Leveraging Corpus Knowledge for Historical Chinese OCR

Donald Sturgeon

As an increasingly large amount of pre-modern Chinese writing is transcribed into digital form, the resulting digitized corpus comes to represent an ever larger fraction of the total body of extant pre-modern material. Additionally, many distinct items from the total set of pre-modern writings to which one might wish to apply OCR are either non-identical editions of the same abstract work, or commentaries on (and thus repeat much or all of the content of) earlier works. As a result, for historical OCR the probability that a text we wish to recognize contains extensive overlaps with what has previously been transcribed in another document is not only significant but also increases over time as more material is digitized. While general techniques for improving OCR accuracy using language modeling can also be applied successfully to historical OCR, it is also possible that more specialized techniques can take greater advantage of our more extensive knowledge of the historical corpus to further improve recognition accuracy. In this paper, I present an initial evaluation of unsupervised techniques that attempt to leverage knowledge extracted from a large existing corpus of pre-modern Chinese to improve OCR recognition accuracy on unseen historical documents.

Enhancement of Characters Recognizing for Ancient Chinese Texts with Bitmap Clustering and Contextual Analysis

Huang Chien-Kang

In this research, we are developing a technology to speed up the digitizing of ancient Chinese texts, especially those written in regular script or carved for printing. Generally, the modern Chinese OCR techniques can be applied to ancient Chinese texts, but the accuracy is much lower compared to modern printed texts. Therefore, we develop two different methods to enhance the OCR results.

First, we develop a better character segmentation algorithm, to get the separated bitmap of each character from ancient Chinese texts. The OCR-rate of these separated bitmaps is better than the original OCR-rate of the whole page. Second, in order to further enhance the accuracy, we cluster the bitmaps of the characters, and train an error-correction model based on the contextual patterns, which looks like a simplified language model. According to our preliminary result, the achieved accuracy rate ranges from 65% (first method) to 70% (second method).

Named Entity Recognition in Classical Chinese Text

Yuanquan Lu[2], Tao Jiang[2], Hou Ieong Ho[1], Shengfa Miao[1,*]

1. Leiden University
2. Beijing Institute of Technology

Named entity recognition (NER) is an important subtask of natural language processing, which aims to map elements in texts into pre-defined categories, such as names of locations, persons, titles, organizations, etc. NER is composed of two parts: entity recognition and entity typing. In this presentation, we apply two NER methods to classical Chinese texts: the pattern-based method and the conditional random fields method. We will go through the workflows of these two methods, illustrate and discuss challenges in each procedure, and evaluate their performances based on some basic measures such as precision and recall.

Bootstrapping Named-Entity Recognition in Difangzhi with CBDB and Semi-Automatically Labeled Data

Liu Chao-Lin

Note: This is a joint undertaking with Chih-Kai Huang of National Chengchi University, Taiwan

We report our experience of bootstrapping the work of recognizing person and location names in *difangzhi* with CBDB and imperfect data. Aiming to expand the contents of CBDB, we wish to identify person names and location names in *difangzhi*. Yet, there is no known syntactic parser for literary Chinese that was used in *difangzhi*. Hence, we tackle the recognition task with two different approaches of our own design. The first approach employs information in CBDB to annotate *difangzhi* texts to learn frequent patterns that are relevant to person and location names. We then take advantage of these patterns to find previously unknown named entities. The second approach uses the information of named entities that were identified by regular expressions as training data. We have not thoroughly verified this source of information yet. Although the information is imperfect, we still use them to train a CRF classifier to identify the named entities in the test data. We will present detailed steps and experimental results of these two approaches in the conference.

From Capital Triangle to Capital Corridor:
Mapping Social Elites in the Early Tang (600-700)

Xin Wen

Where did the social elites in the first century of the Tang rule reside? This issue can be addressed by comparing three geographically locatable and quantifiable datasets in the early Tang: 1. The more than three thousand prefects (*cishi*); 2. The seven hundred militias (*fubing*); 3. The owners of early Tang epitaphs. By mapping these datasets, I argue that the highest stratum of the social elites of the early Tang resided in a "triangle" between the three capitals: Chang'an, Luoyang, and Taiyuan. They lived and died in these regions and worked mostly in the central government; when they served at prefectural posts, they tended to be at prefectures in and around the capital triangle; this triangle also was the heart of the early Tang military presence, as more than 80 percent of the *fubing* militia existed there, providing both security and career opportunity for the capital elite. The transformation of this group of capital elite from the early Tang to the late Tang involved a geographical shift from this "capital triangle" to the "capital corridor" between Chang'an and Luoyang described by Nick Tackett. I argue that this geographical shift underlined a process in which members of the Tang elite from Taiyuan disappeared and dissolved into a new group that derived their preeminence not from association with the first generation of Tang founders, but from their role in the Tang central government.

