On Building a Full-Text Digital Library of Land Deeds of Taiwan

Jieh Hsiang  
National Taiwan University  
jhsiang@ntu.edu.tw

Szu-Pei Chen  
National Taiwan University  
gail@turing.csie.ntu.edu.tw

Hsieh Chang Tu  
National Taiwan University  
tu@turing.csie.ntu.edu.tw

In this paper we present a full-text digital library of Taiwanese land deeds. Land deeds were the only proof of land activities such as transaction of ownership and leasing in Taiwan before 1900. They form a major part of the primary documents at the grassroot level in pre-1900 Taiwan, and are extremely valuable for studying the evolution of the Taiwanese society.

Land deeds, on the other hand, are difficult to study because they are hand-written and hard to read. Furthermore, they are scattered in many different locations and, in some cases, in the hands of families and private collectors.

In order to make the land deeds more accessible to researchers and educators, the Council for Cultural Affairs of Taiwan embarked on a major effort to organize available land deeds and typed them as machine readable full-text. Based on this collection and collections from other sources, the National Taiwan University built a full-text digital library of Taiwanese land deeds. The current size of the collection is over 23,000 which, according to one estimation, cover about 50% of all existing land deeds. The collection will be expanded to around 30,000 by the end of the year.

Our digital library is built with the goal of providing an electronic research environment for historians to conduct research using land deeds. Thus in addition to providing full-text search and retrieval, we developed a concept of regarding the query return as a sub-collection and built tools to help the user find meaning and relationships at the collection level. Post-processing presentation, term frequency analysis and co-occurrence, and relation graphs are some of the tools described in this paper. We believe that our digital library will bring Taiwanese historical research using land deeds to a different horizon.

Land Deeds of Taiwan

Before Taiwan was ceded to Japan by the Chinese Qing Dynasty after the Sino-Japanese War of 1895, land deeds were the only proof of ownership and transaction of land in Taiwan. Land deeds were, thus, a centrally important primary source for studying the 300 years written history of pre-1900 Taiwanese society (Wu, Ang, Lee, Lin, 2004). Even after the modernization of land administration by the Japanese, many families still kept the old land deeds either as part of the family heritage or due to the mistrust of the government. Many, however, were destroyed or discarded since they lost their original significance. During the early stage of the Japanese occupation, the government conducted research on the old administrative systems and customs and produced three series of books totalling 40 volumes, many of which contained transcriptions of land deeds as examples. During the reform of the land administration, the Japanese government sent surveyors to systematically transcribe land deeds so that they can be convert into modern land administrative records. In the latter endeavor, about 16,000 were collected. Their transcribed versions (copied verbatim by hand) are scattered in the 13,855 volumes of the Archives of the Japanese Taiwan Governor-Generals (Wang 1993). After Taiwan was returned to Chinese rule in 1945 after the 2nd World War, some research institutes and researchers recognized the importance of the land deeds and made efforts to collect them. The most notable, and largest scale such effort was conducted between 1976 and 1983 by a team lead by the historian Wang Shih-Ching who, commissioned by the Committee for Taiwan Historical Studies, Association for Asian Studies, U.S.A., collected about 5,600 land deeds and published a six volume catalog Taiwanese Historical Documents in Private Holdings (Wang, 1977). The photocopies of the land deeds were bound into more than 100 volumes. Other notable collections were kept at the National Taiwan University, the Institute of Taiwanese History of the Academia Sinica, the National Taiwan Library, and by various private collectors. Scores of books containing the images of some land deeds have appeared in the past ten years. Wang estimated (1993, pp. 71) that there are 20,000 land deeds in the hands of private collectors, libraries, museums and research institutes that were not included in the official collections. That makes the total number of such land deeds about 40,000. Our experience in the past ten years of digitization tells us there should be more, although we cannot give a reasonable estimation.
What is in a Land Deed?
Land deeds of Taiwan are contracts about various actions involving lands, such as the commission by the government to cultivate previously un-owned land, the division of family properties, the transaction of ownership, the rental of farming right, the pawning of land, etc. They were handwritten and were usually prepared by a scrivener. A land deed usually consists of the following elements:

- The type of the land deed: selling, renting, pawning, etc.
- The “seller” (or owner) of the land,
- The “buyer” (or lender) of the land,
- The location of the land and its boundaries (usually marked on all four directions using neighboring landmarks such as river, road, building, pond, or even trees),
- The cost: money, maybe accompanied by other properties such as houses, cows, storage sheds, and farming tools,
- The names of witnesses and the scrivener, and
- The date.

