ÖAW AUSTRIAN ACADEMY OF SCIENCES

IKGA—INSTITUTE FOR THE INTELLECTUAL AND CULTURAL HISTORY OF ASIA

European Research Council
Established by the European Commission

universität wien Institut für Südasien-, Tibet- und Buddhismuskunde

Advanced Computational Methods for Studying Buddhist Texts

International symposium organised by
Patrick McAllister, Rachael Griffiths, and Markus Viehbeck

(Version: 24th April 2023)

When 27th to 28th April 2023
Where SR 1, Institute for South Asian, Tibetan and Buddhist Studies
Spitalgasse 2, Court 2, Entrance 2.7
1090 Vienna, Austria
Zoom Meeting-ID: 636 8825 7180
Passcode: 684137
https://univienna.zoom.us/j/63688257180
Register Please contact patrick.mcallister@oeaw.ac.at to attend in person.
No registration is required for joining through Zoom.
Updates https://www.oeaw.ac.at/ikga/ -> Events

Map

Zoom

Updates

1 Schedule

27th April 2023

Chair:	*Pascale Hugon*	*Chair:*	*Dorji Wangchuk*
9:15–9:30	Welcome	**14:00–14:30**	Élie Roux
9:30–10:00	Nathan Hill and Julien Baley	**14:30-15:00**	Charles Li
		15:00–15:30	*Coffee break*
10:00–10:30	Marieke Meelen	**15:30–16:00**	Rachael Griffiths
10:30–11:00	*Coffee break*	**16:00–16:30**	Channa Li, Marco Peer and Kaifeng Yang
11:00–11:30	Oliver Hellwig		
11:30–12:00	Kiyonori Nagasaki		
12:00–14:00	*Lunch break*		

Keynote lecture
17:00–18:30

Marcus Bingenheimer:
"On the Use of Historical GIS in the Study of Chinese Buddhism"

19:00: Dinner (speakers invited)

28th April 2023

Chair:	*Rachael Griffiths*	*Chair:*	*Masahiro Shimoda*
9:30–10:00	Orna Almogi	**14:00–14:30**	Tyler Neill
10:00–10:30	Sebastian Nehrdich	**14:30–15:00**	Gérard Huet
10:30–11:00	*Coffee break*	**15:00–15:30**	Amba Kulkarni
11:00–11:30	Markus Viehbeck		— End Zoom —
11:30–12:00	Bruno Lainé	15:30–16:00	*Coffee break*
12:00–14:00	*Lunch break*	**16:00–17:00**	Group discussions

Location: SR 1, Institute for South Asian, Tibetan and Buddhist Studies, Spitalgasse 2, Court 2, Entrance 2.7, 1090 Vienna, Austria

2 International Symposium: *Advanced Computational Methods for Studying Buddhist Texts*

Computational Humanities is a rapidly growing multidisciplinary field that uses computational and quantitative methods for processing, analyzing, and modeling complex data. Within Buddhist Studies, these methods have emerged as an important tool for those working with Buddhist texts, enabling large-scale analysis, facilitating preservation and increased accessibility, and providing new ways of visualizing and understanding data.

The symposium "Advanced Computational Methods for Studying Buddhist Texts" will bring together scholars conducting research on Buddhist texts by computational methods ranging from natural language processing, optical character and handwriting recognition, geographic information systems, cross-linguistic alignment, to content analysis.

Alongside providing an opportunity for researchers to present their most recent work and share their experiences, the symposium aims to facilitate discussion on the challenges, opportunities, and further applications of advanced computational methods for the field of Buddhist Studies.

The symposium is organized by Patrick McAllister (Institute for the Cultural and Intellectual History of Asia IKGA, Austrian Academy of Sciences), Rachael Griffiths (ERC project *The Dawn of Tibetan Buddhist Scholasticism (11th-13th c.)* TibSchol, IKGA), and Markus Viehbeck (*Tibetan Manuscript Project Vienna* TMPV, University of Vienna).