Mapping the Chinese Novel

Margaret B. Wan

The study of traditional Chinese literature tends to focus on a well-known canon, neglecting the vast majority of extant works. Only around 10% of traditional Chinese novels have received scholarly attention in the West. My project aims to bring back the other 90%. By using the MARKUS tool to mark up large numbers of digitized novel texts with place data from the Temporal Gazetteer (TGAZ), I am creating an overview of space in the traditional Chinese novel. This research, inspired by the work of Franco Moretti, will contribute to our understanding of popular literature, history of the book, and regional culture. Genres or texts that are figuratively “off the map” or at the margins of canonical Chinese literature have much to tell us about important questions such as the system of thematic genres that constitutes the Chinese novel or the relationship between local, regional, and national in late imperial China. Some of these works have remained largely unexplored because of the sheer volume of texts one has to master in order to understand these genres. One must read, but to understand the context, one must also count and graph and map. Digital approaches and new digital tools promise to expand what one scholar can accomplish.

Towards a Comparative History of Political Factions in Chinese History

Hilde De Weerdt /Brent Ho

Past and current research on factionalist politics tends to focus on a few select struggles at the Han, Tang, and Song courts and tends to explain their formation and clashes by referencing the social backgrounds of those involved. Factionalism tends to be seen as an extraordinary phenomenon and the moral rhetoric in which factional struggles were conducted (or at least described in contemporary chronicles) was allegedly based on the assumption of the illegitimacy of their existence. However, the existing record shows that networking was an essential feature of the careers of late imperial literati. In order to sit for and pass examinations, to be appointed and re-appointed, scholars and officials kept track of personnel moves at court and networked with each other. Factional politics were common. I will present our attempts to re-invent Chinese political history by focusing on two questions.

First, we will examine when and how the faction lists (*dangbei*, *dangji*) of the Northern and Southern Song periods came into being. Some researchers have asked why literati from different persuasions were listed together; others have argued that early lists were the product of later historiography. Through network analyses of the co-occurrence of the dozens (and in one case hundreds) of names mentioned on faction lists in thousands of Song letters we have developed a set of methods that allows us to examine these questions in the aggregate. We can thus complement and move beyond the select case studies that have been conducted to address them.

Second, and more importantly, I will report on ongoing research to use the same method to investigate how factional politics operated outside the court in the provinces. We will discuss how we examine the bases for support and opposition to central political figures through the analysis of a relatively large number of notebooks and letters dating from the eleventh through the thirteenth centuries, tracing over time which politicians are discussed in these bodies of literature by which clusters of informants and with what level of approval or disapproval.

In addition to preliminary findings, I will also discuss methodological questions relating to the semi-automated tagging of different types of named entities, the analysis of sentiment, and social network analysis.

Reconstructing the Social Networks of Monks in Medieval China

Marcus Bingenheimer

Hidden in the Buddhist biographical literature on eminent monks is a large amount of detailed information about who knew whom. It is especially rich for the time between 300 and 1000, when the four major *Gaoseng Zhuan* collections allow us to situate people in place and time in great detail and trace their relationships to a degree unimaginable for Europe or India in that period.

Using data from the *Gaoseng Zhuan* project at Dharma Drum Mountain, this presentation introduces the social network drawn from the *Gaoseng Zhuan* corpus and what we can hope to learn from it. The network contains c. 6500 vertices with c. 13000 edges. After applying the proposed filters, the main network component contains c. 3500 vertices with c. 10000 edges. Based on visualizations done in Gephi we will try to show the outlines of what we know about the network, both in technical terms and in its consequence for historical research. Special attention will be paid to possible definitions of what counts as "importance" in this field of known relations. It is suggested that network analysis will allow us to see hitherto undetected patterns of influence and "importance" that generate a whole range of new research questions.

