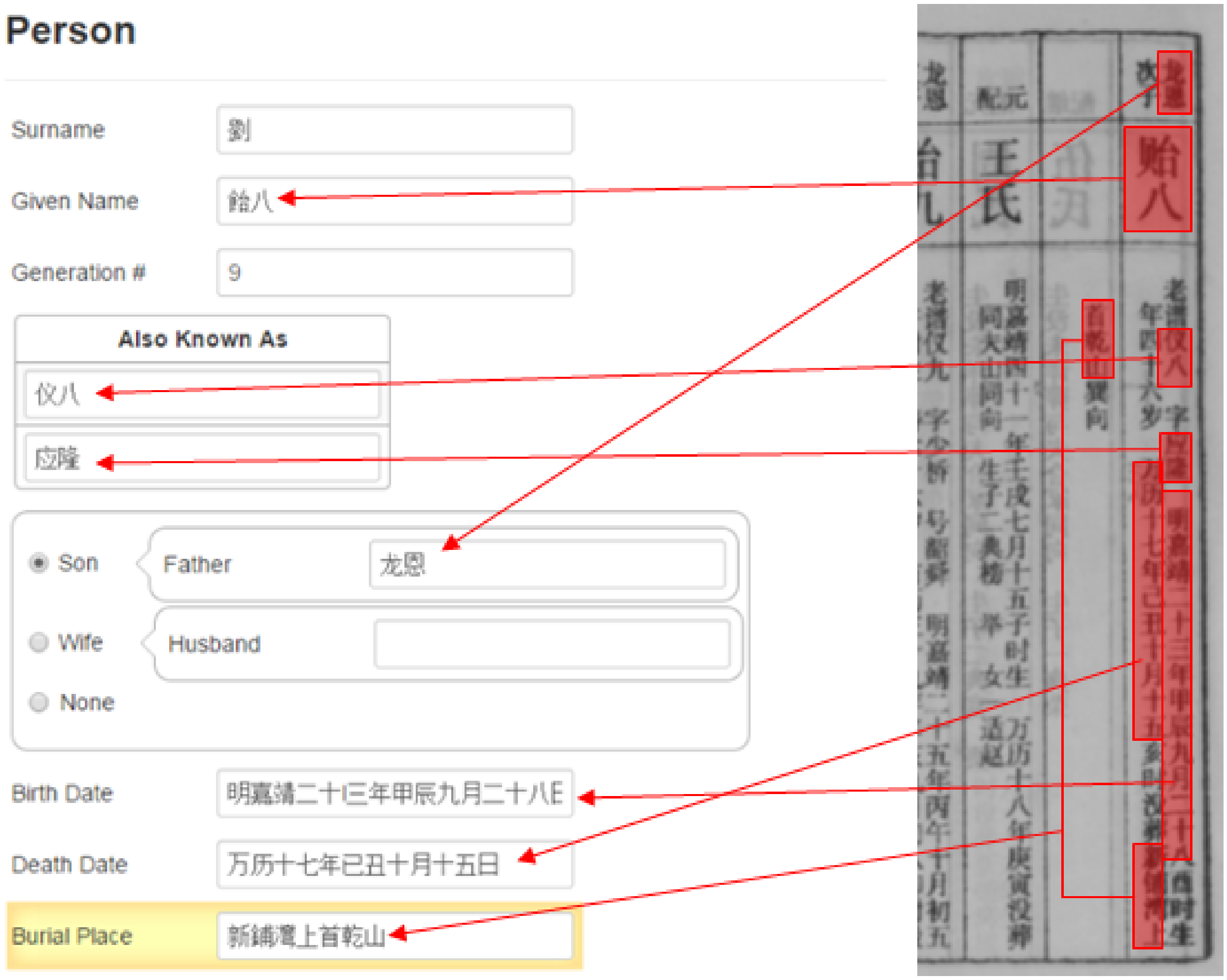
Enabling Efficient Chinese Jiapu Information Extraction
(Extended Abstract)

Stephen W. Liddle
BYU Information Systems Department

Derek Dobson, David W. Embley, Chuck Liu
FamilySearch

1 Introduction

“Jiapu” means “clan family record.” Currently, 12,871,979 images of Chinese Jiapu are available for browsing on FamilySearch.org [6]. Figure 1 shows a snippet of one page image, with the vital information for one person extracted into a form (thus making the information from the image importable into a structured data store for search and query). About half of the nearly 13 million images contain genealogical information, and each of these genealogically rich images asserts about ten facts for about eight people. Thus, as a rough estimate, the FamilySearch collection of Jiapu images contains about a half billion assertions, all needing to be extracted and recorded.



