Charting the European D-SEA:
Workshop Program & Abstracts

July 8-10, 2024

Berlin State Library, Potsdamer Str. 33

Note: Participants for workshops are expected to bring a laptop to follow along.

All workshops are 3.5 hours long with a built in 30-min coffee break.

Workshops: July 8-10

Day 1 (July 8)		
Room	***DBS***	***Simón-Bolívar-Saal***
Morning session: 9:00 - 12:30 ***(Coffee break: 10:30 - 11:00)***	Workshop 1: COMARKUS: Exporting data in context (Sander Molennar)	Workshop 2: Sinographic Historical Documents Automatic Transcription (Colin Brisson)
Lunch break 12:30-13:30		
Afternoon session: 13:30 - 17:00 ***(Coffee break: 15:00 - 15:30)***	Workshop 3: Creating a Data Model and Annotating Images in IMMARKUS (Sunkyu Lee)	Workshop 4: Leveraging Large-Scale Historical Databases with HistText (Chrstian Henriot & Cécile Armand)

Day 2 (July 9)		
Room	***DBS***	***Simón-Bolívar-Saal***
Morning session: 9:00 - 12:30 ***(Coffee break: 10:30 - 11:00)***	Workshop 5: Digital Editions of East Asian Sources (Duncan Paterson)	Workshop 6: Quantitative Analysis on Licensed Materials 6a: CrossAsia Ngram service and data mining (Hou Ieong Brent HO) 6b: Treating a Genre as a Knowledge System with LoGaRT (Shih-Pei Chen)
Lunch break 12:30-13:30		
Afternoon session: 13:30 - 17:00 ***(Coffee break: 15:00 - 15:30)***	Workshop 7: Algorithmic Identification and Elucidation of Intertextual Networks in Digital East Asian Corpora (Jeffrey Tharsen)	Workshop 8: Building a Database for Text-Aligned Reading with DocuSky (Hsieh-chang TU)

Day 3 (July 10)			
Room	***DBS***	***Simón-Bolívar-Saal***	***Rm. 320***

Morning **9:00 - 12:30** ***(Coffee break: 10:30 - 11:00)***	Workshop 9: A practical introduction to topic modelling (Christian Göbel)	Workshop 10: Social Network Analysis with Gephi: Theory and Practice (Song Chen)	Workshop 11: Utilising Prosopographic Databases for Sequence and Spatial Analysis (Thorben Pelzer)
Lunch break 12:30-13:30			
Afternoon **13:30 - 17:00** ***(Coffee break: 15:00 - 15:30)***	Workshop 12: Integrating ChatGPT into Humanities Research: An Introductory Workshop (Calvin Yeh & Jing XIANG)	Workshop 13: China Biographical Database Workshop: Advancing Chinese Studies through Database Development and Prosopography Research (Hongsu Wang & Peter Bol)	Workshop 14: Text and data mining with the Chinese Text Project (Donald Sturgeon)

Day 1, July 8, Morning (9:00 - 12:30)

Workshop 1. **COMARKUS: Exporting data in context**

Sander Molenaar (International Institute of Social History, Amsterdam)

X-MARKUS: Contextual Annotation (COMARKUS) is an annotation platform that facilitates the construction of ontological relations between entities. COMARKUS builds on X-MARKUS: Entity Annotation (ENTMARKUS), the annotation tool that allows users to tag and describe entities in texts. In COMARKUS, users establish and describe relations between entities through a schema that structures data annotated in ENTMARKUS. Entities are dragged from the text and dropped into data fields, and then saved as a cluster of related data. These structured data clusters allow users to interpret the context that is lost when entities are tagged in isolation and extracted from the text.

This workshop will instruct participants in the use of COMARKUS, illustrate how the schema that structures data can be modified to reflect the research designs of different projects, and illustrate how the data can be visualised and analysed.

