive transaction pairs, 1,144 sets of allotment agreements, 165 pairs of red deeds, and 777 sets of identical deeds. The number of deeds we found that are related to others in some way are 7,498 or 24% of the total number of deeds. This high percentage is somewhat unexpected, considering the diverse sources from which they came. Among the successive transaction pairs found, 214 are 'cross-generational', meaning that one deed involves a person A who sold it to B, and another one involves a descendent of B, who made further actions (selling or dividing) on the same piece of land.

Table 2 gives the number of relations we have found that involve land deeds from different sources or are from the same source but were kept in different parts of the original archive. We emphasize that the pairs/sets that are from different sources (the 'cross sources' column in Table 2) would be quite impossible to find by human. Those that are kept in different parts of the original archives are also very difficult to find manually since they are not expected.

While looking for allotment deeds, we also found an anomaly that was quite unexpected. Figure 5 contains two allotment deeds that are identical except for the dates. One is the 12th month of the 10th year of Daoguan (道光), while the other is dated the 6th month of the 12th year of Daoguan. Apparently, at least one of them is a forgery, although we do not know which or why.

While checking for duplications, we made an important discovery. The largest sub-collection of land deeds in THDL, numbered at about 15,000, came from the Archives of the Japanese Taiwan Governor-Generals. These deeds were acquired through transcription, during the land reform around 1905, by the colonial government and used as proof of ownership in case land dispute occurred. It was commonly believed that the surveyors transcribed the deeds and returned the originals to the land owners. However, during the duplication checks (using both the longest subsequence method and the expanded metadata method), we noticed that there was 'no' duplications at all between the 15,000 deeds from the Japanese Archives and the rest of the deeds in THDL. Since the majority of the other deeds in THDL were collected after 1945 by scholars and local historians, it is statistically impossible not to have overlaps if the originals were indeed returned to the land owners. This could only mean that the original deeds were kept by the Japanese colonial government. However, we could not find any record or policy statement regarding their whereabouts.

3 Land Transitivity Graphs

When further examining the transaction pairs, we noticed an interesting transitive phenomenon.

Table 1 Relations among 32,074 land deeds

Relationship	Discovered automatically	Manually checked as correct (*A*)	Accuracy	Pairwise completion (B)	Final result (*A* + *B*)
Successive transaction pairs (pair)	9,630	3,834	39.8%	2,201	6,035
Successive transaction pairs—cross generation (pair)	605	200	33%	14	214
Red deeds (pair)	262	145	55.3%	20	165
Allotment agreements (set)	1,374	1,144	83.3%	–	1,144
Identical deeds (set)	1,045	777	74.4%	–	777

Table 2 Related land deeds from different sources

Relationship	Relations discovered	Cross sources	From same source but files are not adjacent
Successive transaction pairs (pair)	3,834	144	2,963
Allotment agreements (set)	1,144	44	208

There may be a deed of A selling a property to B, and some years later B divided the land among his sons, then one of them, C, rented it to D to farm. Such transitive activities on the same piece of land could last for decades. By connecting all these transactions into a graph, it may capture the evolution of a property over time.

This is exactly what we did. We call these graphs 'land transitivity graphs'. Using the relations that we discovered earlier, we came up with 2,436 such graphs. The result is listed in Table 3. With each node representing a land deed, a land transitivity graph reflects a combination of activities on the same piece of land, such as selling, renting, dividing, or the permission to cultivate. Each arrow represents a successive transaction relationship, pointing from an older deed to a succeeding one. As a small example, the graph in Fig. 6 contains four deeds, which are two title deeds and two identical allotment deeds. The person involved in this graph, Gao Tienxi (高添喜), bought a piece of land in 1903 and another (presumably adjacent) piece in 1907, then merged the two and divided the property among his descendants in the same year.

Figures 7–9 contain three of the largest graphs, each telling a different story. The graph in Fig. 7 captures the evolution of a piece of land within a family, while the one in Fig. 8 shows how a group of land developers (*kenhao*, 墾號) dealt with a piece of land newly acquired from the government. The graph in Fig. 9 represents a unique land development situation that no one had noticed before. We will describe them in more detail.

