The Future of Digital Texts in South Asian Studies — A SARIT Workshop

2017-05-22 to *2017-05-24*

http://www.ikga.oeaw.ac.at/Events/SARIT_Workshop_2017

Birgit Kellner, Patrick McAllister, Andrew Ollett

Version: *2017-05-20*

Contents

1 Introduction and practical information

As a conclusion to a four-year project dedicated to developing and enriching a collection of digital texts in Sanskrit and other Indian languages, the team behind SARIT is convening a workshop called "The Future of Digital Texts in South Asian Studies." The goal is twofold. First, we want to survey and reflect on the current state of digital texts in our field. What is a "digital text"? How are they produced? Who is responsible for them? How are they provided to users? Who are their users, and what do they do with them? How, if at all, have they changed the landscape of research and teaching? Second, we want to reflect on the future of digital texts. What could we be doing with them that we aren't doing yet? What inspiration can we take from projects in other fields? What emerging technologies can we take advantage of? How can we better integrate our various digital projects? How can we involve communities of students, teachers, and researchers in the production, curation, and publication of digital texts?

The workshop is also intended to stimulate discussion on the future of SARIT.

1.1 Practical Information

- Where:
  - 'Seminarraum' (room 2.25), Institute for the Cultural and Intellectual History of Asia (IKGA)
  - Hollandstraße 11+13/2nd floor, 1020 Vienna, Austria. (map)
- When: *2017-05-22* to *2017-05-24*
- Contacts:
  - patrick.mcallister@oeaw.ac.at
  - office.ikga@oeaw.ac.at
  - T: (+43 1) 515 81 / 6400
- Registration: to help us prepare, please register per email to **both** contacts, office.ikga@oeaw.ac.at and patrick.mcallister@oeaw.ac.at no later than *2017-05-07*.

2 Schedule

2.1 *2017-05-22*

Registration will be open from **9:00** in the 'Sekretariat', room 2.49, at the IKGA.

Opening session

1. **10:00–10:30** Birgit Kellner: *The development of SARIT 2013–2017: goals, achievements, problems*

2. **10:30–11:00** Dominik Wujastyk: *What do users want from SARIT in future?*
3. **11:00–11:30** Coffee break

History through Indic Texts (1) // Chair: Masahiro Shimoda

1. **11:30–12:30** Bronner, Yigal: *Indic Prosopography in the Digital Age*
2. **12:30–14:00** Lunch break
3. **14:00–15:00** Kellner, Birgit: *Bibliography and prosopography in the digital age: EAST (Epistemology and Argumentation in South Asia and Tibet) and its challenges*
4. **15:00–16:00** Baums, Stefan: *Documents, Databases and Networks: Scholarly Work on Gāndhārī in the Digital Age*
5. **16:00–16:30** Coffee break
6. **16:30–18:30** Software Demonstrations

2.2 *2017-05-23*

History through Indic Texts (2) // Chair: Nina Mirnig

1. **10:00–11:00** Bajracharya, Manik and Christof Zotter: *Turning pre-modern documents into digital texts: The pragmatics of an approach*
2. **11:00–11:30** Coffee break

Computational Linguistics for Indic texts // Chair: Lou Burnard

1. **11:30–12:30** Scharf, Peter M.: *Creative and intelligent use of linguistic, textual, and bibliographic information to enhance interlinked access to lexical, textual, and image data*
2. **12:30–14:00** Lunch break
3. **14:00–15:00** Kulkarni, Amba: *Bridging the gap between Computational tools and Sanskrit Digital Libraries: Where do we stand?*
4. **15:00–16:00** McAllister, Patrick: *Searching Sanskrit Texts*
5. **16:00–16:30** Coffee break

Computer-assisted Editing of Indic Texts (1)

1. **16:30–17:30** Bellefleur, Tim and Adheesh Sathaye: *Developing Linked Data Standards for Working with Sanskrit Manuscript Traditions*
2. **17:30–18:30** Balogh, Dániel: *Building a Database of Indic Inscriptions*
3. **19:00–22:00** Dinner