COMARKUS was originally designed for the annotation of infrastructural events (the construction, renovation, failure, and destruction of city walls, bridges, and roads) and therefore comes with a default schema that describes an ontology of infrastructural events and that regulates how entities populate event fields. The workings of COMARKUS will be explained on the basis of data gathered within the projects 'The Lives and Afterlives of Imperial Infrastructure in Southeastern China' (InfraLives) and 'Regionalizing Infrastructures in Chinese History' (RegInfra), conceived and led by prof. dr. Hilde De Weerdt. In addition, the workshop will illustrate how the default schema can be modified and participants will be encouraged to develop a schema that reflects the ontological relations between entities in their own research projects. Finally, the workshop offers insight in the analysis and visualization of COMARKUS data clusters.

Prerequisites

It would be great if participants could bring their own primary source material, preferably annotated in MARKUS.

Workshop 2. **Sinographic Historical Documents Automatic Transcription**

Colin Brisson (Ecole pratique des hautes études, Paris)

This workshop aims to equip participants with foundational knowledge and practical skills necessary for the automatic transcription of Sinographic manuscripts and printed books.

The workshop will be conducted in two parts. We will begin with a brief overview of the objectives and the current state of the art in the field of automatic transcription of historical documents, highlighting how it diverges from traditional OCR techniques. Participants will then be guided through the critical steps involved and the data formats employed, through hands-on training using eScriptorium, an open-source annotation platform.

Subsequently, we will introduce the specialized tools developed specifically for processing Sinographic documents. We will guide you through using these tools to transcribe your own images.

Prerequisites:

No prior knowledge is required. Participants are encouraged to bring their own laptops. If desired, they can prepare a selection of images to test the models. The first part of the workshop will be conducted using an online instance of eScriptorium. For the second part will use Jupyter Notebooks, which can be run on your own machine or using Google Colab (a Google account is required).

Reference:

- Colin Brisson, Frédéric Constant, Marc Bui, Chinese Historical Documents Automatic Transcription (CHAT) models, https://github.com/colibrisson/CHAT_models.
- P.A. Stokes, Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, and El Hassane Gargem. “The eScriptorium VRE for Manuscript Cultures.” In: Classics@ 18 (1 2021).

Day 1, July 8, Afternoon (13:30 - 17:00)

Workshop 3. **Creating a Data Model and Annotating Images in IMMARKUS**

Sunkyu Lee (KU Leuven)

The Regionalizing Infrastructures in Chinese History (hereafter Reginfra) project has recently released IMMARKUS: Image Annotation in X-Markus. Different from other existing image annotation tools, including Mirador, the IMMARKUS allows users to design their own data model and to link your annotation to one or more external authority services, including TGAZ (China Historical GIS placename Database), CBDB (China Biographical Database Project), and DILA (Buddhist Studies Authority Database Project). In this workshop, participants will learn how to use IMMARKUS to create a customized data model, annotate images with different types of properties, add metadata, visualize relationships between different elements, and export annotations, data model and metadata. This session is intended for researchers, scholars and practitioners who may not have previous experience in image annotation but are interested in using digitized images for their research.

The tentative structure of the workshop is as follows:

- A 30-minute presentation about the basic interface of IMMARKUS, the concept of entities and properties in a hierarchical data model, possible input and output formats, etc.
- A 60-minute hands-on session on how to use IMMARKUS. Using sample images, participants will learn how to create or edit a data model, annotate multiple images, add notes, enter metadata information at both folder and image level, define relationships between different entities, create knowledge graphs, and export their annotations.
- A 60-minute session during which participants will have the opportunity to create their own data model, upload and annotate their own images, export their annotations, and analyze their annotations.
- The final 30-minute discussion session for questions and answers, and for participants to share their experiences of using IMMARKUS.

Prerequisites

Participants are required to bring their own laptop with a Google Chrome browser running. In addition, participants are strongly encouraged to prepare for a folder with more than three image files (.jpeg or .png files) that they would like to use in future research. During the second half of the workshop, participants will use their own image files to annotate and create a data model.