Figure 7, the 3rd largest graph, contains 36 deeds, dating from 1850 to 1910. The head of the family, Liao Jiafu (廖佳福), was among the shareholders who received a cultivation permit from the Qing government and obtained his share through allotment in 1850 (the first deed). Liao farmed the land for 50 years and divided it among his descendents in 1901 (the second deed). The rest of the deeds described the various activities such as further divisions or selling in the next 10 years. By 1906, only two of the eight parcels of land remained in the Liao family (Tu *et al.*, 2011).

Figure 8 contains 65 deeds, spanning from 1867 to 1911, and are located in the current Taipei County. The first deed is an allotment agreement (1867), which indicated that the permit of cultivation was in fact received in the 1840s (although we could not find that record). But for some reason the company, Jinfuan (金福安), that acquired the land, never developed it. Five years after the division (1872), there was suddenly a flurry of activities, most of which were further divisions of the property (as evident by the number of allotment agreements). Although Taipei is now the most populated area of Taiwan, it was not the case before the Qing government established the seat of the newly founded provincial government at Taipei fu in 1894. (The Province of Taiwan was founded in 1885.) This might explain why the land was not developed

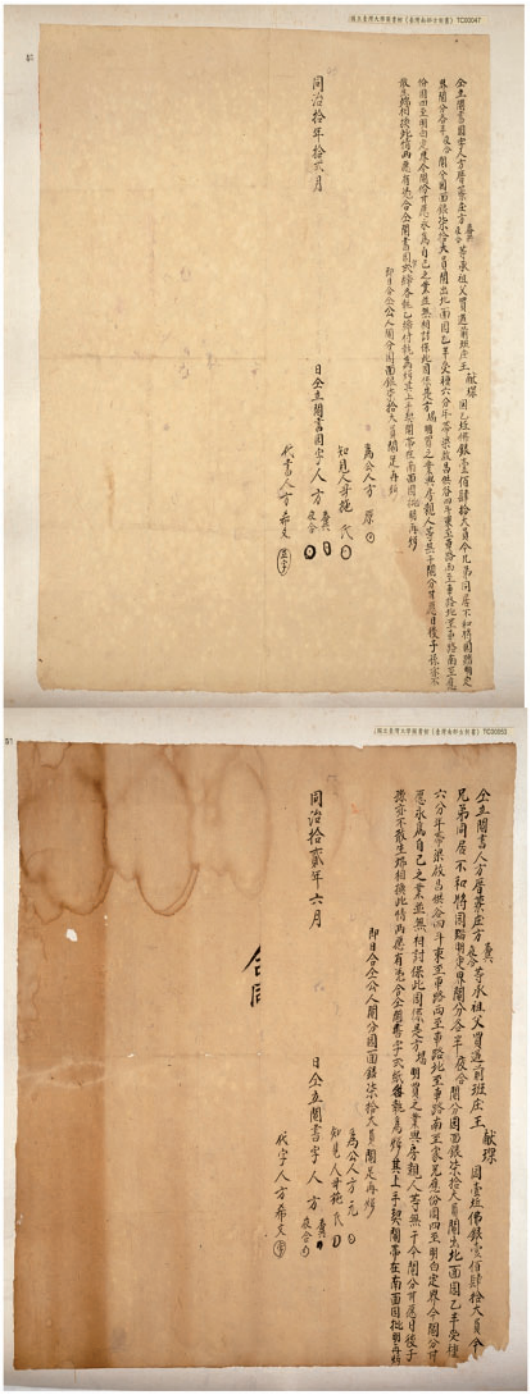

Fig. 5 Two identical allotment deeds with different dates

Table 3 Land transitivity graphs constructed among the land deeds in THDL

Deeds involved	Number of graphs
104	1
65	1
36	2
23	1
15–19	5
10–14	29
6–9	119
5	103
4	214
3	475
2	1,486
Total	2,436

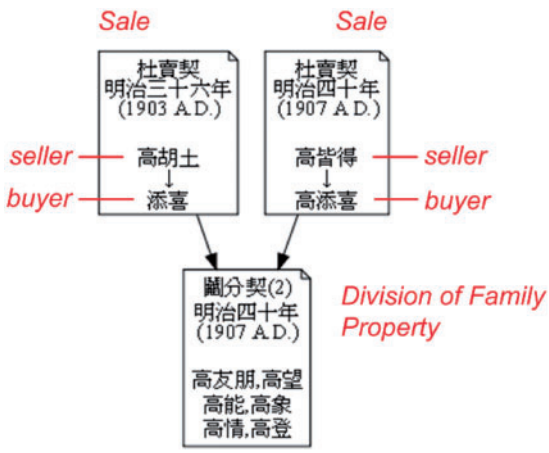

Fig. 6 A small example of land transitivity graph.

earlier. It might also because of some dispute among the group of developers. The actual reasons need further investigation.

