Digital Scholarship Symposium 2019

(Re-) Mining Text:

From Traditional to Digital

文字(再)勘探: 由傳統到數位

Date: 19 March 2019 (Tuesday)
Time: 09:15—16:30
Venue:
Digital Scholarship Lab,
G/F, University Library,
The Chinese University of Hong Kong

Programme

Time	Programme	Speaker(s)
09:00–09:15	**Registration**	
09:15–09:25	**Opening Ceremony**	
Keynote Speech (Moderator: Dr. Maria L.C. LAU)		
09:25–10:15	HathiTrust Research Center: Creating New Opportunities in Support of Scholarly Text Mining	Prof. Stephen DOWNIE Professor and Associate Dean for Research School of Information Sciences University of Illinois at Urbana-Champaign
10:15–10:40	**Coffee Break**	
Panel 1: English Text and Social Media (Moderator: Prof. YE Jia, Michelle)		
10:40–12:00	Automatic Text Classification – Models, Applications, and Recent Trends	Prof. LAM Wai Professor Department of Systems Engineering and Engineering Management The Chinese University of Hong Kong
	Bilingual Text-mining for Inspiring Personalized Reflection	Prof. KONG Siu Cheung Professor Department of Mathematics and Information Technology The Education University of Hong Kong
	Text Mining for Communication Measures on Social Media	Prof. LIANG Hai Assistant Professor School of Journalism and Communication The Chinese University of Hong Kong
	Discussion	
12:00–13:30	**Lunch**	

Programme

Time	Programme	Speaker(s)
Panel 2: Hong Kong Literature (Moderator: Prof. FAN Sin Piu)		
13:30–14:40	Network Theory, Plot Analysis: A Case Study of Lu Lun's "Poverty-Stricken Alley/Dead End"	Prof. WONG Nim Yan Associate Professor Department of Chinese Language and Literature The Chinese University of Hong Kong
	(Re-)Mining Text: An Experiment on Visualising Author Network in *An Annotated Bibliography of the Classical Writings of Hong Kong Poets*	Ms. Kitty K.Y. SIU Digital Scholarship Librarian The Chinese University of Hong Kong Library Ms. Daphne T.Y. SO Research Assistant The Chinese University of Hong Kong Library
	A Pilot Study on Topic Modeling of *The Chinese Student Weekly*	Dr. Wendy H.Y. WONG CLIR Postdoctoral Fellow The Chinese University of Hong Kong Library
	Discussion	
14:40–15:00	**Coffee Break**	
Panel 3: Chinese Text (Moderator: Dr. Maria L.C. LAU)		
15:00–16:20	Using CORPRO to Revisit the Authorship Controversy of *Dream of the Red Chamber*	Prof. CHUEH Ho-chia Associate Professor Department of Bio-Industry Communication and Development National Taiwan University
	Extracting Stylistic and Intertextual Markers from Chinese Text	Prof. Paul VIERTHALER Assistant Professor Centre for Digital Humanities Leiden University
	Unveiling the Sheng XuanHuai Archive @ CUHK – Experience Sharing on Using Corpus Data for Text Analysis and Beyond	Mr. Jeff LIU Associate Librarian Lingnan University Library Mr. Tony TSANG Assistant Computer Officer The Chinese University of Hong Kong Library Mr. SHENG Chang-Hung Descendant of SHENG XuanHuai
	Discussion	
16:20–16:30	**Closing Remarks**	

HathiTrust Research Center: Creating New Opportunities in Support of Scholarly Text Mining

Prof. Stephen DOWNIE

Professor and Associate Dean for Research
School of Information Sciences
University of Illinois at Urbana-Champaign