Person

Surname	劉
Given Name	飴八
Generation #	9

Also Known As

- 仪八
- 应隆

● Son — Father: 龙恩
○ Wife — Husband:
○ None

Birth Date	明嘉靖二十三年甲辰九月二十八日
Death Date	万历十七年已丑十月十五日
Burial Place	新鋪灣上首乾山

Figure 1: Filled-in Form from Chinese Jiapu.

Information entry via a keyboard is tedious and even more so for Asian Script Languages than for Western European Languages. To enter native scripts, users acquire an add-on program called the Input Method Editor (IME). Phonetic entry is the most common way to enter a Chinese character. Users type the pronunciation of a character on the keyboard and select the intended one from the list of IME-displayed characters. Many Chinese characters used in the ancient books such as genealogy collections are unfamiliar to most users, which renders phonetic input enigmatic. The alternative is to use radicals for character construction, which some IMEs support, but the construction process is tedious, time-consuming, and unpleasant.

2 Experimental Prototype

In search of a more efficient way to extract information from Chinese Jiapu, we are exploring the possibility of using COMET, our **C**lick-**O**nly (or at least **M**ostly) **E**xtraction **T**ool [3], coupled with our approach to automated information extraction [4]. In Figure 1 the form on the left is filled in by selecting and clicking on characters in the image on the right. To make this work, we preprocess images by running them through an OCR engine, aligning the resulting output characters with the characters in the image, and superimposing them in an HTML page over the original image. Then when the focus is on a form field (e.g. the highlighted last field in Figure 1), a user can click on a character in the image or mouse-select a group of characters in the image to transfer their OCR equivalents to the form field. Although no typing is needed, the OCR is not always correct, and COMET has ordinary type-in facilities for editing any mistakes made in transferring information from image to form. Also, users may type in off-page information such as the surname and generation number in Figure 1; COMET also lets a user scroll to previous or subsequent pages where the information appears and can thus be obtained in a click-only mode of operation.

The manual system, by itself, should mostly obviate the need for tedious keying—a huge win for Asian Script Languages. The automated information-extraction component of the system can potentially further reduce the work of the user by automatically filling in the form and leaving it for the user to verify and correct any mistakes the system makes. We rely on semi-supervised machine-learning [9] and cognitive-based natural language processing [7] for automated form-filling. Ideally, both the OCR and the system-learned extraction are perfect, and there is thus no need for human intervention. More practically, however, human involvement will continue to be necessary, albeit with ever-decreasing time and effort as automated extraction tools improve.

We have built an experimental prototype of COMET for Chinese Jiapu [5]. We adapted COMET for Chinese Jiapu rather rapidly, with about 75 man-hours of work, and we learned a lot about our proposed approach for efficiently extracting information from Chinese Jiapu.

3 Discussion

Given several example sets of Jiapu pages from various source documents in TIFF format, our Chinese-COMET prototype converts these images into PDF documents with accompanying OCR information. COMET takes PDF documents that have a text layer and converts them into web pages that present both the page image along with the page content represented as text in the DOM tree. By default these text nodes are hidden, although COMET allows the user to see them if desired. Figure 2 shows a snippet of a Chinese Jiapu with OCRed glyphs superimposed in red over the original black glyphs.

In our initial implementation of Chinese-COMET both OCR accuracy and character alignment became problematic:

- The OCR processing step is nontrivial. We tried three different OCR tools for converting images to text: Adobe Acrobat Professional [2], Tesseract [10], and ABBYY FineReader [1].