2.3 *2017-05-24*

Computer-assisted Editing of Indic Texts (2) // Chair: Karlheinz Mörth

1. **10:00–11:00** Li, Charles: *Editors as Maintainers*
2. **11:00–11:30** Coffee break
3. **11:30–12:30** Tomabechi, Toru: *TEI Markup of Abhayākaragupta's Āmnāya-mañjarī: An Attempt to Create an "Open Research Note" for the Study of Late Indian Buddhism*
4. **12:30–14:00** Lunch break
5. **14:00–15:00** Maas, Philipp: *Sanskrit Textual Criticism in the Digital Age – Will Really Everything Change?*
6. **15:00–16:00** Ollett, Andrew: *A Less Distant Future: Sanskrit Texts for Scholarly Communities in the Digital Age*
7. **16:00–16:30** Coffee break

Closing

1. **16:30–18:00** Closing words, round-table discussion: *What next for SARIT?*

3 Participants and Abstracts

3.1 Balogh, Dániel

- British Museum
- danbalogh@gmail.com

Building a Database of Indic Inscriptions

This paper introduces a recent initiative in digital epigraphy under the aegis of the ERC Synergy project 'Beyond Boundaries – Religion, Region, Language and the State.' The project as a whole aims to re-evaluate the social and cultural history of the Gupta period in South, Central and Southeast Asia and approach an understanding of the region as an interconnected cultural network. One component of this project is the 'Siddham' database of Indic epigraphic texts. Its development was commenced in the summer of 2015 with the encoding of previously published Sanskrit inscriptions created under the imperial Gupta rulers. It will be progressively expanded both horizontally (by adding inscriptions from other dynasties and regions) and vertically (by accumulating metadata, gradually increasing the granularity of markup, and through re-editing crucial inscriptions).

EpiDoc (an application of TEI for encoding epigraphic documents) serves as the flesh and blood of our corpus: texts are stored in XML snippets, each

comprising the edition division of a full EpiDoc file. Siddham's skeleton is made up of relational database tables. The edition snippets, along with other snippets containing translations (and, optionally, critical apparatus and commentaries), are referenced from an "Inscriptions Table" that additionally stores metadata pertinent to each inscription, such as layout and hand description, language and date. A separate "Objects Table" serves as the repository of metadata pertaining to inscription-bearing objects, such as physical properties (material, dimensions and freeform description) and history. The separation of object metadata from inscription metadata is conceptually desirable as it brings objects to the fore as entities in their own right rather than mere dismissible substrates of the texts they carry. It is our hope that Siddham will not only become a useful reference tool for textual scholars of Indic languages, but will, through the foregrounding of the objects themselves and through the inclusion of translations, also encourage the formulation of new types of questions that scholars of various disciplines may ask of text-bearing objects of this region.

3.2 Baums, Stefan

- Bavarian Academy of Sciences and Humanities, Ludwig Maximilian University of Munich
- baums@lmu.de

Documents, Databases and Networks: Scholarly Work on Gāndhārī in the Digital Age

Gāndhārī is a Middle Indo-Aryan language attested from the third century BCE until the fifth century CE. Until quite recently, it was known almost exclusively from inscriptions, administrative documents and a single literary manuscript, and this scarcity and specialization of sources meant that no comprehensive dictionary or grammar of the Gāndhārī language were ever attempted. The situation changed radically with the discovery, in the 1990s, of about one hundred long manuscripts containing Buddhist and literary texts, which are now in the process of being edited. The new manuscript discoveries prompted Andrew Glass and myself to undertake (in the year 2002) the compilation of a digital corpus of Gāndhārī texts as the basis for our Dictionary of Gāndhārī. Our corpus reached completion several years ago (with a total of currently 2,751 texts), and we have made significant progress in lemmatization, morphological marking and article writing (currently 6,713 articles covering 33,403 references) with incipient support for syntactic treebanking. We make our complete Gāndhārī corpus as well as our in-progress lexicographic work available from our website http://gandhari.org (which also provides a selection of Old and Middle Indo-Aryan dictionaries for the convenience of users).