The following is an example of a typical land deed.

What Kind of Research can be done with Land Deeds
While each land deed may have significance only to its owner, the collection as a whole provides a fascinating glimpse into the pre-1900 Taiwanese grassroots society. Through these land deeds, one can study the development of a region, or the overall land management, society, economy, and law of pre-1900 Taiwan. Furthermore, since many of the deeds were contracts between indigenous people of Taiwan and the Han immigrants from China, they also provide clues to the intricate relationship among the various peoples of Taiwan (Hong, 2002), the transition of rights to land, and the gradual assimilation of the indigenous people (in particular the Pinpu 平埔族群) into the Han society.

<table>
<thead>
<tr>
<th>Org.</th>
<th>Collection</th>
<th>Number of Land Deeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTL</td>
<td>The Archives of the Japanese Taiwan Governor-Generals</td>
<td>15,899</td>
</tr>
<tr>
<td></td>
<td>Published Materials</td>
<td>1,831</td>
</tr>
<tr>
<td></td>
<td>Private Collections</td>
<td>171</td>
</tr>
<tr>
<td>NTU</td>
<td>Anli Dashe Archive</td>
<td>2,653</td>
</tr>
<tr>
<td></td>
<td>Published Materials</td>
<td>767</td>
</tr>
<tr>
<td></td>
<td>The Collection of Zheng Family of Xinzhu</td>
<td>383</td>
</tr>
<tr>
<td></td>
<td>Land Deeds from the Department of Anthropology of National Taiwan University</td>
<td>362</td>
</tr>
<tr>
<td></td>
<td>Land Deeds from the Taipei City Archive</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>Land Deeds of Southern Taiwan</td>
<td>87</td>
</tr>
<tr>
<td>TCCC</td>
<td>Published Materials</td>
<td>944</td>
</tr>
<tr>
<td></td>
<td>Total Number of Land Deeds</td>
<td>23,250</td>
</tr>
</tbody>
</table>

Our Collections of Land Deeds of Taiwan
In 2003 and 2004, the Council for Cultural Affairs (CCA) of Taiwan commissioned the National Taichung Library (NTL) and Professor Lee Wen-Liang of NTU 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 15,899 land deeds from the Archives of the Japanese Taiwan Governor-Generals. In the meantime, National Taiwan University (NTU) and Taichung County Cultural Center (TCCC) also digitized their own collections of land deeds, most notably the Anli Dashe Archive. Together, NTL, NTU, and TCCC have collected more than 23,000 land deeds in Taiwan, all incorporated into Taiwan History Digital Library (THDL), a full-text digital
library of primary historical documents that we built to serve as a research environment for researchers in Taiwanese history and other disciplines. All of the deeds are available in searchable full text, with metadata and, in some cases, images.

The building of content is an on-going effort. We project that the size of our collection will reach 30,000 by the end of the year.

**A Research Environment for Land Deeds**

We have incorporated the above-mentioned collections of Taiwanese land deeds into Taiwan History Digital Library (THDL) (Chen, Hsiang, Tu, and Wu, 2007), which is built with the goal of providing an electronic research environment for historians. Since our primary goal is to build a digital library to be used by researchers, we spent a great amount of time interacting with historians and built tools that they would find useful in their research. Full-text search is, in our view, the most basic facility. However, what is more important is how to help the user analyze the query results once they are retrieved.

We developed a methodology that treats query returns as a *sub-collection*, instead of as individual (and independent) documents. This seems to reflect better the need of researchers, who usually look for significance emerged from a set of land deeds. Under this philosophy, we have built extensive *post-query classification* facilities, which classify and present the query results according to attributes such as year, type of deeds, origin, etc. We also provide *term frequency analysis* which, using the 50,000 terms (names and locations, mostly extracted automatically) appeared in the collection, analyzes relationships such as geographic locations, co-occurrences, people involved.