3 Abstracts

3.1 Almogi, Orna

Title **BuddhaNexus: A Tool for Studying the History of Composition of Buddhist Texts**

Presenter Orna Almogi (Universität Hamburg)

Abstract BuddhaNexus (https://buddhanexus.net/) is an online tool particularly designed to facilitate the study of Buddhist texts in the four Buddhist classical languages of Pāli, Sanskrit, Tibetan, and (Buddhist) Chinese. The database supports the study of various research areas and helps answer research questions of various kinds, such as the history of composition of a certain text or its transmission and impact. In my talk I shall discuss how the database can be used for the study of the history of composition of any given text, or group of texts, and provide some examples from the Tibetan corpus. While presenting the database's strengths in this regard, I shall also point out its current shortcomings, and attempt to offer some solutions as well.

3.2 Griffiths, Rachael

Title **Towards transcribing handwritten Tibetan manuscripts with HTR**

Presenter Rachael Griffiths (IKGA, ERC project *The Dawn of Tibetan Buddhist Scholasticism (11th-13th c.)*, TibSchol)

Abstract The use of advanced computational methods for the analysis of digitized texts is becoming increasingly important in the humanities and social sciences. This is also the case in Tibetan and Buddhist studies. As part of the ERC-funded project "The Dawn of Tibetan Buddhist Scholasticism (11th-13th c.)" (TibSchol, 101001002), conducted at the Institute for Cultural and Intellectual History of Asia (IKGA), Transkribus is being used to train HTR model(s) on Tibetan cursive handwritten texts in order to facilitate the analysis of a large corpus of unedited manuscripts from the 10^{th}-15^{th} centuries. In this talk, I will present the project's approach to HTR and some of the challenges encountered, and discuss our initial results.

3.3 Hellwig, Oliver

Title **Dating text corpora with Bayesian models**

Presenter Oliver Hellwig (University of Zürich; University of Düsseldorf)

Abstract In spite of over 150 years of research the chronology of the Vedic literature is still not fully understood. This presentation introduces a quantitative model that aims at deriving such a chronology from linguistic features (https://chronbmm.phil.hhu.de/). It focuses on the development of syntactic markers and the computational framework required for their collection. In addition, it will briefly touch upon the question of how such a framework can be used for studying the history of Buddhist literature.

3.4 Hill, Nathan and Julien Baley

Title **Graph theory approaches to Buddhist transcriptional Chinese**

Presenters Nathan Hill (Trinity College Dublin), Julien Baley (SOAS University of London)

Abstract In 1983 W. S. Coblin published *A handbook of Eastern Han sound glosses*. This book contains a collection of Indic terms from early Chinese Buddhist translations and attempts to use them to determine how Chinese was pronounced at the time of each translator. Progress in Buddhist philology, in Chinese historical phonology, and in digital humanities methods all call for this work to be done afresh.

This paper provides an update on an ongoing project to systematically reexamine Han dynasty Chinese transcriptions of Indic Buddhist texts. Our novel contributions include the incorporation of texts translated by An Shigao 安世高 (circa 148-180 CE) rediscovered in 1999 in the Kongōji 金剛寺 temple in Osaka and a full use of comparisons to Gāndhārī, the suspected source language for most early translations.

To model Indic transcriptions within Buddhist texts we deploy bipartite networks, with two types of both nodes and edges. The nodes will be (1) Indic syllables (bud, dha, etc.) and (2) Chinese characters (佛, 陀, etc.) and the edges will respectively link (1) an Indic syllable to the character that transcribes it (bud→佛, dha→ 陀) and (2) a syllable in a word with the syllable

that follows it (佛⇒陀, bud⇒dha). This method will improve upon those approaches to Buddhist transcription that ignore the place of syllables inside of words (Yu 1999[1979]), a highly relevant context for sound changes in the Middle Indic varieties that provided Chinese with its Buddhist vocabulary.