Three case studies surrounding important figures from the 4th to the 6th century show how network visualization can be used to aid historical inquiry and generate new research questions.

Temporal Comparison between Two Southern Song Notebooks

Chu Ping-tzu

My recent project is a comparative study of two Southern Song notebooks, the *Jianyan yilai chaoye zaji* 建炎以來朝野雜記 (Miscellaneous notes on the court and society since the Jianyan reign) by Li Xianchuan 李心傳 (1167-1244) and the *Sichao wenjian lu* 四朝聞見錄 (Record of things seen and heard over four reigns) by Ye Shaowong 葉紹翁 (1190?-). In my presentation, I will demonstrate how to use MARKUS (http://dh.chinese-empires.eu/beta/) and HuTime (http://www.hutime.jp) to compare temporal expressions between these two notebooks which generally cover the same period of time. By doing a topical and temporal comparison, we can examine the foci of these two literati to create a portrait of their time: the major events, the most important figures, and the most discussed time periods during the first four reigns of the Southern Song. In this talk, I will deal with the temporal part and examine how tools of the digital humanities can help in the comparison process. For the procedure, I will utilize MARKUS to tag the dates. Then I will reorganize the two texts by dates and make them into tables and feed the tables into HuTime to plot the timelines. Finally, I will make some observations in HuTime and conclude by demonstrating how these time expressions can help us understand the texts better.

Clustering Late Imperial Chinese Texts by Style:
Principal Component Analysis and t-SNE

Paul Vierthaler

As large corpora of late imperial Chinese texts become more readily available, they open up exciting new possibilities for digital research. They offer an opportunity to grasp large stylistic trends that are invisible at narrower levels of analysis. However, their number and highly variable content introduce computational and visualization difficulties. Fortunately, a variety of linear algebraic and machine-learning algorithms exist that facilitate these tasks. In this talk, I will compare and contrast several of these algorithms in the context of late imperial Chinese literature. In the first part of the talk, I will discuss the benefits and drawbacks of using PCA (Principal Component Analysis), a type of linear algebraic transformation of document-term matrixes, to analyze a variety of historical and quasi-historical texts. In the second portion, I will focus on analyzing these same documents with a related machine-learning technique called t-SNE (t-distributed Stochastic Neighbor Embedding). This technique may offer significant advantages over older clustering methodologies, while producing easy to understand, meaningful visualizations. I will finish by discussing the insights these algorithms offer into the nature of late imperial stylistics.

Identifying Long-term Trends in the Qing *Veritable Records*

Ian Miller

The Chinese archive boasts several astonishingly large and complete sets of state records. The Qing *Veritable Records* alone contains more than 300,000 individual reports covering two and a half centuries. These records potentially allow us to track changes in areas of interest to the Qing state over very long periods of time, and in an astonishingly close level of detail for a premodern corpus. This paper will discuss some attempts to use unsupervised and semi-supervised machine learning to extract information from the *Veritable Records* to use in analyzing the causes and consequences of long-term change.

This paper begins with my earlier work using the Latent Dirichlet Algorithm (LDA) topic model - an unsupervised model - to approximate long-term trends in the topics discussed in the Qing *Veritable Records*. Based in the LDA topic model, I am able to identify three types of trends: topics with small variation around a baseline (such as “crime”); secular changes in topic composition (including the instance of Han and non-Han name); and topics dominated by extreme events (including “rebellion”).

For this early work, I make up for errors in topic assignment at the document level by working with monthly sums of topic proportions. In an ongoing collaboration with Carol Shiue of the University of Colorado, we are using semi-supervised classifier training to improve the accuracy of document-level topic identification in order to enable more rigorous statistical analysis. Better topic identification will add granularity to trend-identification. This will enable regression of document topics (such as “rebellion”) against other long-term data sets, including price data, to help identify the precursors to large-scale historical events and factors that correlate with secular change.

The application of digital humanities in the stylometric analysis of the *Book of Odes* (***Shijing*** 詩經)

Louis Shueh-Ying LIAO

Recitation has been a cultural practice to tokenize verses in order to craft poetic language. My research is focused on the effect of such practice in *The Book of Odes*'s readings at the level of the hemistich. The question is: How do we handle literary rhythms which could be engendered during the reading of the corpus, and how do we establish a non-semantic connection between patterns?