Figure 9 is the largest land transitivity graph with 104 deeds and is also the most interesting one. The root of the graph is a permit of cultivation issued in 1894 to a person called Lin Renwen (林人文) that involves a piece of land size of 0.5 km^2 located in the current Tainan County. This graph is intriguing for several reasons. First, the area where the land is located was considered heavily cultivated as early as 1720s. (In fact, it was the first agricultural area in Taiwan.) It is thus surprising to see a cultivation permit of an area of such a size at this location some

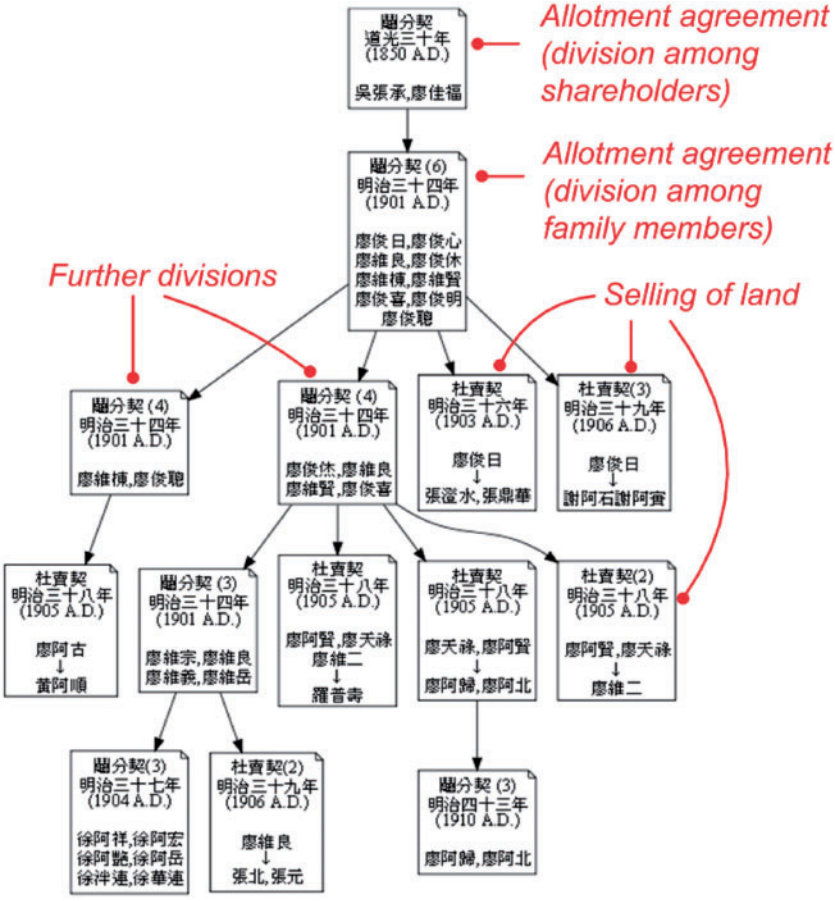

Fig. 7 The 3rd large graph, containing thirty-six deeds (including duplicates of allotment deeds)

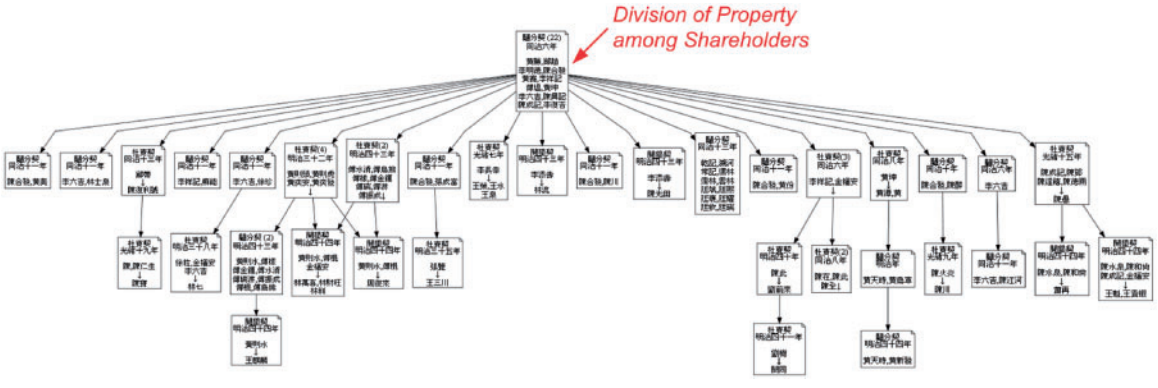

Fig. 8 The 2nd large graph, containing sixty-five deeds (including duplicates of allotment deeds)