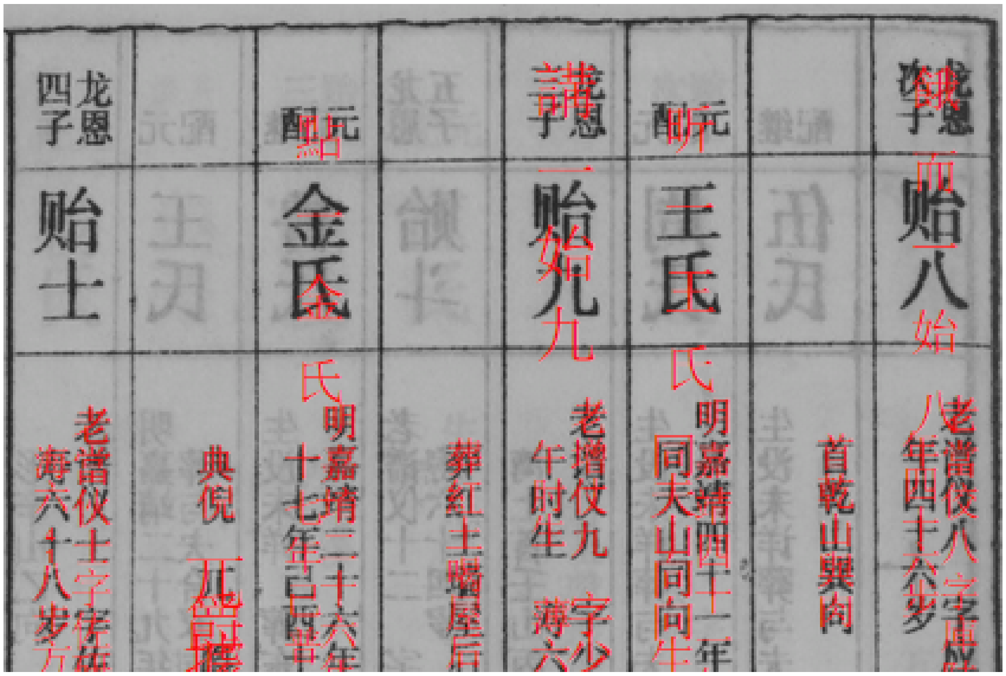


Figure 2: OCR and Alignment Quality.

In our informal comparisons, ABBYY FineReader was best, Tesseract performed moderately well, and Adobe Acrobat Professional did a bit better than Tesseract. The best, however, was far from acceptable. As Figure 2 shows, many glyphs are missed, sometimes several glyphs are read as one (e.g. the four glyphs in the upper right corner), and other glyphs are interpreted incorrectly.

- Providing accurate location information for recognized characters is much harder than expected. PDF is an ever-evolving standard with numerous variations. We used an open source library, PDFBox, from the Apache Foundation to read the PDF documents and extract character information. It seems that Tesseract and ABBYY do not provide accurate character positions, at least in a way that PDFBox can understand. PDFBox is able to handle Chinese documents processed by Acrobat Professional. However, as Figure 2 shows, it misplaces many characters.

4 Lessons Learned and Future Directions

Lessons learned from our initial exploration point to three issues to be resolved before COMET can enable users to more efficiently extract information than the alternative of keying all the information by hand.

1. *Better OCR.* Do better OCR engines for Chinese exist? Can we make the OCR engines more accurate?—for example, by image binarization techniques, by zoning different-size glyphs before applying OCR, by using an ensemble of OCR engines [8], by training OCR engines on the peculiarities of the family-history domain.

2. *Better alignment of OCR output with glyphs* Can we resolve the esoteric intricacies of the various tools to accurately align OCR character output to image character input?

3. *A cleaner user interface attuned to Asian conventions.* Can we fix finicky browser issues such as the difficulty of selecting a column of glyphs? How should we rework annotation conventions

to best suit Asian participants? Should we, for example, favor a document-markup motif over the form-fill motif?

Given that these issues can be resolved, we are excited and encouraged about the possibilities for automating Chinese Jiapu extraction. For example, could we automatically zone the various Jiapu pages and pre-populate portions of the form based on regions that we know represent surnames or relationships? How many different kinds of Jiapu pages are there, and how much variability exists in the set of nearly 13 million? Can we automatically identify generation numbers and help fill in the father/husband field when we process a son/wife record? This appears to be a rich space to explore.

References

[1] Abbyy finereader. http://finereader.abbyy.com/professional.

[2] Adobe acrobat professional. http://www.adobe.com/products/acrobatpro.html.

[3] COMET: Click-Only (or at least Mostly) Extraction Tool. http://dithers.cs.byu.edu/annotator2.

[4] Data Extraction Group web site. http://deg.byu.edu.

[5] COMET for Jiapu. http://dithers.cs.byu.edu/annotator.jiapu.

[6] FamilySearch Jiapu Collection. https://familysearch.org/search/collection/1787888.

[7] Peter Lindes. Ontosoar: Using language to find genealogy facts. Master's thesis, Brigham Young University, 2014.

[8] William B. Lund. *Ensemble Methods for Historical Machine-Printed Document Recognition.* PhD thesis, Brigham Young University, 2014.

[9] Thomas L. Packer. *Scalable Detection and Extraction of Data in Lists in OCRed Text for Ontology Population Using Semi-Supervised and Unsupervised Active Wrapper Induction.* PhD thesis, Brigham Young University, 2014.

[10] Tesseract. http://code.google.com/p/tesseract-ocr.