In order to trace rhythmic phenomena in the text, I propose to focus on repetitious elements like syntactic formula and other frequent combinations of sinograms. A full analysis allows revealing dynamic patterns in a targeted string by contrast of occurrence.

I thus propose an experimental method of reading which sets forth two aspects: on the one hand the frequent pattern that creates the dominant rhythm, and on the other hand the rare pattern that constitutes the mean ideas.

On the basis of such an inquiry, we are able to evaluate the language style distance between authors, anthologies, dynasties, regions… etc. This approach also makes possible a new perspective to review some enigmatic literary phenomena, such as the ambiguous connection between verses produced with the " incentive process (*xing* 興)": a fundamental process of classical literary theory.

The Rise of the *Letterlet*:
Generic Analysis of Northern Song Epistolary Literature

Liu Chen

Although personal letters, what I call letterlets (*jian* 簡 or *chidu* 尺牘) in this project, had been written long before, it was during the Song Dynasty (960-1279) that they became a genre, signaled by their regular anthologization in literary collections. Existing scholarship has different interpretations of this phenomenon, especially regarding whether the *letterlet* was stylistically distinct from *shu* 書 letter, a long-established genre in Chinese history. In this project, I use Python to compare letterlets with both contemporary *shu* letters as well as pre-Song epistolary literature, focusing on the works of Su Shi 蘇軾 (1037-1101), the leading literatus of his day. By incorporating close reading with a digital analysis of topics, word frequency and N-grams, I attempt to demonstrate that the "birth" of the *letterlet* was a complicated process involving factors such as the circulation of letters and the development of the aesthetics of the spontaneous. This also serves as a case study for how to use digital tools for generic studies of Chinese literature.

The Reconstruction of Historical Data in Korea with Historical Keywords

Kim Baro

The emergence of the digitization of historical records in Korea dates back to 1967, when Harvard Professor Edward Willet Wagner (1924-2001) initiated the Munkwa Project. In 1995, the Korean government completed the “Digital Annals of Joseon Dynasty” project successfully, serving as one of the most significant achievements from the government-initiated promotion known as the “Public Service Project.” Based on these previous accomplishments, the National Institute of Korean History constructed the “Korean History On-Line” system in 2000, which is the first integrated system for Korean historical sources.

Current approaches on the digitalization of Korean sources mostly focus on building different digital databases of raw texts, often including scanned images of the original text, digitalized text (classical Chinese), and supplementing the text with English translation. However, connecting the raw texts with keyword-tagged metadata is another aspect which deserves more attention from Korean scholars. Accordingly, the methodology of historical data processing has remained stagnant for decades, as seen in the examples where individual research projects only add XML markers on historical keywords (time, place, person and other metadata) to the text. There have hardly been any attempts to build a database that enables a more systematic management of these metadata, such as assigning universal ID attributes to the keywords for cross reference and further application beyond the individual databases.

This presentation examines and introduces my work-in-process, proposed prototype for the construction of a more advanced and integrated historical keyword system for reconstruction Korean historical data.

Intellectual History and Computing:
Modeling and Simulating the World of the Korean Yangban

Javier Cha

The intellectual history of early modern Korea is defined by the coalescence of four major schools of Neo-Confucian thought and a number of literary trends. These developments took place at a time of increasing localization of population, material resources, state institutions, and what this paper will foreground: intellectual culture. The connections between the material and ideational aspects of the *yangban* aristocracy have been unclear, owing in large part to the reliance on case studies and the exclusive attention given to a small number of personalities, sources, and locations. To address this shortcoming, this talk presents some quantitative analysis of structured data as well as semi-supervised and predictive interpretation of a large corpus on the basis of diction, style, and figures of speech. The long-term objective is to create a simulation model for the social environment and textual world of the *yangban* in early modern Korea. The pilot run draws from three data sources: (1) the roster of 5,000 civil service examination degree holders; (2) 12,000 nodes representing their extended kinship network; and (3) an estimated 7 million characters of prose extracted from 200 collected works.