References

Platform:

Hilde De Weerdt, Rainer Simon, Lee Sunkyu, Iva Stojević, Meret Meister, and Xi Wangzhi. IMMARKUS: Image Annotation in X-MARKUS. 2024. immarkus.xmarkus.org

Instructions (wiki):

Hilde De Weerdt, Rainer Simon, Lee Sunkyu, and Iva Stojević. Image Annotation in IMMARKUS Wiki. 2024. github.com/rsimon/immarkus/wiki

Workshop 4. **Leveraging Large-Scale Historical Databases with HistText**

Christian Henriot, Cécile Armand (Aix-Marseille University)

This workshop introduces HistText, an application designed to leverage large-scale, multilingual, digitized corpora, with a particular emphasis on Chinese and English sources for studying modern China. Developed by the ENP-China team ("Elites, Networks, and Power in modern China") through a longstanding interdisciplinary collaboration between historians and computer scientists, HistText employs natural language processing and other computational techniques.

The application offers a range of functions enabling researchers to:

1. Build a corpus tailored to their research questions using advanced keyword search, multifaceted filters and queries, concordance, and word embedding.

2. Explore their corpus through diverse statistics and visualizations, such as word clouds and document frequency over time.

3. Extract, analyze, and visualize information such as names of persons, organizations, locations, and other named entities.

The workshop will be divided into two main parts:

1. In the first part, we will provide a brief overview of the genesis of HistText and demonstrate its functionality.

2. In the second phase, participants will have the opportunity to test the application with their own research cases.

A final wrap-up session will be dedicated to discussion and feedback.

Participants will have access to explore the corpora included in the Modern China Textbase, which currently comprises dozens of reference texts in Chinese and English, including the Chinese newspapers *Shenbao* and *Dongfang zazhi*, the ProQuest Chinese newspaper collection, student and economic journals, diaries, directories, who's who publications, archives, and Wikipedia.

HistText offers two main modes: (1) A beginner mode with a user-friendly R Shiny interface. (2) An expert mode in the form of an R package, utilizing R Studio. The workshop will focus on the interface and **requires no programming skills** from the participants. A demo of the expert mode can be provided upon request.

References

Paper

Blouin, Baptiste, Christian Henriot, and Cécile Armand. 2023. "HistText: An Application for Leveraging Large-Scale Historical Textbases." Journal of Data Mining and Digital Humanities. https://shs.hal.science/halshs-04178820

Manuals

Cécile Armand, Baptiste Blouin, and Christian Henriot, "HistText Interface: A User Guide", 13 novembre 2023. https://bookdown.enpchina.eu/Histtext/HistText_interface.html

Cécile Armand, Baptiste Blouin, and Christian Henriot, "HistText Manual", 15 novembre 2023. https://bookdown.enpchina.eu/HistText_Book/

Day 2, July 9, Morning (9:00 - 12:30)

Workshop 5. **Digital Editions of East Asian Sources**

Duncan Paterson (Berlin State Library)

[TEI](https://tei-c.org/) forms the backbone of diverse kinds of textual scholarship globally. Activity around East-Asian documents has intensified recently with the formation of the SIG Asia and its collaboration with the TEI technical council. This hands-on workshop for generating Digital Editions with [TEI-Publisher](https://teipublisher.com/exist/apps/tei-publisher-home/index.html) demonstrates means of enriching digital editions when reusing published documents, or when generating TEI as part of your own research.

Prerequisites:

Participants are required to download and install [docker](https://www.docker.com/) on their machines ahead of the workshop. Custom images will be shared during the workshop. Participants are encouraged to bring their own collection of TEI files to work on. A collection of sample files is available for those without their own TEI collection. GitHub accounts are recommended but not necessary.

Workshop 6a. **Leveraging CrossAsia N-gram Service for DH Research (90-mins)**

Hou Ieong Brent HO (Berlin State Library)

N-grams are an essential tool in linguistic research and digital humanities, enabling the analysis of language patterns, textual variations, terminology shifts, and more. However, obtaining meaningful N-gram data from commercial databases presents significant challenges, including technical difficulties and access restrictions due to licensing rights. CrossAsia's on-demand N-gram service aims to overcome these obstacles by providing researchers with accessible, high-quality N-gram data tailored to their specific needs. By offering a flexible and customizable solution, CrossAsia empowers researchers to make substantial advancements in their respective fields while adhering to licensing regulations.