3.5 Huet, Gérard

Title **Hypertext grammatical tools for Sanskrit digital libraries: the Sanskrit Heritage experience**

Presenter Huet, Gérard (Emeritus, Paris Inria Center)

Abstract We shall briefly describe and demonstrate the facilities offered by the Sanskrit Heritage suite of tools for managing classical Sanskrit corpus in grammatically informed ways (`https://sanskrit.inria.fr/index.en.html`). The talk will discuss shortcomings of the methods and their possible remedy, as well as the possible future development of digital libraries consistent with such facilities.

3.6 Kulkarni, Amba

Title **START: Sanskrit Text Annotation and Research Tool**

Presenter Amba Kulkarni (Department of Sanskrit Studies, University of Hyderabad)

Abstract In this talk we demonstrate 'START - Sanskrit Text Annotation and Research Tool,' a tool being developed at the University of Hyderabad. The back end of START is a Sanskrit Computational Linguistics platform which integrates the distributed services of two platforms viz. *The Sanskrit Heritage* platform (https://sanskrit.inria.fr) and the *Saṁsādhanī* platform (see https://sanskrit.uohyd.ac.in/scl).

Given a Sanskrit verse or a Sanskrit sentence, START provides a complete analysis: segmentation, morphological analysis (both inflectional and derivational, linking the nominal and verbal stems to the various monolingual and bilingual dictionaries), as well as a sentential analysis providing the *śābdabodha* graphically. The user interface of START also allows the user

to edit the annotation. The edited annotation for a given text can be compiled automatically into useful formats, such as for E-readers.

Such semi-automatically annotated texts, perhaps after further manual editing, would contribute to Gold Standard corpora. Being in a uniform format, without any manual errors, these corpora are useful for training machines as well as for researchers studying linguistic divergences, stylistic variations, etc.

3.7 Lainé, Bruno

Title **Optimisation of e-texts for an efficient quotation search in the rKTs database**

Presenter Bruno Lainé (Tibetan Manuscript Project Vienna, University of Vienna)

Abstract Various versions of e-texts of Tibetan canonical literature help researchers to identify and locate specific textual passages. At present, the available material includes three sets of Kanjurs, one set of Tanjur, and one set of Old Tantra. Given the vast amount of textual material, efficient quotation search queries require e-texts that have been optimized for those queries.

In this paper, I will present different methods of storing and processing textual data employed in the Resources for Kanjur & Tanjur Studies (rKTs) database (http://www.rkts.org/). I will mainly focus on the accuracy of the results and other aspects of the performance of the different methods.

3.8 Li, Channa, Marco Peer and Kaifeng Yang

Title **Looking for Chos grub: Proposing An AI-Assisted Method for Recognizing Scribal Hands of Dunhuang Tibetan Manuscripts**

Presenters Channa Li (IKGA, Austrian Academy of Sciences), Marco Peer (Technische Universität Wien), and Kaifeng Yang (Upper Austria University of Applied Sciences)

Abstract The field of manuscript studies is evolving through the wide implementation of digital tools. Our pilot study aims to contribute to this process by developing AI techniques for identifying Tibetan manuscripts written by Chos grub, a famous Buddhist translator active in 9th-century Dunhuang. In a first

step, we propose to use Deep Learning-based, image processing algorithms to group, retrieve, and extract all accessible Dunhuang Tibetan manuscript images on the basis of handwriting styles. In a second step, we sparsify the outputs obtained from the first step in interaction with explainable AI techniques and Bayesian optimization. This will allow us to construe optimized models of Chos grub's cursive Tibetan handwriting, which are interpretable and verifiable by human experts. With the aid of AI-based hand recognition techniques, we hope to identify more philosophical works written by Chos grub (especially, works on the *Yogācārabhūmiśāstra*, a fundamental treatise for the Buddhist Yogācāra school), thereby complementing our knowledge of Buddhist intellectual history.