Abstract

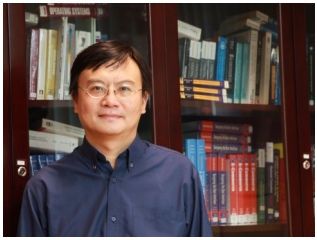
The HathiTrust Research Center (HTRC) is the research arm of the HathiTrust. As of March 2019, the HathiTrust Digital Library contains 16.9 million volumes (some 5.9 billion scanned pages). HTRC's mission is to provide "non-consumptive research" access to the HathiTrust collection. The non-consumptive research model is one where researchers can conduct text mining operations against the items found a given collection but cannot copy, read or redistribute the copyright-restricted materials contained within. Because scholars and students can have different levels of text mining experience, the HTRC has developed a suite of tools and services to support their various research capabilities and needs. This talk will highlight the technological, content, legal, social and human factors that shape HTRC's services and guide its staff. On the technological side, we will highlight some of HTRC's text mining and analysis tools including the Data Capsule virtual computing environment and the Bookworm trend analysis tool. On the human side, we will discuss the HTRC's Advance Collaborative Support (ACS) and other outreach programs including our "Digging Deeper, Reaching Further" train-the-trainer project. We will also introduce the Extracted Feature (EF) datasets which provide users with the freedom of "open data" while still respecting the HTRC's non-consumptive imperatives. The presentation will show how the EF datasets are being used by researchers and suggest some possible future development directions.

Biography

J. Stephen Downie is the Associate Dean for Research and a Professor at the School of Information Sciences, University of Illinois at Urbana-Champaign. Dr. Downie is the Illinois Co-Director of the HathiTrust Research Center (HTRC). Downie is the leader of the Hathitrust + Bookworm (HT+BW) text analysis project that is creating tools to visualize the evolution of term usage over time. Professor Downie represents the HTRC on the NOVEL(TM) text mining project and the Single Interface for Music Score Searching and Analysis (SIMSSA) project, both funded by the SSHRC Partnership Grant programme. Professor Downie was also the Principal Investigator on the Workset Creation for Scholarly Analysis + Data Capsules (WCSA+DC) project, funded by the Andrew W. Mellon Foundation. All of these aforementioned projects share a common thread of striving to provide large-scale analytic access to copyright-restricted cultural data. Stephen has been very active in the establishment of the Music Information Retrieval (MIR) community through his ongoing work with the International Society for Music Information Retrieval (ISMIR) conferences. He was ISMIR's founding President and now serves on the ISMIR board. In the recent past, Professor Downie worked with Dunhuang Academy on the "Digital Dunhuang" project to help connect Digital Humanities scholars with the high-resolution digital materials capturing the Mogao Caves. Professor Downie holds a BA (Music Theory and Composition) along with a Master's and a PhD in Library and Information Science, all earned at the University of Western Ontario, London, Canada.

Panel 1: English Text and Social Media

Automatic Text Classification – Models, Applications, and Recent Trends



Prof. LAM Wai

Professor
Department of Systems Engineering
and Engineering Management
The Chinese University of Hong Kong

Abstract

Text classification aims at assigning category tags or labels to text documents based on the content. Text documents typically contain unstructured text content often in natural language form. The tags come from a set of pre-defined categories. Text classification is an important part of text analysis or natural language processing. It can be employed whenever there are some tags or labels to map to some pieces of texts. Consequently it has a wide range of applications such as topic labeling, intent detection, sentiment analysis, social media monitoring, etc.

In this talk, I will discuss the general settings of automatic text classification which attempts to discover text classification models based on historical data or past experience. Such settings are also related to a popular topic in artificial intelligence (AI) known as machine learning. There are some specific representations for text data, which can be manipulated by computer. I will also present some basic concepts and overview of some techniques of text classification model learning. Some recent trends of automatic text classification are also mentioned.

Biography

Wai Lam received a Ph.D. in Computer Science from the University of Waterloo. He is currently a professor at the Department of Systems Engineering and Engineering Management in the Chinese University of Hong Kong (CUHK). He has published extensively in the areas of information retrieval, text mining, and natural language processing with h-index of 41 as at 2018. He has served as Area Chairs or Senior PC members in various prestigious conferences such as ACM Special Interest Group on Information Retrieval (SIGIR), Association for the Advancement of Artificial Intelligence (AAAI), ACM International Conference on Wed Search and Data Mining (WSDM), Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), etc.