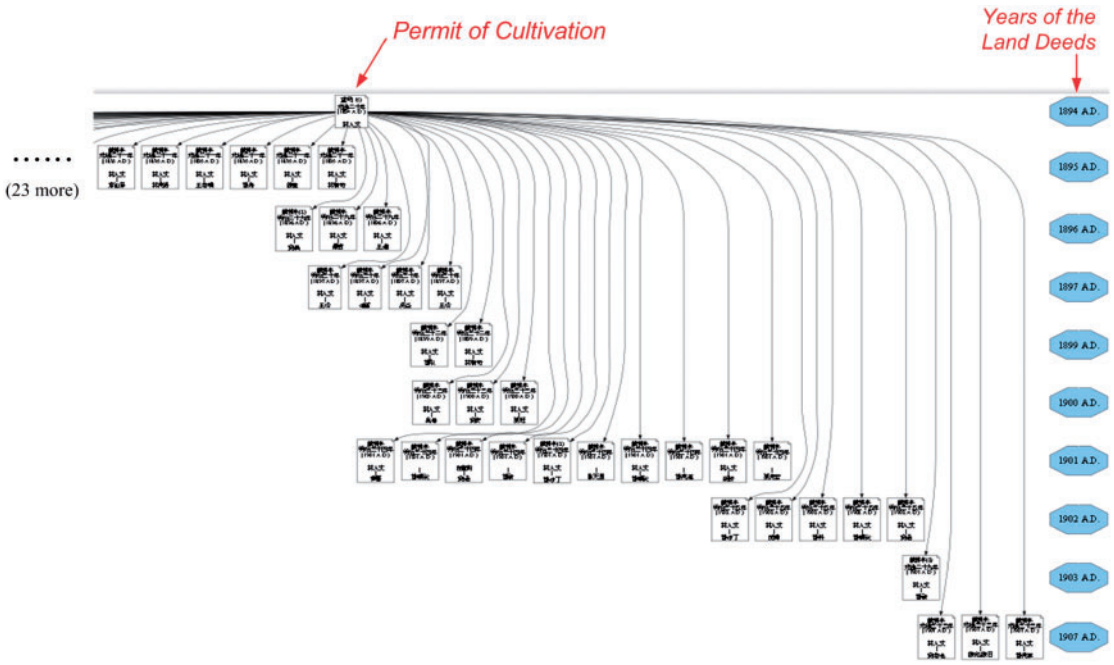

Fig. 9 The largest graph, containing 104 deeds

250 years later. It is reasonable to speculate that Lin was an important person of his time, who had received the permit through personal influence. However, further investigations revealed very little information about him (Lu, 1974). Tu studied this graph and discovered that the deeds involved demonstrated a unique case of land use that had never been noticed before (Tu, 2010). Tu discovered that the land was actually located near the river bank of Zengwen River (曾文溪), and the soil was considered uncultivable due to flash floods caused by typhoons that (still) visit the area every few years. In the next several years, after Lin received the cultivation permit, he carved up the land in pieces and rented them out to tenant farmers. The lending deeds comprise the majority of the nodes in the graph. But that is not the end of the story, because there are in fact another 330 documents involving Lin in THDL. What happened was that the river changed course after a severe storm in 1899. As a result, Lin's land became much more valuable. That initiated a litigation battle from a Yang family that claimed that Lin did not own rights to the land. It is unlikely for human to notice this possibility without the computer-generated transitivity graph.

To help historians take advantage of these graphs, we developed an integrated environment to analyze the information embedded in each graph (Fig. 10). In addition to the graph itself and its zoomable navigation facility, we also added tag cloud, chronological distribution, and a location map.