3.9 Li, Charles

Title **Reading around the edges: doing research with manuscript metadata**

Presenter Charles Li (National Centre for Scientific Research, Paris)

Abstract In the rush to create corpora for large-scale textual analysis, many details are lost, out of necessity. Electronic text corpora generally deal with edited texts, rather than manuscript transcriptions, dropping all the textual variants and scribal peculiarities in the process. Moreover, paratextual content — such as scribal colophons, the initial namaskāra, or marginal annotations — is usually ignored. However, given the research potential inherent in these paratexts, the Texts Surrounding Texts project has created a manuscript database — focusing on the BnF Paris and SUB Hamburg collections — devoted to studying this often neglected material (https://tst-project.github.io/). As with most projects of this genre, we are producing detailed records using TEI XML. But we also aggregate the XML-encoded metadata to look for patterns in scribal practices and generate graphs of relations between scribes, owners, and collectors. In this paper, I will present some of the data that we have already harvested and look at what paratexts can tell us about how manuscripts were produced, who they were made by, and what they were made for.

3.10 Meelen, Marieke

Title **NLP for Tibetan corpus creation: how to get more out of our data?**

Presenter Marieke Meelen (University of Cambridge)

Abstract This talk will focus on answering questions about the Tibetan languages and their history, and, more specifically, what kind of data and annotation we need to be able to address those in the first place. Ultimately, if we want to learn how, for example, evidentiality and egophoricity emerges in their various forms, we need both deeply-annotated historical corpora, but also transcribed and annotated data from (often endangered) modern spoken Tibetan varieties. I will present several case studies on how to exploit state-of-the-art NLP techniques to create annotation pipelines to facilitate diachronic linguistic research: from basic POS tagging & parsing to Animacy detection and Automatic Speech Recognition (ASR) in historical and endangered low-resource languages.

3.11 Nagasaki, Kiyonori

Title **Revisiting Text Encoding for Buddhist Studies**

Presenter Kiyonori Nagasaki (International Institute for Digital Humanities, Tokyo)

Abstract Electronic texts have the potential for a variety of uses beyond the traditional boundaries of text. Uniform encoding methods are crucial to the effective sharing of texts: these methods make it possible to process texts in an integrated manner, even if they are from different fields or languages. Therefore, we at the International Institute for Digital Humanities are actively engaged in the international standardization of character and text encoding for Buddhist scriptures, with the aim of defining international methods that are applicable to encoding texts from other disciplines as well. Recently, a useful international standard for treating digital facsimiles as witnesses has also emerged. I will introduce our current and future activities on Unicode, TEI (Text Encoding Initiative) and IIIF (International Image Interoperability Framework) in this presentation.

3.12 Nehrdich, Sebastian

Title **Creating a Shared Semantic Vector Space for Buddhist Languages**
Presenter Sebastian Nehrdich (Düsseldorf University, Hamburg University)
Abstract Philological work in Buddhist Studies usually requires work with material preserved in various languages. In recent years, significant progress has been made in creating shared semantic spaces for contemporary high-resource languages, enabling effective semantic search across language boundaries. With the amount of digitally available primary Buddhist material constantly increasing, mapping these languages into a shared vector space becomes more and more feasible. The most important prerequisite for the successful creation of a shared semantic space is the accumulation of high-quality parallel data. In this talk, I will discuss what strategies can be used to accumulate this data, what amounts of parallel data for Buddhist primary languages are currently available, and what further steps are necessary in order to further increase the amount of data. I will also discuss strategies to evaluate the quality of the created language models and how they could potentially be integrated into downstream applications.

3.13 Neill, Tyler

Title **Vātāyana: A window onto the Pramāṇa NLP corpus**
Presenter Tyler Neill (Independent Researcher)
Abstract Vātāyana (`https://vatayana.info/`) is a digital humanities project that grew out of experimenting with a number of natural language processing (NLP) techniques on Sanskrit textual data while also writing a philological dissertation focused on one *pramāṇa* text (Bhāsarvajña's Nyāyabhūṣaṇa). Once it appeared that these NLP techniques could help in locating certain instances of intertextuality between such Sanskrit philosophical works, the effort was expanded to be the second half of that dissertation (defended Dec 2022) and now also a free-standing system.