This workshop is designed to introduce participants to the CrossAsia’s N-gram service and demonstrate how to leverage it for linguistic research and digital humanities projects. The workshop will be structured into three main parts:

1. **Introduction to CrossAsia N-gram Service and its Data Formats (15 mins)**
   We will begin with an overview of CrossAsia's N-gram service, explaining how the data is collected, how to access them and its data formats. This session provides the foundational understanding necessary for the hands-on activities that follow.
2. **Hands-on Session with Dataset Sample in Google Colab (45 mins)**
   Participants will engage in a hand-on session using Google Colab. Working with sample datasets, attendees will learn how to access and manipulate N-gram data, building confidence in using cloud-based tools for textual analysis.
3. **Introduction to Orange Data Mining Tool and Hands-on Session (45 mins)**
   The workshop will briefly introduce Orange, a local data mining tool written in Python. Participants will work with a sample dataset, covering basic textual analysis and visualization techniques using Orange.

Prerequisites:
Participants are required to have a Google Colab account (https://colab.google/) and to download and install Orange (https://orangedatamining.com/) on their local machines prior to the workshop.

Reference:

CrossAsia N-gram Service (https://crossasia.org/en/service/crossasia-lab/crossasia-n-gram-service/)

Workshop 6b. **Treating a Genre as a Knowledge System with LoGaRT (90 mins)**

Shih-Pei Chen (Max Planck Institute for the History of Science)

This workshop will introduce the Local Gazetteers Research Tools (LoGaRT), which currently hosts 4,410 titles of full-text digitized Chinese local gazetteers (difangzhi 地方志) published during late Imperial China and the Republican era. The 4410 titles are offered from two sources: 4000 are licensed materials from Zhongguo Fangzhi Ku (Beijing Erudition) and the license only covers CrossAsia users. The other 410 titles are digitized by MPIWG from Harvard Yenching Library's rare book collection and is open access. Users who sign up for a LoGaRT account from outside of MPIWG can immediately see the 410 open access titles. In the past year, we have also made the metadata of the Zhongguo Fangzhi Ku available to general users, which include the book metadata and the section headings of the entire collection within LoGaRT.

In this workshop, in addition to show the basic functions of LoGaRT, I will show how a general user can already make use of the metadata within LoGaRT to observe general patterns in this big collection. I will showcase how running a section search in this collection can help us understand the knowledge structure encompassed by this genre and how it changed over time.

Prerequisites:

Participants can sign up for a LoGaRT account prior to the workshop at this page: https://logart.mpiwg-berlin.mpg.de/LGServices2/#/signin. They can also watch the existing recorded tutorials online before coming to the workshop, but it's optional.

Online tutorials:

- https://content.mpiwg-berlin.mpg.de/mpiwg/online/permanent/Media_Online/Video/2020/2020-06-19_TALK_HarvardYenchinLoGaRT/2020-06-19_TALK_HarvardYenchinLoGaRT_SChen.mp4
- https://www.youtube.com/playlist?list=PLhOCf20UlVNtwaQwegrBfzOxSgWit4UZY

Day 2, July 9, Afternoon (13:30 - 17:00)

Workshop 7. **Algorithmic Identification and Elucidation of Intertextual Networks in Digital East Asian Corpora**

Jeffrey Tharsen (University of Chicago)

Intertextuality has been a significant concern of scholarly communities around the world for centuries; fields like *Redaktionsgeschichte* in Germany and *jiaokanxue* 校勘學 in China have long provided evidence-based foundations for debates on the relationships between works, editions and authors. With the advent of digital texts and computational tools, new avenues for research into intertextuality have recently emerged. To this end we developed TextPAIR, a language-agnostic open-source unsupervised approach to detecting "text reuse" in any language or script. TextPAIR enables new forms of algorithmically-based research into and large-scale network visualizations of relationships between textual communities, traditions and sources, detection of correspondences (from direct quotations to imperfect citations to allusions) across multiple languages and through various intellectual traditions, new ways to map the development of ideas and concepts (cooccurrences, direct and indirect) over the longue durée, and insights into the sources of many of our most classic works, long obscured by time, space and/or lack of prestige.