4 Discussions

In this article, we described work in THDL that is based on its land deed corpus, which currently contains 39,455 land deeds and other social contracts. (The number increases every month.) The deeds are gathered from more than 100 sources and are hundred times the number used in most research papers. One question that naturally arises is what kind of research can emerge from using such a collection. To answer this question, one needs to first discover the 'contexts' that are hidden among the deeds. However, a collection of this magnitude and diversity simply cannot be explored manually. Thus,

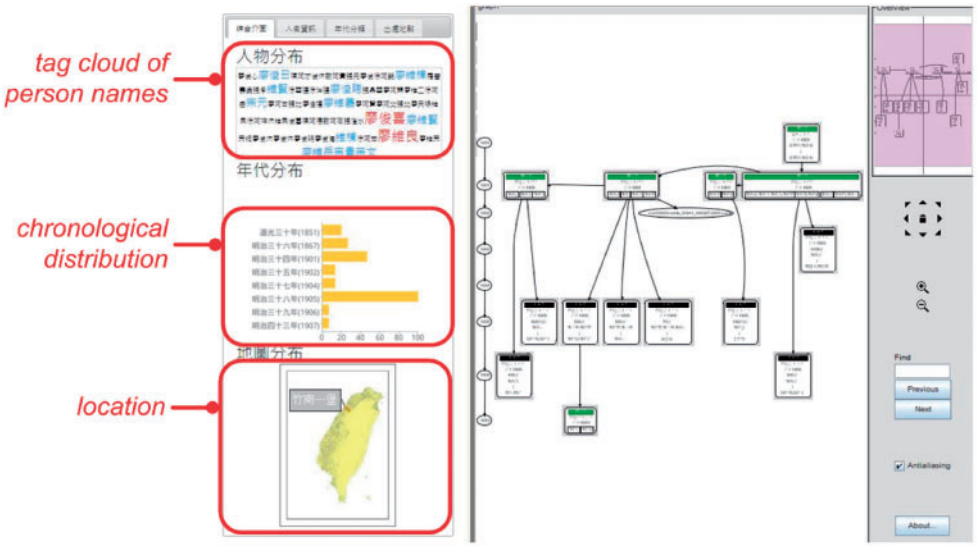

Fig. 10 The integration environment for land transitivity graphs

automated methods such as text mining must be used in order to find a large number of possible relations simultaneously. The results that we have reported in this article present some of the work in this direction. We have designed a method that has discovered 6,035 successive transaction pairs and 1,144 sets of allotment agreements, many of which are from different sources and would be almost impossible to find manually. They, in turn, are transformed into 2,436 land transitivity graphs, each of which describes the transaction evolution of a piece of land. One such graph has already led to the discovery of a unique pattern of land development that had not been studied before (Tu, 2010). Our work used 32,074 deeds of those in THDL, among which 7,498 are related to others in one way or another. As part of the process, we extracted over 90,000 person names and locations (Shieh, 2011), with which we designed a co-occurrence analysis mechanism so that a user can analyze the co-occurrence relations among the deeds that she retrieved through a query (Hsiang *et al.*, 2009).

The approach we have presented here is part of an attempt to address a broader question that has emerged in the digital age. The digitization efforts of many government agencies, institutions, and individuals in recent years have created a new form of archives, the digital archives. Unlike traditional archives whose materials are usually collected and organized in a systematic way, the contents in a digital archive may have been gathered from different sources and do not have a rigorous organization under which the documents are arranged. The land deed collection in THDL is a typical example. The lack of a clear organization among the documents makes it difficult to use such an archive. Furthermore, when a scholar uses such an archive, she is usually not simply looking for a single piece of document but rather a group of documents under certain context. Conventional retrieval systems (such as Web search engines or library automation systems) cannot deal with this problem because these systems do not consider the possible relations among the returned documents. (The concept of 'ranking' of a query result, for instance, treats documents as competing entities as opposed to related ones.) THDL is designed on a different principle. The basic assumption is that documents are 'related', and that the retrieval system should provide context of the retrieved documents in addition to retrieving them. However, since the system does not know what context the user is looking for, it should provide as many contexts about the query return as possible so that the user can observe and explore further on her own. Thus, the core of our approach is to 'treat a query result as a

sub-collection' in itself. That is, instead of returning a ranked list of documents to the user according to some internal priority function, we return the query result as a whole. We then provide additional analytical and observational tools such as post-query classification (using attributes such as year, source, type, and location), co-occurrences of names and locations, and statistical analysis and visualization tools to reveal possible collective meanings of the query result so that the user can explore further (Hsiang *et al*., 2009). For instance, the chronological distribution of the set of land deeds resulted from using with an important land developer as the keyword may reveal the pattern of his land development behavior. Term co-occurrence analysis on the same set will tell who his important associates were as well as the most significant locations in his acquisitions. Combining the two can reveal his land development strategy. All these contexts can be provided by the system and do not involve human effort.