The co-occurrence and temporal relationships are further analyzed in the *line chart of temporal distribution* facility provided, which gives a visual representation that makes observation easier. As mentioned before, each land deed features a list of attributes. These attributes can, in principle at least, be extracted from a deed. This work is quite laborious and is still under way. But we have developed an XML format that captures the attributes and, more importantly, makes it easy to build *relation graphs* that show the relationship among land deeds. Our preliminary experiments show that these graphs can play a significant role in the study of land deeds.

In the following, we present the aforementioned features in more detail.

**Query Returns as a Sub-Collection**

Historians usually do not look at a single document but, rather, a group of documents and try to find significance through their relationship. For example, land deeds from the same region as a whole may reveal the gradual change of land ownership from one ethnic group to another which, obvious, cannot be observed from a single document. For this purpose, we developed a concept that regards the query returns as a *sub-collection* and built tools to help the user find meaning and relationships at the collection level. This is done in THDL mainly through *post-processing* a query’s returns, presenting and analyzing them as a whole.

**Post-Query Classification**

Figure 2 is an example of how query results are presented in THDL. After the query results (the sub-collection) are returned, THDL classifies the resulting land deeds according to three predefined dimensions (year, origin, and type) on the left of the web page, while presenting summaries of the land deeds themselves on the right. Each class is followed by the size of the class (Fig. 3). By representing post-query classification, the historian can observe the distribution and behavior of the sub-collection, and see if there is anything that contradicts what the historian predicts. At a first glimpse, the historian can quickly capture the outline of the sub-collection. It’s helpful especially when the query results of full-text search is too large to manage. Furthermore, the post-query classification can also be used as a faceted search: simply click on a class will refine the user’s query.

Note that the three dimensions are chosen because they are important characteristics of land deeds. A different corpus could define a completely different set of dimensions to reflect the characteristics of the content.

![Fig. 2 An outline of THDL right after a query](image-url)
Line Chart of Temporal Distribution
The post-query classification on year reviews the temporal distribution of a sub-collection. To better visualize the temporal distribution of a query’s returns, we have built a tool to draw a line chart for any given query. It is especially useful when comparing the temporal distributions of two queries at a same time. For example, when a historian suspects that there is dependency between two concepts and wants more analysis, she can simply input each concept as a query, and then get a line chart (Fig. 4). The line chart of Fig 4 suggests that the two concepts are quite correlated.

Term Frequency Analysis
We have developed a term extraction method for extracting noun phrases from old Chinese text (Chang 2006). In the land deed corpus, we have successfully extracted 40,000 names of people and 7,000 names of locations from metadata records and from full text (Chang, 2006). THDL uses the names to provide term frequency analysis by calculating the numbers of times each name appears in the sub-collection and representing the result in tables alongside the full text of resulting land deeds (Fig. 5). The names are listed in descending order according to their document frequency (DF, the number of documents in which a name appears). The user can use the tables to observe the relevance among locations and people in the sub-collection. Fig. 5 shows the returns of the query “Jin Guang Cheng” (金廣成), a local reclamation cooperative in the Guanxi (閩西) area. At a closer look at the tables of names (Fig. 6) shows that the people on top are indeed the major shareholders of Jin Guang Cheng. Similarly, the locations on top are exactly the locations where Jin Guang Cheng claimed lands back to 1880s. However, the person with the highest DF only appears in 38 documents, while the size of the sub-collection is 61, showing that none of the people appear dominantly in the sub-collection. On the contrary, the locations with the highest DF, “Shiliao Zhuang” (石燎莊) and “Zhubei Er Bao” (竹北二堡), appear in 57 and 56 documents accordingly, showing the lands Jin Guang Cheng claimed were mostly around the same area.