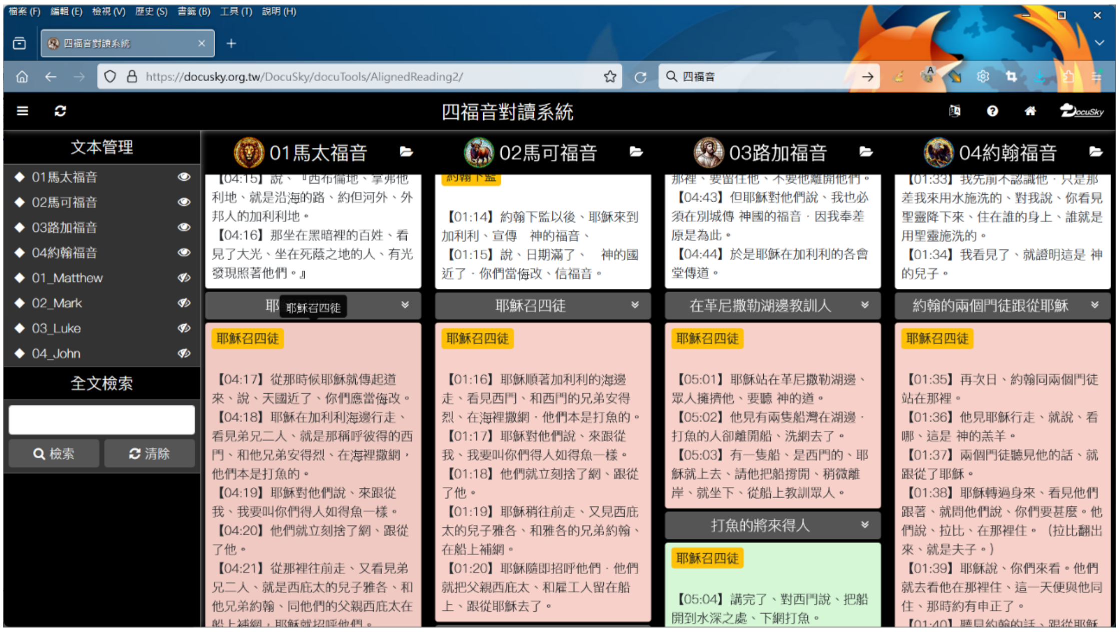
References:

https://textual-optics-lab.uchicago.edu/

Workshop 8. **Building a Database for Text-Aligned Reading with DocuSky**

Hsieh-Chang Tu (National Taiwan University)

DocuSky is a research platform developed by the NTU Research Center for Digital Humanities that enables users to build personal textual databases that support search and analysis over texts. The texts are organized as a set of documents. Each document contains metadata that describe properties of the text, and a document can contain markup to support fine-grained text analysis.



01馬太福音

【04:15】說、『西布倫地、拿弗他利地、就是沿海的路、約但河外、外邦人的加利利地。
【04:16】那坐在黑暗裡的百姓、看見了大光、坐在死蔭之地的人、有光發現照著他們。』

耶穌召四徒

耶穌召四徒

【04:17】從那時候耶穌就傳起道來、說、天國近了、你們應當悔改。
【04:18】耶穌在加利利海邊行走、看見弟兄二人、就是那稱呼彼得的西門、和他兄弟安得烈、在海裡撒網，他們本是打魚的。
【04:19】耶穌對他們說、來跟從我、我要叫你們得人如得魚一樣。
【04:20】他們就立刻捨了網、跟從了他。
【04:21】從那裡往前走、又看見弟兄二人、就是西庇太的兒子雅各、和他兄弟約翰、同他們的父親西庇太在船上補網，耶穌就招呼他們。