In parallel with our lexicographic work, the field of Gāndhārī manuscript studies has broadened significantly in the last twenty years, with two major projects (the Early Buddhist Manuscripts Project in Seattle and the Buddhist Manuscripts from Gandhāra Project in Munich) now involved in the edition of the new discoveries. To support the work of these two projects in particular, and of scholars of early South Asian documents more generally, we spearheaded the development of a new digital toolset called Research Environment for Ancient Documents (READ) that has received funding and adoption from a number of institutions. Design decisions for READ were shaped by an analysis of the workflows of the target user communities and of the intended scholarly products, as well as by social considerations such as the representation of individual scholarly work and distinct project identities. Two defining characteristics of READ are the linked parallel storage of multiple editions of the same source text, and the linking of images and transcriptions as a basis for navigation, paleographic work and pedagogical applications. The core technologies behind READ are relational databases (PostgreSQL) for storage, a PHP/HTML/JavaScript system for interaction with stored content and for content creation, and TEI P5 XML documents for export, import and archival purposes. Capabilities for networking textual corpora and analytical works on different servers (either running READ or providing other TEI-based interfaces) are currently under development. At the same time as this technical work, the editors of the Gandhāran Buddhist Texts series made new publication and archival arrangements ensuring the perpetual availability of editions in the series under open-access licensing and in open data formats, both online and through print on demand.

The present paper will give an overview of the history and prospects of these digital initiatives for Gāndhārī studies and related fields. It will consider some technical differences between relational and document databases and how they relate to the traditional organization of scholarly work. It will argue that the combination of open formats and licenses with a document level of organization and networking capabilities is key to ensuring a proper balance between content sharing on the one hand and local resource management and the integrity of individual scholarly work on the other.

3.3 Bajracharya, Manik and Christof Zotter

- manik.bajracharya@adw.uni-heidelberg.de
- christof.zotter@adw.uni-heidelberg.de
- Documents on the History of Religion and Law of Pre-modern Nepal, Heidelberg Academy of Science and Humanities

Turning pre-modern documents into digital texts: The pragmatics of an approach

Digital humanities provide powerful tools that can facilitate the research work and enhance the options of scientific analysis, also in South Asian studies. But, having their own requirements, digital solutions can also consume a lot of man power, especially if they need to be tailored to the needs of researchers and users. Furthermore, it requires "esoteric" technical knowledge to understand what is possible and how it can be realised; a knowledge that is not a usual part of the curriculum in South Asian Studies. With a special focus on the production of digital texts, this contribution will report about the experiences of a humanities' project that decided to enter the challenges of working digitally. It will address some of the questions that accompanied the process of setting up, extending and refining the project's IT structure and will give an account of a still ongoing quest for workable solutions, the cost-benefit-ratio in defining standards, the organisation of workflows, collaborative work and authorship, and the unavoidable compromises.

The research unit "Documents on the History of Religion and Law of Premodern Nepal" (see http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/nepal/welcome.html) of the Heidelberg Academy of Science and Humanities employs digital humanities in various ways. It has built up a MySQL database to enter catalogue data of a huge corpus of historical documents in order to make them accessible to the project members but also to the general public, and it provides digital texts, namely editions of selected documents prepared according to the standards of the Text Encoding Initiative (TEI). These two building blocks are linked with each other and associated with a bibliography, a glossary and, as latest feature, an ontology of named entities. The planning and implementation of all these components went along with sometimes extensive discussions about the issues mentioned above. Similar processes will accompany future steps of the project, such as the involvement of a lemmatizer, OCR and the automatized annotation of at least some formal features.

One big advantage of digital texts is that associated information is extensible and that they allow for different reuses. However, to keep a text 'alive' it needs not only convertible formats and a manageable technical habitat, but also an, at least partly, new understanding of what is actually being produced.

3.4 Bellefleur, Tim and Adheesh Sathaye

- tbelle@alumni.ubc.ca
- University of British Columbia
- Adheesh.Sathaye@ubc.ca
- Dept. of Asian Studies, University of British Columbia

Developing Linked Data Standards for Working with Sanskrit Manuscript Traditions

In the world of digital textual scholarship, we generally focus either on creating artifacts (e-texts, collations, stemmata) or applications—the software tools that create, process, and present these artifacts. While SARIT has, quite rightly, focused its efforts on developing and maintaining encoding standards for artifacts—especially e-texts—the needs of modern digital philology require a more robust set of standards for how different applications may interface with such artifacts, as well as with one another. The development of such standards, we suggest, may help us get one step closer to the "holy grail" of an extensible and user-friendly digital environment for Sanskrit textual scholarship. In this joint presentation, we will first present the perspective of the end-user, the textual scholar, and explore why one would need networked textual data that is streamlined, multi-dimensional, and responsive to human decision-making and workflows. We will then offer some solutions that we have been developing as part of the Vetāla Project at UBC through the use of Linked Data standards, such as Open Annotations, to facilitate the robust connections between artifacts and applications.