Relation Graphs and Role Analysis
Each land deed should, in principle, include all the attributes we mentioned earlier in the paper. It is thus desirable to extract them so that analysis can be done more easily. We have developed an XML format and have already extracted and analyzed about 13,000 of these deeds (with the attributes of such land deed represented as an XML file). Fig. 7 shows an example of these XML
We have also developed a way to show the inter-relationship of the XML files via a notion of relation graphs. These graphs have been used to conduct role analysis, which shows how a specific person, family, or cooperative is involved in the development of a certain region through time. It is done by unfolding the roles they played in land deeds. For example, Fig. 8 shows the role analysis of “Lin Benyuan” (林本源), a cooperative that the well-known Lin family of northern Taiwan set up to represent the family in land acquisition. We found that most of the land deeds involving Lin Benyuan in our collection are sales of lands or certificates of lands from the government. Furthermore, Lin Benyuan was the buyer in all the sales and was the landowners in all the certificates. This observation matches the general impression of the Lin family, which has been one of the wealthiest families in Taiwan since late 19th century till now. The timeline of the deeds also shows that they focused on (and systematically) acquiring lands from one geographic location before moving on to the next.

**Concluding Remarks**

In this paper we described the land deed portion of THDL (Taiwan History Digital Library) that we build with the goal of providing a research environment with primary documents in full-text for research in Taiwanese history. We have developed a concept of regarding query results as a sub-collection, and have built tools that help users observe the relationship and collective meaning of a set of documents. On the aspect of land deeds, we have noticed that most of the existing research in Taiwan on this subject had used no more than a few old deeds (often within a hundred). It would be interesting to see, with over 20,000 land deeds available in searchable full-text and with tools to help discovering and analyzing their relationship, what kind of research issues can emerge and what kind of observations can be made.

**Notes**

This estimation, however, could be significantly lower than the real number.

**Acknowledgements**

The authors gratefully acknowledge the National Taichung Library for their effort in producing the full text of many of the land deeds mentioned in this article; the Council of Cultural Affairs of Taiwan for authorizing the inclusion of those deeds into our system; the staff of the Special Collections Division of the National Taiwan University Library in producing the full text of their collection and for further checking the typing correctness of all the deeds; Professors Wu Micha and Lee Wen-Liang of the Department of History of NTU for their continuing support; the members of the Lab of Digital Archives and Automated Reasoning of the Department of Computer Science of NTU for producing many of the tools mentioned here; and the staff of the Research Center for Digital Humanities for their general support.
Predicting new words from newer words: Lexical borrowings in French

Paula Horwath Chesley
University of Minnesota
ches0045@umn.edu

This study models the integration of new lexical borrowings into French, a language in which new lexical borrowings are common. Our goal is to predict whether or not a new lexical borrowing will "survive" the onslaught of time and be integrated into French.

In linguistics, most theories of word formation have been conducted in the generative tradition, such as those taken by Aronoff (1976), Selkirk (1982), Halle & Marantz (1993), and Ussishkin (2005). These approaches work well for new words formed by affixation and address, for example, how to form the neologism hateable from hate according to the same rules from which we have love --> loveable. Yet these theories have not addressed the productivity of borrowings. Although borrowings may have internal morphological structure in the donor language, their adoption in French is not governed by structural rules as studied in theoretical morphology. The goal of the present study is to address the non-structural factors that codetermine whether a borrowing will find its way into the vocabulary of the recipient language.

Although many words from other languages enjoy ephemeral use, the borrowings that become entrenched in the language are a highly constrained subset of the possible borrowings: new words do not occur indiscriminately. Several factors may promote entrenchment in the recipient language’s lexicon.

First, the DONOR LANGUAGE of a borrowing may play a role in lexical integration. For example, borrowings from a prestigious language like English could be more likely to be integrated into the French lexicon than borrowings from a less prestigious language like Polish. Second, a borrowing’s FREQUENCY at a given moment in time could be an influential predictor about the borrowing’s integration into the language at a later point in time. Third, a borrowing’s DISPERSION—the number of different text chunks a word occurs in if a text is divided into several sub-parts—also promises to be an worthwhile predictor. The more writers/speakers use a borrowing, the greater likelihood it has of becoming entrenched in the language community. Fourth, since shorter borrowings require less processing effort, we...