His research projects have been funded by the Hong Kong SAR Government General Research Fund and Defense Advanced Research Projects Agency (DARPA, USA). He also managed industrial projects funded by Innovation and Technology Fund (industrial grant) and IT companies.

Bilingual Text-mining for Inspiring Personalized Reflection

Prof. KONG Siu Cheung

Professor
Department of Mathematics and Information Technology
Centre for Learning, Teaching and Technology
The Education University of Hong Kong

Abstract

The vision of education in the 21st century is to prepare learners for domain knowledge development through an active, constructive and interactive learning process in technology-enhanced environments. There is an increasing trend of encouraging learners to participate in discussion forums in Learning Management Systems (LMSs) for active, constructive and interactive learning. This creates new demands on teachers for a meaningful interpretation of the large amounts of learner-generated data in the LMSs to transform pedagogical practices as well as inspire reflective engagement. This presentation shares the experience of EdUHK in developing and implementing a bilingual text-mining system in course delivery for addressing the abovementioned teaching demands in higher education. The system adopts the bilingual text-mining technique for analyzing learner-generated text on Moodle. It incorporates a bilingual taxonomy of domain-specific keywords for an automatic identification and counting of matching keywords in learner-generated text, and provides hierarchical visualization for an informative and interactive display of counting results of matching keywords. Each learner can access the hierarchical visualization for an interactive comparison of his/her own usage of matching keywords with the whole-class usage. The system is evaluated to effectively support an efficient and evident tracking of learners' conceptual change in their course learning. Learners are inspired by the bilingual text-mining results to make concrete reflection on their personalized progress in learning course concepts. This data-oriented pedagogical initiative benefits teachers' pedagogical decision-making and learners' reflective learning engagement.

Panel 1: English Text and Social Media

Biography

Prof. KONG Siu Cheung is currently Professor of the Department of Mathematics and Information Technology, as well as Director of Centre for Learning, Teaching and Technology of the Education University of Hong Kong. He has over 100 academic publications in the areas of pedagogy in the digital classroom, policy on technology-transformed education, IT in mathematics education, programming for computational thinking development, and computational thinking education. He is currently the Editor-in-Chief of the international journal *Research and Practice in Technology Enhanced Learning* and *Journal of Computers in Education*. Prof. Kong was in the presidential roles for the Asia-Pacific Society for Computers in Education (APSCE) for six years, as the President-Elect in 2012 and 2013, the President in 2014 and 2015, and Past-President in 2016 and 2017.

Text Mining for Communication Measures on Social Media

Prof. LIANG Hai

Assistant Professor
School of Journalism and Communication
The Chinese University of Hong Kong

Abstract

Text-mining techniques have been demonstrated useful in measuring communication concepts and testing social science theories. This presentation will show examples of how text mining and social media data can facilitate the measurement of several key concepts in human communications, which were difficult in the past. First, this presentation will show how the measurements of political common ground and incivility can be incorporated into political communication studies. And then, the presentation will examine the roles of content similarity and redundancy in online conversations and information diffusion.

Biography

Professor Liang is an Assistant Professor in the School of Journalism and Communication at the Chinese University of Hong Kong. His research interests include computational social science, political communication, and public health. Currently, he is working on several interdisciplinary projects at the intersection of computational social science (analytical approach) and social media studies (data source). He has published numerous articles in the top communication journals such as *Journal of Communication*, *Communication Research*, *Human Communication Research*, *Journal of Computer-Mediated Communication*, and *New Media & Society*.

Network Theory, Plot Analysis: A Case Study of Lu Lun's "Poverty-Stricken Alley/Dead End"

Prof. WONG Nim Yan

Associate Professor
Department of Chinese Language and Literature
The Chinese University of Hong Kong

Abstract

'Network Theory, Plot Analysis' was an attempt to sketch out some hypotheses for quantitative analysis on plot: space and time, network regions, central characters, and periphery. The paper engages the network theory to Hong Kong writer Lu Lun's novel "Poverty-Stricken Alley" (1948) and a recently discovered screenplay of the same title by the writer. The networks help us to understand the roles of four main characters fully by its basic form of visualization—vertices (or nodes) and edges—that the temporal flow of a dramatic plot can be turned into a set of two-dimensional signs and can be grasped at a single glance. The network theory brings orders into literary evidence and suggests new courses of analysis to this exemplary work of a pioneer writer in Hong Kong literature.