3.5 Bronner, Yigal

- The Hebrew University of Jerusalem
- yigal.bronner@mail.huji.ac.il

Indic Prosopography in the Digital Age

Panditproject.org is a digital humanities project with a unique and ambitious task: to create a database for the vast world of South Asian letters. The name stands for the Sanskrit title of a virtuoso scholar with full mastery of traditional knowledge systems, but as an acronym it also expresses the project's main objective: the creation of a Prosopographical Database of Indic Texts. In brief, Panditproject.org seeks to store, curate, and share reliable data on works, people, places, institutions, and manuscripts from premodern South Asia, in addition to relevant secondary sources, and to do so across period, language, discipline, and subject matter. It is designed as an interactive web-based repository that scholars of every South Asian specialty and interest can contribute to and as a basic tool on which they will routinely come to rely. I propose to give a hands-on presentation of the database and its possibilities and remaining challenges.

3.6 Burnard, Lou

- lou.burnard@retired.ox.ac.uk

- Former Assistant Director of Oxford University Computing Services (OUCS)
- Central contributor to the Text Encoding Initiative (TEI)

Lou Burnard will chair the panel Computational Linguistics for Indic texts.

3.7 Hellwig, Oliver (*Cancelled*)

Unfortunately, this presentation had to be cancelled.

- SFB 991, University of Düsseldorf
- oliver.hellwig@indsenz.com

Machine Learning Techniques in an Indological Context

Although Digital Humanities as a research paradigm have promoted the interaction between text-oriented Philology on one and quantitative Natural Language Processing (NLP) and Machine Learning (ML) on the other side, there is still an enormous conceptual gap between these two approaches to texts and language. In most research scenarios, neither side is aware of specific problems and solutions presented by the other one. While NLP and ML provide efficient frameworks for learning and for reasoning based on partly observable information, philological disciplines assemble the specialized knowledge and the – partly undigitized – resources for understanding historical texts. With the exception of few "fashionable" areas such as topic or graph analysis, Digital Humanities as the "boundary discipline" has not been able to connect both fields effectively.

This paradigmatic separation has considerable consequences. Most NLP studies work with the same datasets that cover a limited set of modern languages[1] and of intellectual domains (mainly newspaper texts). They often silently assume that mechanisms working for the benchmark data sets of modern newspaper English will be equally efficient on out-of-domain texts and/or texts in other (ancient) languages. Philology, on the other hand, frequently does not make use of standard methods from NLP or ML that could be helpful in structuring the available data and in drawing scientifically sound conclusions from them.

The presentation will focus on two quantitative approaches that can strongly increase the efficiency of philological reasoning.

1. **Supervised classification** deals with predicting the class of a new instance, given a set of previously labeled instances. Supervised classification is especially useful in the context of corpus annotation, where

1. English and Chinese; modern German, for example, is almost considered as an under-resourced language.

new instances (e.g., unannotated words or syntactic structures) should be labeled automatically by using an ML model. The presentation will introduce several research cases, in which **Deep Learning** models are used for the morphological, lexical, semantic, and syntactic annotation of Sanskrit texts.

2. Although Topic Models are quite popular in Literary Studies, the underlying field of **Graphical Models** has not found much attention in philological research. Graphical models provide a principled method of evaluating causal relationships between large numbers of variables in textual data, and allow to draw conclusions about "hidden factors" such as authorship given only a limited set of observed data (words, topics). The presentation will given an informal introduction into the theory underlying Graphical Models, and will sketch how problems of authorship attribution and text stratification can be formulated in this framework.