02馬可福音

約翰下監

【01:14】約翰下監以後、耶穌來到加利利、宣傳　神的福音、
【01:15】說、日期滿了、　神的國近了．你們當悔改、信福音。

耶穌召四徒

耶穌召四徒

【01:16】耶穌順著加利利的海邊走、看見西門、和西門的兄弟安得烈、在海裡撒網．他們本是打魚的。
【01:17】耶穌對他們說、來跟從我、我要叫你們得人如得魚一樣。
【01:18】他們就立刻捨了網、跟從了他。
【01:19】耶穌稍往前走、又見西庇太的兒子雅各、和雅各的兄弟約翰、在船上補網。
【01:20】耶穌隨即招呼他們．他們就把父親西庇太、和雇工人留在船上、跟從耶穌去了。

03路加福音

那裡、要留住他、不要他離開他們。
【04:43】但耶穌對他們說、我也必須在別城傳 神國的福音．因我奉差原是為此。
【04:44】於是耶穌在加利利的各會堂傳道。

在革尼撒勒湖邊教訓人

耶穌召四徒

【05:01】耶穌站在革尼撒勒湖邊、眾人擁擠他、要聽 神的道。
【05:02】他見有兩隻船灣在湖邊．打魚的人卻離開船、洗網去了。
【05:03】有一隻船、是西門的、耶穌就上去、請他把船撐開、稍微離岸、就坐下、從船上教訓眾人。

打魚的將來得人

耶穌召四徒

【05:04】講完了、對西門說、把船開到水深之處、下網打魚。

04約翰福音

【01:33】我先前不認識他．只是那差我來用水施洗的、對我說、你看見聖靈降下來、住在誰的身上、誰就是用聖靈施洗的。
【01:34】我看見了、就證明這是 神的兒子。

約翰的兩個門徒跟從耶穌

耶穌召四徒

【01:35】再次日、約翰同兩個門徒站在那裡。
【01:36】他見耶穌行走、就說、看哪、這是 神的羔羊。
【01:37】兩個門徒聽見他的話、就跟從了耶穌。
【01:38】耶穌轉過身來、看見他們跟著、就問他們說、你們要甚麼。他們說、拉比、在那裡住。（拉比翻出來、就是夫子。）
【01:39】耶穌說、你們來看。他們就去看他在那裡住、這一天便與他同住、那時約有申正了。
【01:40】聽見約翰的話、跟從耶穌

Figure 1. The Four Gospels database for text-aligned reading

In this workshop, we shall introduce DocuSky and illustrate its power by several use cases. In particular, we discuss an interesting application whose goal is to construct a DocuSky database for text-aligned reading. Figure 1 illustrates an example that aligns texts from The Four Gospels (四福音書). Each column displays documents from a gospel, and those with the same topic are aligned for comparative reading. The database of this example actually contains 8 sets of documents, 4 in Chinese and 4 in English. A user can use the left control panel to choose which sets to show in the right area.

During the practical session of this workshop, we will use the first three chapters of The Four Gospels as an example to practice, step by step, how to construct a DocuSky database for text-aligned reading. A participant is expected to know how to use a text editor (such as Notepad) and Excel. They should also know some Chinese for better understanding of the tool used to convert data from Excel to DocuSky. After the workshop, one should be able to construct a complete database functioning as the one shown in Figure 1.

References:

https://docusky.org.tw/DocuSky/docuTools/AlignedReading2/

Day 3, July 10, Morning (9:00 - 12:30)

Workshop 9. **A practical introduction to topic modelling**

Christian Göbel (University of Vienna)

How to quickly make sense of a body of text that is too large for a single human to read? This workshop provides a practical introduction to topic modeling, a form of text mining that uncovers hidden semantic structures ("topics") in a corpus of documents. Topic modelling is a form of unsupervised machine learning suitable for eliciting how prominent certain topics are in a corpus, how they are connected, and how they develop over time. With a bit of caution, researchers can also use topic modelling algorithms to classify documents and thereby make them amenable to statistical analysis.

The workshop consists of four parts. First, participants will receive a brief introduction into the use cases of topic modelling, the most commonly used algorithms, and their strengths and weaknesses. Second, participants will learn how to preprocess text for analysis (remove stop words, lemmatise words, segment Chinese language documents), select the hyperparameters of Latent Dirichlet Allocation (LDA) models and decide on an appropriate number of topics. In the third part, we will use the fitted model to classify documents, inspect a random sample of classified documents and discuss the accuracy of classification.