Biography

Wong Nim-yan is an associate professor in the Department of Chinese Language and Literature at the Chinese University of Hong Kong. Her research interests include Hong Kong literature, women literature, political discourse analysis, and archival study. Her work of "Late Style: Discourses on Three Hong Kong Women Writers" (2007) won the recommended award of the tenth 'Hong Kong Biennial Awards for Chinese Literature (Literary Criticism)'. She has recently focused on discursive studies of Hong Kong literature in the 1990s and has edited various anthologies including "The Collection of Hong Kong Literature: Novels, 1942-1949" (2015).

(Re-)Mining Text: An Experiment on Visualising Author Network in *An Annotated Bibliography of the Classical Writings of Hong Kong Poets*

Ms. Kitty K.Y. SIU

Digital Scholarship Librarian
The Chinese University of Hong Kong Library

Ms. Daphne T.Y. SO

Research Assistant
The Chinese University of Hong Kong Library

Abstract

An Annotated Bibliography of the Classical Writings of Hong Kong Poets 《香港古典詩文集經眼錄》compiled by YW Chau (鄒穎文) in 2011 collects and annotates over 800 titles of classical Chinese writings published since 1842 by 514 Hong Kong authors. In this presentation, the Digital Scholarship Team of CUHK Library will demonstrate the use the valuable information from *The Bibliography* to experiment in building networks between the authors, their affiliated institutions, groups, places of origins, and other related information with Gephi, a software in creating network visualisation. A geographical display related to the authors with the use of GIS software for potential spatial analysis is also prepared. It is hoped that this pilot will demonstrate the use of scholar's research data with new tools to inspire new forms of research or digital scholarship.

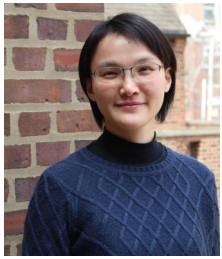
Panel 2: Hong Kong Literature

Biography

Kitty SIU is the Digital Scholarship Librarian of CUHK Library. She was a Geography graduate in CUHK with a Postgraduate Diploma in Applied Geoinformatics before acquiring Librarianship. Before entering the Librarian profession, she has been working for research and e-learning projects in Geography subject. After becoming a librarian, she has helped producing e-learning materials such as online video and website, etc. as a Reference Librarian. As the Digital Scholarship Librarian since July 2015, she has been providing digital scholarship service including workshops on digital scholarship research related tools, taking care the Digital Scholarship Lab as a space for researchers, working with Faculty members in conducting digital scholarship research by giving advice, providing assistance on software and tools for research data, and collaborating with Faculty members in research projects. Meanwhile, it is important to start working on Library materials and resources in giving inspirations to researchers in conducting digital scholarship research.

Daphne SO is a research assistant in the Digital Scholarship Team, CUHK library. Before joining the CUHK Library, she was a research assistant in the Academy of Visual Arts, Hong Kong Baptist University, and has presented her paper on Chinese material culture in Feng Chia University, Taiwan, during her undergraduate study in CUHK.

A Pilot Study on Topic Modeling of *The Chinese Student Weekly*



Dr. Wendy H.Y. WONG

CLIR Postdoctoral Fellow (Digital Scholarship)
The Chinese University of Hong Kong Library