These tools, however, are still syntactic in nature. One has to issue a query first, and only documents that are syntactically related to the query are returned and analyzed. The methods presented in this article go one step further. The relations and land transitivity graphs we described are 'semantic' contexts among the documents that cannot be obtained through syntactic query and retrieval. Thus, if querying a person's name and a land deed is returned, the land transitivity graph of that deed is also returned to the user as part of the query result. Since the graph may contain earlier deeds, which may involve the ancestors of the queried person, or later deeds, which may involve the descendents of the person, a great deal of additional information can be revealed through a syntactic query. The analytical tools that THDL provide can also include the deeds in the land transitivity graphs returned in the overall post-query processing and analysis. Thus, richer contextual analysis can be obtained and observed.

Acknowledgements

The authors would like to thank anonymous referees for their constructive comments, the members of the Research Center for Digital Humanities of the National Taiwan University, and the students of the Laboratory of Automated Reasoning and Digital Archives of the Department of Computer Science of NTU for their help in running the experiments and feedback on THDL.

Funding

The research reported in this article was supported by funding from the National Science Council and the National Taiwan University.

References

Baker, E. W. (1989). ''A Scratch with a Bear's Paw'': Anglo-Indian Land Deeds in Early Maine. *Ethnohistory*, **36**: 235–56.

Billot, C. P. (1979). *John Batman: The Story of John Batman and the Founding of Melbourne.* Hyland House, p. 330.

Bradley, J. (2007). Text Tools. In: Schreibman, S., Siemens, R., and Unsworth, J. (eds), *A Companion to Digital Humanities*. Malden, MA: Blackwell Publishing Ltd, pp. 505–22.

Chen, C. K. (1997). *Taiwan's Aboriginal Proprietary Rights in the Ch'ing Period: Bureaucracy, Han Tenants and the Transformation of Property Rights of the Anli Tribe, 1700–1895*. Taipei: Academia Sinica.

Chen, Z. (2004). *Social Contracts and the History of Tax Laws of Ming and Qing (Min jian wen shu yu Ming Qing fu yi shi yan jiu)*. Hefei: Huang Shan shu she.

Chun, A. J. (1986). The land revolution in twentieth century rural Hong Kong. *Bulletin of the Institute of Ethnology in Academia Sinica*, **61**: 1–40.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2003). Longest Common Subsequence. In: *Introduction to Algorithms*. Boston: MIT Press and McGraw Hill, pp. 350–5.

Fu, G. and Luke, K. (2005). Chinese named entity recognition using lexicalized HMMs. *ACM SIGKDD Explorations Newsletter*, **7**: 19–25.

Fu, Y. (1961). *Rural Social Economy of Ming and Qing (Ming Qing non cun she hui jing ji)*. Beijing: Shen huo, du shu, xin zhi san lian shu dian.

Goetze, A. (1939). Cuneiform inscriptions from Tarsus. *Journal of the American Oriental Society*, **59**(1): 1–16.

Hoang, P. (1920). Notions techniques sur la propriété en Chine. Variétés Sinologiques, 11. Chang-Hai: Tou-Se-We.

Hone, E. W. (2008). *Land & Property Research in the United States*. La Vergne, TN: Ingram Pub Services, p. 517.

Hong, L. W. (2005). *A Study of Aboriginal Contractual Behavior and the Relationship between Aborigines and Han Immigrants in west-central Taiwan*, vol. 1. Taichung: Taichung County Cultural Center.

Hsiang, J., Chen, S. P., and Tu, H. C. (2009). *On Building a Full-Text Digital Library of Land Deeds of Taiwan, Digital Humanities 2009 Conference*. Maryland, pp. 85–90.

Hsiang, J., Chen, S. P., Ho, H. I., and Tu, H. C. (2012). Discovering relationships from imperial court documents of Qing China. *International Journal of Humanities and Arts Computing*, **6**: 22–41.