Finally, participants will learn how to visualise the prevalence, development and connection of topics as a bar chart, line diagram and correlation plot, respectively.

Workshop 10. **Social Network Analysis with Gephi: Theory and Practice**

Song Chen 陳松 (Bucknell University)

This workshop explores the use of social network analysis (SNA) in Chinese studies. Designed for novices, it weaves together theory and practice. The first half of the workshop introduces the basic concepts of SNA and its application to prosopographical research, guiding participants through the basics of the Gephi software. The second half delves into SNA as a graph-theoretical approach to data modeling and its applications beyond interpersonal relationships. During the second half, participants will also learn about various Gephi plug-ins that enable the analysis of two-mode networks and the integration of SNA with spatial analysis.

Prerequisites:

Participants are expected to download and install Gephi before the workshop using the following link: https://gephi.org/users/download/ Gephi is an open-access network visualization program, compatible with both Windows and Mac OS. It also provides the most common analytical utilities for network data. Participants are also expected to have Microsoft Excel (or a similar spreadsheet program) installed prior to the workshop.

Workshop 11. **Utilising Prosopographic Databases for Sequence and Spatial Analysis**

Thorben Pelzer (Leipzig University)

In this three-hours workshop, attendees will learn how to analyse large biographical datasets with a focus on uncovering spatial distributions (domestic and international mobilities) and identifying typical career patterns. As a case study, attendees will use exported data from the "Chinese Engineers Relational Database" (Pelzer et al. 2022) to untangle the prosopography of a Republican-era elite profession. Although the example relies on historical data, the methods may also be of interest to researchers of contemporary East Asia.

We will first go over the basic theory of sequence and spatial analysis. Then, we will discuss the fundamentals of structuring a relational database and how one should arrange and format data to make sure they can be used for the desired purpose. Then, we will run some example code using the "Simple Features" and "Life Trajectory Miner" packages for R. Finally, we will discuss some of the limits of quantitative data analysis to better understand the extent to which datasets can be used to augment or challenge our arguments.

Prerequisites:

Attendees are requested to bring a laptop. To allow for a swift start of the seminar, you are kindly asked to install the newest version of R and R-Studio (both available via https://posit.co/download/rstudio-desktop/) beforehand. No prior knowledge in coding is required.

Day 3, July 10, Afternoon (13:30 - 17:00)

Workshop 12. **Integrating ChatGPT into Humanities Research: An Introductory Workshop**

Calvin Yeh (Max Planck Institute for the History of Science) and Xiang Jing (University of Chinese Social Sciences)

This session will delve into both basic and advanced features of ChatGPT, tailored to both ChatGPT non-paying and paying users, emphasizing practical, exploratory learning. As AI continues to grow in importance, grasping its application within humanities research becomes crucial.

For non-paying users, we will showcase how to employ Microsoft Copilot with GPT-4 technology for text-based tasks. The workshop will address creating effective prompts, summarizing texts, and simulating group discussions. Techniques like prompt chaining and text data extraction will be introduced, providing foundational insights into how AI can support humanities research.

For paying users, the opportunity to explore ChatGPT's more advanced capabilities will be presented, though with an emphasis on introductory exploration. This includes data conversion, geo-coordinate extraction, and basic data analysis. We will also explore GPTs, focusing on their distinctive features: "Knowledge files" and "Actions," which enhance and customize the ChatGPT experience.

Designed as an introductory session, this workshop is ideal for researchers eager to learn how AI tools like ChatGPT can be integrated into their work. Participants will gain an understanding of both the potential and the limitations of AI in humanities research, preparing them for further exploration and application.

Prerequisites:

To maximize the benefits of this workshop, all participants are required to have either an active ChatGPT Plus subscription or a registered Microsoft Copilot account prior to attending. This will enable full access to the features and exercises planned during the session. Please ensure your subscription or registration is set up before the workshop date.