Abstract

The Chinese Student Weekly 《中國學生周報》, a magazine published in Hong Kong between 1952 and 1974, was an influential magazine on the literati and students during its twenty-three years of publication and thereafter. The Chinese University of Hong Kong Library has digitized *The Weekly* in early 2000 with its full text available in the Library's Hong Kong Literature Database. With techniques on text mining, *The Weekly* could be investigated more comprehensively. This presentation aims to suggest some of the possible methods in text-mining *The Weekly* from the perspective of topic modeling. Topic modeling is a statistical model to investigate the topics that come about in a corpus. Analyzing publications in Chinese, particularly on those published in Hong Kong, using topic modeling is rare. Taking *The Chinese Student Weekly* as an example, this pilot study attempts to illustrate how topic modeling could be applied in Chinese-text corpus. With "distant reading" implied, this study targets at opening up new research areas on *The Weekly* and Hong Kong literature. While visualization based on the analysis are also exemplified, the demonstration sets as an example on how text mining techniques could be applied to other corpus available in the Hong Kong Literature Database and other datasets in the Chinese language.

Biography

Wendy Hoi Yan Wong is now working as the CLIR Postdoctoral Fellow (Digital Scholarship) at The Chinese University of Hong Kong (CUHK) Library. Wong holds PhD in Music, with focus on music theory, from CUHK. She was also awarded the MS in Library and Information Science (LIS) from the School of Information Sciences, University of Illinois at Urbana-Champaign. Before her study at Illinois, Wong worked as an adjunct lecturer at the Music Department, CUHK, and a research associate at the Special Collections unit of the CUHK Library. While she was pursuing her LIS degree, she worked as a graduate assistant for the HathiTrust Research Center and the Music and Performing Arts Library at Illinois. Currently at the CUHK Library, she is working on projects related to the Hong Kong Literature Database and the Rulan Chao Pian Collection (卞趙如蘭特藏).

Using CORPRO to Revisit the Authorship Controversy of *Dream of the Red Chamber*

Prof. Ho-chia CHUEH

Associate Professor
Department of Bio-Industry Communication and Development
National Taiwan University

Abstract

This presentation will introduce CORPRO, a Chinese language corpus-based social science software tool, and use *The Dream of Red Chamber* as a text corpus to demonstrate its possible implication. Given the advantage of using software for the textual data analysis in social sciences study, CORPRO provides an approach based upon corpus linguistics. CORPRO is designed to incorporate self-complied corpus and self-setup corpus analysis condition, including self-defined dictionary, stop-word, words-grouping. Corpus analysis functions include term frequencies, collocations, keywords in corpus and concordance, all with related statistics.

CORPRO provides humanities and social researchers (without computer science background) to conduct textual-mining independently. Its interface allows researchers to easily explore interesting features of corpus, and to repeatedly analyze corpus by setting different analysis conditions. That is, humanities and social science researchers can repeatedly look back and forth between observing characteristics of corpus and their subject domain knowledge. It is in this way to increase dialogues between corpus analysis and subject theory. This presentation will use the *Dream of Read Chamber* as a text corpus to illustrate possible contribution to its authorship controversy.

Biography

Ho-chia Chueh is currently Associate Professor at Department of Bio-Industry Communication and Development, National Taiwan University. Her research interests fall under four main areas: Rural Studies, Digital Humanities, Ethical issues of Agriculture, Philosophy of Education. Trained as an educational philosopher, Ho-chia's early work was focused on critical deconstruction of theories of identity through post-structuralist thinking. She published a book in this field, *Anxious Identity* from the Praeger Publishers. Later, she applied poststructuralist analysis to the field of rural studies, in particular examining how new agricultural concepts are represented in social texts. She is now focused on humanities approach to agricultural practices and rural activities. Ho-chia is also interested in digital humanities. She has developed a corpus analysis software CORPRO for humanities/social scientists to able to conduct text-mining analysis of a large amount of texts 'independently'. CORPRO is designed in the consideration of special contexts of Chinese language use.