Huang, Y. M. (2009). *On Reconstructing Relationships among Taiwanese Land Deeds.* Master thesis, National Taiwan University, Taipei, Taiwan.

Ikeda, O. (1986). A Study on Chinese Social Agreements. *Research on Historical Materials of China and Korea (Chugoku chosen monjo shiryo kenkyu)*. Tokyo: Institute of East Asian Culture, University of Tokyo.

Ka, C. M. (2001). *The Aborigine Landlord: Ethnic Politics and Aborigine Land Rights in Qing Taiwan (Fan tou jia: Qing dai Taiwan zu qun zheng zhi yu shu dan di quan)*. Taipei: Academia Sinica.

Kishimoto, M. (1997). *Price and Economic Fluctuation of Qing China (Shindai Chugoku no bukka to keizai hendo)*. Tokyo: Kenbun Shuppan.

Li, J. (2010). Research trends on Taiwan's social contracts. *Taiwan Fen Wu*, **60**: 101–59.

Li, W. L. (2004). Land deeds and land administrative documents—interpreting the Archives of the Japanese Taiwan Governor-Generals. *Taiwanese History Research*, **11**(2): 221–40.

Lu, C. C. (2008). *Automated Classification of Taiwanese Land Deeds.* Master thesis, National Taiwan University, Taipei, Taiwan.

Lu, J. X. (1974). Lin Renwen, a Tainan teacher who wrote revised Three-word Chant. *Nanying Documentary*, **19**: 1–12.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, **30**: 3–26.

Nguyen, T. and Shimazu, A. (2007). *Acquisition of Named-Entity-Related Relations for Searching, Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*. Seoul, Korea, pp. 349–57.

Shieh, Y. P. (2011). Appositional Term Clip: A Subject-Oriented Appositional Term Extraction Algorithm. In: Hsiang, J. (ed.), *New Eyes for Discovery: Foundations and Imaginations of Digital Humanities*. Taipei: National Taiwan University Press, pp. 133–64.

Shiga, S. (1967). *Principle of Chinese Family Laws (Chugoku kazokuho no genri)*. Tokyo: Sobunsha.

Shih, T. F. (1995). Historical Research with Regional Geography—A Case Study with Qing Anli Region. In: Huang, Y. G. (ed.), *Space, Force, and Society*, vol. 84. Bulletin of the Institute of Ethnology.

Shih, T. F. (2001). *Local Society in Qing Taiwan*. Hsinchu: Cultural Affairs Bureau.

Sun, H., Zhou, M., and Gao, J. (2003). A class-based language model approach to Chinese named entity identification. *Linguistics and Chinese Language Processing*, **8**: 1–28.

TB. (1960–63). *Taiwan Private Laws (Taiwan si fa), Taiwan wen xian cong kan,* no 79, 91, 117, 150, Taiwan Bank.

TB. (1963). *Investigation of Taiwan's Grand Leases (Qing dai Taiwan da zu diao cha shu). Taiwan wen xian cong kan,* no 152, Taiwan Bank.

Terada, H. (2005). The nature of social agreements (yue) in the legal order of Ming and Qing China (Part 1). *International Journal of Asian Studies*, **2**: 309–27.

Terada, H. (2006). The nature of social agreements (yue) in the legal order of Ming and Qing China (Part 2). *International Journal of Asian Studies*, **3**: 111–32.

Tu, F. E. (2010). *Environmental Change, Land Development and Dispute over Property Rights in southern Taiwan (1890–1920), The Sixth Conference of Taiwan Colonial Government Archives*, Taiwan Historica.

Tu, F. E., Tu, H. C., Chen, S. P., Ho, H. I., and Hsiang, J. (2011). Information Technology and Open Problems in the Taiwan History Digital Library. In: Hsiang, J. (ed.), *New Eyes for Discovery: Foundations and Imaginations of Digital Humanities*. Taipei: National Taiwan University Press, pp. 21–44.

Wang, Z. (2002). *The Social and Cultural History of Huizhou (Huizhou she hui wen hua shi tan wei)*. Shangai: Shanghai she hui ke xue yuan chu ban she.

Yang, G. (1988). *Research on the Ming and Qing Land Deeds (Ming Qing tu di qi yue wen shu yan jiu)*. Beijing: Ren min chu ban she.