Workshop 13. **China Biographical Database Workshop: Advancing Chinese Studies through Database Development and Prosopography Research**

Hongsu Wang & Peter Bol (Harvard University)

The China Biographical Database (CBDB) is a freely accessible relational database with biographical information about approximately 535,181 individuals as of February 2024, primarily from the 7th through 19th centuries. With both online and offline versions, the data is meant to be useful for statistical, social network, and spatial analysis as well as serving as a kind of biographical reference.

This workshop will cover the foundational methodologies of the CBDB, demonstrate how to utilize it for prosopographical research, and highlight the latest advancements in our project. In recent years, CBDB has collaborated with numerous scholars to expand its data sources, ranging from early China to the Qing dynasty. Furthermore, CBDB has explored innovative data mining methodologies employing traditional machine learning models, large language models, and digital humanities tools like LoGaRT and Markus. These developments and their applications will be introduced during the workshop.

We will introduce examples of research utilizing the China Biographical Database based on several studies conducted in recent years. These cases will illustrate the diverse ways in which CBDB has been employed to advance research in Chinese studies."

These years, CBDB has fostered a vibrant open-source development community on GitHub. The array of tools and manuals available in the CBDB GitHub community might benefit various other projects. These include disambiguation tools, a practical tool for applying transformers to identify personal names, addresses, and official titles, the transformer model to separate author and book titles for unpunctuated Chinese, the tool to submit prompts to free large language model API by batch, etc.

Installation and download before the workshop:

- QGIS: https://www.qgis.org/en/site/forusers/download.html
- Gephi: https://gephi.org/users/download/
- Microsoft Access
- CBDB Access Database: https://www.dropbox.com/scl/fi/itcaauz35orn48qtsegp5/CBDB_bg_20240208_build20240411.7z?rlkey=us92t0xbjnwdp2qycccrgfilq&dl=0
- CHGIS V6 Layers: https://www.dropbox.com/scl/fi/fviyyklh9s7uf1qt6i1a9/GIS-materials.rar?rlkey=041ga2ngdfxqqpplixfir088b&dl=0
- Installation tutorials: https://www.youtube.com/playlist?list=PLGgLzyv7BMgY0fs-Un-UAieCrZPKG-Fp7

Workshop 14. **Text and data mining with the Chinese Text Project**

Donald Sturgeon (Durham University)

This hands-on workshop introduces participants to complete text and data mining workflows, from digital transcription and annotation of premodern works through to extraction of data derived from their contents, using materials and tools from the Chinese Text Project (https://ctext.org). It consists of four parts:

1. **Using the Chinese Text Project**: how to use this crowdsourced editing platform to create and obtain accurate, linked digital transcriptions of premodern Chinese texts.

2. **Interactive text mining**: extracting and visualizing statistical properties and relationships from transcribed texts. Types of analysis include pattern matching of words and phrases, identification of text reuse, and patterns of vocabulary usage; visualizations include summarization via networks, charts, and textual heatmaps.

3. **Annotating, disambiguating, and linking references to entities** (such as names of people, places, and eras) in a premodern text to authority databases, extracting knowledge claims about these entities (such as dates of birth, death, or appointment to a particular bureaucratic office) and contributing them to a crowdsourced knowledge base.

4. **Interactive data mining**: extracting and visualizing data from annotated texts and extracted knowledge claims. This includes simple querying of the knowledge base for particular types of information through the online interface, as well as the basics of the widely used SPARQL query language.

This workshop does not assume any prior background in digital methods, and requires only a web browser (recommended: Google Chrome or Firefox). Participants are encouraged to create a free account on ctext.org prior to the workshop: https://ctext.org/account.pl

Further information:

- https://ctext.org/dh
- Crowdsourcing the Historical Record: Creating Linked Open Data for Chinese History at Scale, *International Journal of Humanities and Arts Computing* 16:1, 2022.
- Chinese Text Project: a dynamic digital library of premodern Chinese, *Digital Scholarship in the Humanities*, 2019.

- Digital Approaches to Text Reuse in the Early Chinese Corpus, *Journal of Chinese Literature and Culture* 5:2, 2018.