Extracting Stylistic and Intertextual Markers from Chinese Text

Prof. Paul VIERTHALER

Assistant Professor
Centre for Digital Humanities
Leiden University

Abstract

Ever more expansive digital Chinese corpora and the development of increasingly sophisticated tools are now offering scholars of Chinese literature significant new avenues for research. In this talk, Paul Vierthaler will discuss the text-mining methods he uses on large corpora of late imperial Chinese documents to quantitatively extract stylistic signals and instances of intertextuality. In combination, these markers facilitate his research into late Ming and early Qing literary history and offer new insights on the impact extensive heteroglossic text reuse has on the style of late imperial vernacular fiction. He will begin his talk by quickly introducing the methods themselves; first, he will discuss a quantitative means of measuring style based on variable use of vocabulary, an approach known as stylometry. Then he will discuss his recent implementation of an algorithm based on the Basic Alignment Search Tool (originally developed by biologists) that allows him to extract instances of text reuse in very large corpora, even when modified by later authors. For the remainder of the talk, he will focus on how he uses these two methods to explore the nature of the *Plum in the Golden Vase*'s composition by extracting and then comparatively analyzing the style of the *Plum* against its sources.

Biography

Paul Vierthaler is an Assistant Professor of the Digital Humanities at Leiden University. In his current monograph project, he analyzes how historical events are represented in "quasi-histories" written in late imperial China. In this work, he studies how information transforms in genre- and time-dependent ways across thousands of semi- to un-trustworthy texts. Paul is interested in developing and adapting computational methods to analyze and visualize large natural language corpora. In his other work, he has been developing machine learning models to study the authorship of the famous late-Ming novel the *Plum in the Golden Vase*. Additionally, Paul is developing an extensible and mineable bibliographic database on public domain Chinese texts. Paul has held a Digital Humanities postdoctoral fellowship at Boston College and an An Wang postdoctoral fellowship at the Fairbank Center for Chinese Studies at Harvard University. He holds a Ph.D. in East Asian Languages and Literatures from Yale University.

Unveiling the Sheng XuanHuai Archive @ CUHK – Experience Sharing on Using Corpus Data for Text Analysis and Beyond

Mr. Jeff LIU

Associate Librarian
Lingnan University Library

Mr. Tony Tsang

Assistant Computer Officer
The Chinese University of Hong Kong Library

Mr. SHENG Chang-Hung

Descendant of SHENG Xuanhai

Abstract

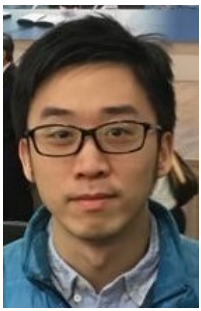
What can be done from the 4 million Chinese characters corpus data from over 13,000 correspondences housed in the "Sheng XuanHuai Archive@CUHK" (http://repository.lib.cuhk.edu.hk/en/collection/shengxuanhuai)?

This presentation will introduce the Archive and demonstrate how researchers can perform text analysis by using modern digital humanities tools that are freely available and visualize the result of selected corpus data. We will also share the experience on capitalizing the corpus by digital methods for some interesting cases and applications to demonstrate the potential of this Archive for the research community as well as promote this Archive into the general public.