In the first part of this presentation, I will lay out the various components of Vātāyana: the Pramāṇa NLP corpus on which it is based, the NLP techniques and tools used, the overall intertextuality detection algorithm, and the

web-app interface. In the second part, we will use both the online web-app interface and an offline batch-processing interface to explore some examples together.

3.14 Roux, Élie

Title **OCR at the Buddhist Digital Resource Center: status and perspectives**

Presenter Élie Roux (BDRC)

Abstract The Buddhist Digital Resource Center (BDRC, https://www.bdrc.io/) has a collection of 20+ millions of images of texts in the Tibetan, Khmer and Burmese scripts. While the catalog of the collection covers hundreds of thousands of texts, some parts of the collection are still unexplored and their contents difficult to discover. In order to unearth these treasures, BDRC is engaging in a large-scale effort to extract data from these images, the first and primary step being the extraction of text through OCR and HTR. We will present BDRC's efforts in that direction, as well as perspectives on the use of this new corpus.

3.15 Viehbeck, Markus

Title **Preserving and Disseminating Tibetan Canonical Literature Through Digital Means: The Resources for Kanjur and Tanjur Studies (rKTs) Archive**

Presenter Markus Viehbeck (*Tibetan Manuscript Project Vienna* TMPV, University of Vienna)

Abstract The Tibetan Manuscript Project Vienna (TMPV, https://tmpv.univie.ac.at) is a long-term research initiative aiming at fostering preservation, dissemination, and investigation of Tibetan canonical literature. Under its aegis, numerous Himalayan manuscripts collections, often in highly endangered conditions, have been preserved in a digital format. Images of these collections along with electronic catalogues and other digital tools are made openly available through the online archive Resources for Kanjur and Tanjur Studies (http://rkts.org), which has developed into a powerful resource, in particular for philological studies of Buddhist canonical texts.

This talk will introduce the scope and recent activities of the TMPV, provide concrete guidelines on how to use the rKTs archive, and reflect about future tasks and possibilities of this research initiative.

4 Keynote: Marcus Bingenheimer, "On the Use of Historical GIS in the Study of Chinese Buddhism"

Can historical GIS contribute to our understanding of Chinese Buddhism? The talk will describe some of the methods and data currently available for the application of historical GIS in a number of case studies. The aim is to show that geographical visualization and analysis can be a useful research tool, and that mapping large collections as well as individual itineraries can provide new, at times surprising, insights.

The cases we will discuss include: **1)** A comparative mapping of the "geographic range" of different collections of biographies of eminent monks: this reveals a shrinking of the Buddhist world in the second millennium. **2)** How visualizing the "biographical topologies" of monks can easily generate new research questions. **3)** How the mapping of pilgrimage routes in 19th and 20th century China renders a religious map of China from the perspective of pilgrimage, one of the most popular forms of Buddhist practice. We can, for instance, compare the more than 50 itineraries of "Knowing the Paths of Pilgrimage" (Canxue zhijin 參學知津) by Ruhai Xiancheng 如海顯承 (fl. 1800–1826) with those described in the "Records of Travels to Famous Mountains" (Mingshan youfang ji 名山遊訪 記) by Gao Henian 高鶴年 (1872–1962), and show that, whereas the destinations of Buddhist pilgrims in the early 19th century did not differ much from those in the early 20th centuries, some routes changed considerably with the arrival of new modes of transport. **4)** How one can understand the very uneven distribution of overseas Chinese temples in Bangkok by combining modern survey data with historical maps.