Useful Links

- *ML* Graphically appealing and non-technical introduction into Machine Learning: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- *Graphical Models* A comparatively non-technical introduction to Graphical Models, with some interesting pointers to phylogenetics and document processing: https://projecteuclid.org/download/pdfview_1/euclid.ss/1089808279
- *Deep Learning* The "Deep Learning Tsunami" and Computational Linguistics: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00239

3.8 Kellner, Birgit

- Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences
- birgit.kellner@oeaw.ac.at

Bibliography and prosopography in the digital age: EAST (Epistemology and Argumentation in South Asia and Tibet) and its challenges

In 1995, Ernst Steinkellner's and Michael Torsten Much's systematic survey of the literature of the logico-epistemological school of Buddhism, chiefly in India, was published in print, as part of a larger endeavour to systematically document Buddhist Sanskrit Literature. (Systematischer Überblick über die Literatur der erkenntnistheoretisch-logischen Schule des Buddhismus. Göttingen 1995: Vandenhoeck & Ruprecht; Systematische Übersicht über die buddhistische Sanskrit-Literatur 2).

The survey offered whatever biographical information was available at the time about individual thinkers chiefly of the *pramāṇa* tradition. But its main goal was to comprehensively document publications in a well-defined system informed by the textual scholar's main interests, using works as the main anchor of classification. Publications were categorized in terms of whether they contained full and partial editions or translations, whether they offered textual fragments, or contained glossaries and indices.

Efforts to transform the data provided by Steinkellner and Much into a database structure date back to the early 2000s (with financial support by the Austrian Science Fund FWF), but it was only within the framework of, first, the Cluster of Excellence "Asia and Europe in a Global Context" of the University of Heidelberg and, then, the DFG-NEH supported SARIT project that a major push could be made to produce a new interactive digital resource pursuing the same overarching scholarly goal – comprehensive documentation of logico-epistemological literature – while making use of new technological possibilities to continually update information and offer it in a form that was better attuned to the dynamics of the Web. The resource EAST (Epistemology and Argumentation in South Asia and Tibet, http://east.uni-hd.de) was released in 2011, and has been updated ever since.

In this contribution I shall offer a brief presentation of EAST, but mainly use EAST as an example for a more general discussion of how the move from print to digital offers challenges for the production and maintenance of prosopographical and biographical resources (with particular focus on South Asian Studies).

3.9 Kulkarni, Amba

- Indian Institute of Advanced Study, Shimla
- ambapradeep@gmail.com

Bridging the gap between Computational tools and Sanskrit Digital Libraries: Where do we stand?

The last decade has witnessed vibrant activities in the field of Sanskrit Computational Linguistics. Several tools have been developed performing various tasks such as word analysis, word generation, segmenting a sandhied text into meaningful components, compound analysis and dependency parsing. Unlike other natural languages, Sanskrit has the advantage of having an almost exhaustive grammar. At the same time it is unique in allowing parallel meanings running across the texts. Both these features pose challenges for a computational linguist, since now there is a demand to produce all possible analyses in parallel by providing the justification in terms of grammar rules.

In this paper I describe the efforts in building **Saṁsādhanī**, a platform for Sanskrit Computational Linguistics that has the features described above.

The platform operates interactively with the **Heritage** segmenter to produce various possible analyses. The interactive user interface is developed to share the load between man and machine in such a way that tasks hard for human being are done by the machine and vice versa.

This platform is being used to develop e-readers for various Classical texts such as Śrīmad Bagavadgītā, Śiśupāvadham, Bhaṭṭikāvya, Mahābhārata, etc. serving dual purpose. On the one hand they demonstrate the effective use of technology in reviving the traditional methods of teaching following the Khanḍānvaya and on the other hand they help in generating annotated texts that will help bootstrapping the Machine learning efforts.

3.10 Li, Charles

- University of Cambridge
- cchl2@cam.ac.uk

Editors as Maintainers

The emergence of electronic texts, and our increasing reliance upon them, has made the shortcomings of printed editions readily apparent. However, most electronic texts are still conceived of as digital facsimiles of printed books; they are not authoritative in their own right, and they usually strive only to reproduce the printed text faithfully, even if that text might be incorrect. There is no framework in which corrections to published editions can be suggested by third parties, or in which new evidence can be incorporated into an existing edition, other than by re-editing it and re-publishing it. In order to change this, we would need to change our conception both of the edition itself — not as a fixed text, but as a body of evidence and hypotheses that are progressively improved — and of the role of an editor — not as a compiler of a fixed text, but as a maintainer of a textual tradition, who takes on the task of reviewing suggestions, corrections, and new evidence, and updates the edition as necessary. In this model, an edition would have one or more maintainers, who oversee the project, and a virtually unlimited number of contributors, whose work — in the form of transcriptions, emendations, testimonia, critical notes, etc. — would be recorded with a version control system.