A Digital Research Platform for Studying Chinese Buddhist Literature

Hung Jen-jou

In the process of spreading to China, Buddhism gave rise to a vast enterprise for the translation of Buddhist scriptures. These texts later became the main source of information used by China's great patriarchs and masters in their study of Buddhism. Today, a large number of Buddhist texts have been collected, becoming a precious resource for researchers, but also creating new puzzles. In order to decipher these puzzles, researchers have used the traditional philological tools, reading and comparing different text. But these traditional methods are highly labor-intensive. The digital age that we have now entered has provided us with tools which can help us in conducting surveys of Buddhist texts on a scale larger than ever before, thus allowing us to overcome one of the main obstacles to further progress in the field. It is with this goal in mind that our team has been making use of these new tools of the digital age to create a digital research environment tailored to the needs of research in the Humanities. This research environment will provide high-quality digital content, combining relevant reference material with the latest research findings, thus becoming an integrated and content-rich Buddhist digital platform. Additionally, we will combine tools for quantitative analysis with tools for the editing and annotating of texts with the ultimate goal of creating a digital research platform which will assist scholars in their study of Chinese Buddhist texts.

Kanseki Repository: An Online Text Archive for Research and Collaboration

Christian Wittern

The Kanseki Repository has been developed by a research group at the Institute for Research in Humanities, Kyoto University, under the leadership of Christian Wittern. It features a large compilation of premodern Chinese texts collected and curated using firm philological principles based on more than 20 years of experience with digital texts. Among its unique features is the fact that the texts can be accessed, edited, annotated and shared not only through a website, but also through a specialized text editor, which thus morphs into a powerful workspace for the reading, research and translation of Chinese texts. The Kanseki Repository includes all texts in the *Daozang* and *Daozang jiyao* and a large collection of Buddhist material, including all texts created by the CBETA team, where applicable enhanced through the inclusion of recensions from the *Tripitaka Koreana*.

Following the first public release of the Kanseki Repository in March 2016, this presentation will introduce the methodological framework, contents and some use cases of the repository. Specifically, it will outline the overall architecture, which consists of a publicly accessible repository of the texts and applications that access these texts. Two of such applications will be introduced, the website www.kanripo.org and the Emacs module Mandoku. These are intended also as examples that show how the architecture could be used for other projects.

Understanding the Databases of Premodern China : Harnessing the Potential of Textual Corpora as Digital Data Sources

Jeffrey R. Tharsen
The University of Chicago

The rise of digital media and digital toolkits which allow the application of computational methods to analyses of language, literature and history has led to a variety of innovative approaches to textual corpora. Recent large-scale digitization projects have provided us with a wealth of outstanding textual resources from premodern China but we are only just beginning to understand how best to approach them and how to design custom algorithms to exploit their underlying structures. In this presentation I provide examples of a few core types of premodern Chinese texts, including lexica and dictionaries, prognostication manuals, literary anthologies, commentaries and historiographies, and propose a few basic strategies for designing data architectures and computational tools to more fully exploit their untapped potential. As each text has its own unique ontology, these general strategies must be customized for each specific case, but when designed and applied with care, recent research has shown that computational approaches can lead to striking new insights into both the texts themselves and the intellectual environments in which they were composed and transmitted.

The East Asian Studies Macroscope:
Infrastructure for Collaborative Scholarship across Corpora and Institutions

Peter Broadwell

The East Asian Studies Macroscope (EASM) is a joint effort by faculty and staff from the UCLA Department of Asian Languages and Cultures, the UCLA Library, and the UCLA Center for Digital Humanities to build partnerships with institutions in East Asia with significant digitized text archives for the purpose of developing software tools and practices for advanced collaborative research using digital corpora. These efforts build on the field's notable successes in creating single-corpora digital collections and interfaces, seeking to develop technological infrastructure and methods that can work with multiple corpora held at different institutions.

This paper will review briefly the results of EASM pilot projects conducted with large digitized collections of poetry from the Tang Dynasty and Heian-period Japan. These examples highlight the key infrastructural elements of the proposed platform and their contributions to scholarship: 1) remote, authorized computational access to multiple large-scale corpora, especially those that cannot be shared in full due to their size and/or access restrictions; 2) support for analytical tools that operate across collections, such as multi-corpus topic modeling and network analysis; and 3) features for scholarly collaboration at all stages of the research process, enabling sharing and critiquing of experimental workflows, results, and visualizations.