ÖAW AUSTRIAN ACADEMY OF SCIENCES

27-28 APRIL 2023
UNIVERSITY OF VIENNA
INSTITUTE FOR SOUTH ASIAN, TIBETAN AND BUDDHIST STUDIES, SR 1
SPITALGASSE 2, COURT 2, ENTRANCE 2.7
1090 VIENNA, AUSTRIA
HYBRID EVENT

ADVANCED COMPUTATIONAL METHODS FOR STUDYING BUDDHIST TEXTS

INTERNATIONAL SYMPOSIUM ORGANISED BY PATRICK MCALLISTER, RACHAEL GRIFFITHS, AND MARKUS VIEHBECK

For details:

https://oeaw.ac.at/ikga

Please contact patrick.mcallister@oeaw.ac.at if you wish to attend in person. No registration is required for joining through Zoom.

Hosts:

- Institute for the Cultural and Intellectual History of Asia (IKGA)
- ERC project "The Dawn of Tibetan Buddhist Scholasticism (11th-13th c.) (TibSchol)", IKGA
- Tibetan Manuscript Project Vienna, University of Vienna

IKGA

European Research Council
Established by the European Commission

universität wien
Institut für Südasien-, Tibet- und Buddhismuskunde

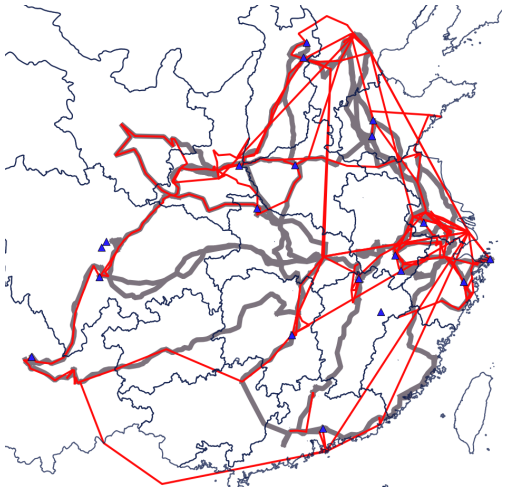
WWW.OEAW.AC.AT

ÖAW AUSTRIAN ACADEMY OF SCIENCES

27 APRIL 2023, 5PM
UNIVERSITY OF VIENNA
INSTITUTE FOR SOUTH ASIAN,
TIBETAN AND BUDDHIST STUDIES, SR 1
SPITALGASSE 2, COURT 2, ENTRANCE 2.7
1090 VIENNA, AUSTRIA
HYBRID EVENT

ON THE USE OF HISTORICAL GIS IN THE STUDY OF CHINESE BUDDHISM

KEYNOTE BY MARCUS BINGENHEIMER



Can historical GIS contribute to our understanding of Chinese Buddhism? The talk will describe some of the methods and data currently available for the application of historical GIS in a number of case studies. The aim is to show that geographical visualization and analysis can be a useful research tool, and that mapping large collections as well as individual itineraries can provide new, at times surprising, insights.

This lecture is the keynote speech for the international symposium "Advanced computational methods for studying Buddhist texts."

Marcus Bingenheimer 馬德偉 was born in Germany. He obtained an MA (Sinology) and a PhD (History of Religions) from Würzburg University, and an MA (Communication Studies) from Nagoya University. Marcus works as Associate Professor in the Department of Religion at Temple University in Philadelphia, but is currently on sabbatical. He taught Buddhism and Digital Humanities in Taiwan at Dharma Drum and held visiting positions and fellowships at universities in Korea, Japan, Thailand, Singapore, and France. Since 2001 he has supervised various projects concerning the digitization of Buddhist culture. His main research interests are the history and historiography of Buddhism, early sūtra literature, and how to apply computational approaches to research in the humanities.

For details:

https://oeaw.ac.at/ikga

IKGA

European Research Council
Established by the European Commission

universität wien Institut für Südasien-, Tibet- und Buddhismuskunde

IKGA – INSTITUTE FOR THE CULTURAL AND INTELLECTUAL HISTORY OF ASIA