Technology Demonstration

In the process of preparing a new critical edition of Bhartṛhari's Dravyasamuddeśa with the commentary of Helārāja, I have developed some open source tools to collate diplomatic transcripts of manuscript witnesses and to display an interactive apparatus alongside the text. Compared to a traditional, printed edition, this digital edition does not treat the apparatus as part of the

content, but as a dynamically-generated analysis of the content, which consists of the witnesses themselves. Each witness is considered a text in its own right, and each transcript a faithful representation of that text. In this way, the edition becomes a collection of documents that can be easily added to if new witnesses are discovered. I will be demonstrating the user interface of the edition both from the point of the view of an editor — covering the transcription process and methods of collaboration —, and from the point of view of a reader — exploring the interactive tools for researching textual variation. The online edition can be found at http://saktumiva.org/wiki:dravyasamuddesa:start, and the source code for the software is on GitHub (http://github.com/chchch/upama).

3.11 Maas, Philipp

- University of Leipzig
- Philipp.A.Maas@gmail.com

Sanskrit Textual Criticism in the Digital Age – Will Really Everything Change?

Indology and South Asian Studies research the cultures of South Asia in their historical contexts. To this end, these disciplines strongly depend on information contained in primary sources written in Sanskrit and other languages, among which the works of the multiple genres of literature are highly important. However, most works of Sanskrit literature are no longer available in the version in which they were originally composed and written down. All that is available are printed editions based on mostly unidentified manuscript copies. Moreover, approximately six million mostly unsought manuscripts exist, which are mostly copies produced from previous copies along unknown (but not unknowable) lines of transmission. This manuscripts heritage, the largest of all cultures worldwide, is the object of research of Sanskrit textual criticism. Its double interrelated objective is traditionally regarded as (1) reconstructing as exactly as possible a text version that resembles as closely as possible a text version that would have been acceptable to the author or final redactor of a given work, and (2) investigating the transmission history of the work under research.

In spite of the fundamental importance of critical editing for any kind of text based research, only very few critical editions of Sanskrit works have been produced so far. And virtually all existing critical editions were designed to be published in print. Now, the so-called digital revolution provides scholars with completely new options and possibilities for critically editing and for the publication of their work. Peter Robinson, a pioneer and leading expert in digital editing works of English literature, has recently argued that new technical tools and web-based publication facilities may

fundamentally change the methods and aims of critically editing (P. Robinson, "The Digital Revolution in Scholarly Editing." Eds. Barbara Crostini et al., Ars Edendi Lecture Series. Vol. 4. Stockholm 2016, pp. 181–201 [http://www.stockholmuniversitypress.se/site/books/10.16993/baj/]). The present talk will critically examine Robinson's assessments with special reference to the aims and objectives for Sanskrit textual criticism as it is (and will be) applied in the DFG-sponsored project "A Digital Critical Edition of the Nyāyabhāṣya" at the University of Leipzig, Germany.

3.12 McAllister, Patrick

- Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences
- patrick.mcallister@oeaw.ac.at

Searching Sanskrit Texts

Searching electronic data is a large and complex field in modern information technology. Simple search interfaces belie the sophistication of both the theoretical foundation and the practical engineering that make these searches possible. Over the last decades, important parts of this technology, on the theoretical as well as on the practical side, have become publicly available. This has made it possible to apply these tools to the searching of texts in languages that are decidedly not in the center of interest for most search companies and also most professional programmers, such as Sanskrit.