Panel 3: Chinese Text

Biography

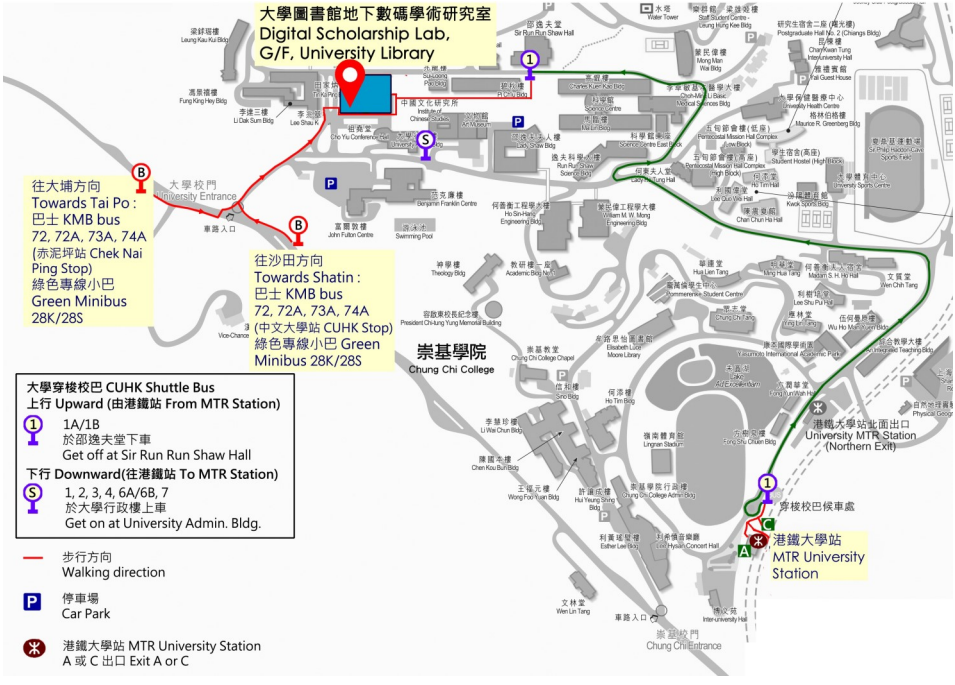
Jeff LIU is the Associate Librarian of Lingnan University Library in Hong Kong. He is responsible for overseeing the overall planning, development and practices of library services in learning and research support services and other user services in the digital era. Prior to joining Lingnan, he was the Digital Services Librarian of CUHK Library, responsible for the development of Library's digitization projects and development of the CUHK Digital Repository. His recent accomplishment is the building and development of the Sheng XuanHuai Archive with the collaboration from Art Museum of CUHK.



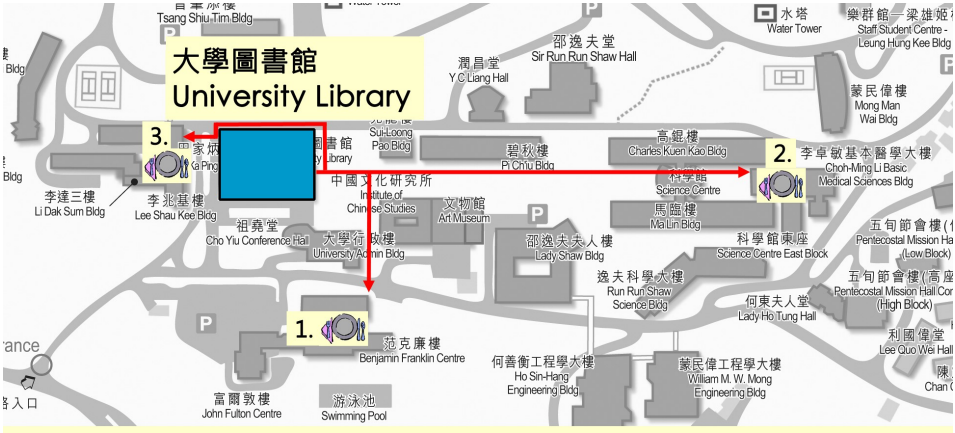
Tony TSANG is the Assistant Computer Officer of the Chinese University of Hong Kong Library. He handles all the technical issues of building of Sheng XuanHuai Archive and other collections in the CUHK Digital Repository (http://repository.lib.cuhk.edu.hk).

SHENG Chang-Hung is the great grandson of Sheng XuanHuai. He has great interest in reading the correspondence of his ancestor that has kept in good condition for more than 100 years ago. He is also a researcher and conservator of other Sheng's collection.

Transportation Directions



Nearby Lunch Venues



附近食肆 Lunch venues nearby:

1. 范克廉樓 Benjamin Franklin Centre:
   - 教職員餐廳 Staff Canteen (G/F)
   - 咖啡閣 Coffee Corner (G/F)
   - 學生膳堂 Student Canteen (G/F)
   - 素食餐廳 Vegetarian Food Shop (LG/F)
   - 女工合作社 Women Cooperative Store (LG/F)
2. 李卓敏基本醫學大樓小食店 Basic Medical Sciences Building Snack Bar
3. 李兆基樓咖啡室 Lee Shau Kee Building Coffee Shop (LG/F)

Hong Kong Literature Research Centre
http://hklrc.hk

The Chinese University of Hong Kong Library
http://www.lib.cuhk.edu.hk