This paper will first give an overview of the search techniques most commonly used for Sanskrit texts, with a specific focus on the ones used on SARIT's public interface (http://sarit.indology.info), arguably the most advanced public search engine for Sanskrit texts, along with their respective advantages and limitations. After that, several less used search methods will be investigated as to their utility for searching Sanskrit texts, particularly ones based on synonyms, phonetic algorithms, and translations. From an evaluation of these various approaches, guided by some practical considerations, the reasonable expectations for future improvements to the computer-assissted searching of Sanskrit texts will become clearer.

3.13 Mirnig, Nina

- Nina.Mirnig@oeaw.ac.at
- Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences

3.14 Mörth, Karlheinz

- Karlheinz.Moerth@oeaw.ac.at
- Director of the Austrian Centre for Digital Humanities (ACDH-OeAW)
- Austrian Academy of Sciences

Dr. Karlheinz Mörth will chair the panel Computer-assisted Editing of Indic Texts (2).

3.15 Ollett, Andrew

- andrew.ollett@gmail.com
- Harvard University

A Less Distant Future: Sanskrit Texts for Scholarly Communities in the Digital Age

In the current funding cycle for SARIT, the Columbia University subproject has prepared a series of texts, with a focus on poetics (*alaṃkāraśāstra*) and hermeneutics (*mīmāṃsā*). At the start of the project, we were relatively new to TEI, and believed that it could improve on the existing models, both formal and informal, of how people interact with texts in our field. First, TEI texts have an advantage over printed texts in that their structure and content is machine-readable. For most users, this simply means "searchable," but we were interested in a wide range of other possible applications: named entity recognition, alignment, identification of quotations, word cooccurrence patterns, and so on. Second, TEI can represent features of a printed edition that plain text files typically don't, including notes, front and back matter, a critical apparatus, pagination and lineation, and so on. Our hope was that, by putting all of this information into the digital text, the digital text would be as "citeable" for scholarly purposes as the printed edition on which it was based. That is, in addition to being "machine-readable," the text would be "scholar-readable." Third, we hoped that our texts would be dynamic rather than static, open to the scholarly community for further improvement and annotation. Our test-case would have been Abhinavagupta's *New Dramatic Art* (*Abhinavabhāratī*), in which a careful reader can conjecturally improve the printed edition on almost every single page. These texts should therefore also be "community-readable." This talk will cover the progress we've made, and the challenges we've faced, in producing machine-, scholar-, and community-readable texts. We have found that there are two limiting reagents in this process: the considerable human labor involved in converting printed editions to high-quality TEI texts (which grows exponentially when, as is often the case in our field, the typographic and editorial conventions of the source edition are inconsistent), and the total inadequacy of existing applications for interacting with these kinds of texts. How can these limitations be addressed or

overcome? We'll share some suggestions from our experience, and from new approaches to TEI publishing that SARIT is now taking advantage of. We'll also offer a model for interacting with digital texts—reading with a selection of commentaries, facsimiles of printed editions and manuscripts, navigable cross-references, a critical apparatus, annotations, and bibliographic information at one's fingertips—that is much closer to realization now than it was when this project started.

3.16 Sathaye, Adheesh and Tim Bellefleur

- Adheesh.Sathaye@ubc.ca
- Dept. of Asian Studies, University of British Columbia

See Bellefleur, Tim and Adheesh Sathaye for the abstract.

3.17 Scharf, Peter M.

- scharfpm7@gmail.com
- President, The Sanskrit Library; Visiting Professor, IIT Bombay

Creative and intelligent use of linguistic, textual, and bibliographic information to enhance interlinked access to lexical, textual, and image data

It is obvious to participants in the SARIT workshop on the future of digital texts in South Asian studies that we are in the midst of a media transition. Just as there was a transition in the primary mode of transmission of knowledge from oral to written, and from written to printed, we are now undergoing a transition from the printed to digital medium. What is not obvious is how the new digital medium liberates us from the conventions appropriate for the written and print media dictated primarily by visual factors, and how to maximize the potentialities of the digital medium by utilizing linguistic, text-structural, and bibliographic information. Clarification of encoding principles resolves on linguistically precise character encoding for functions such as searching, morphological identification, and parsing. Clear delineation of data-entry and display functions from linguistic processing grants freedom of these functions to conform to human efficiency considerations and user preferences. Judicious use of Text-Encoding Initiative (TEI) markup likewise separates XML text markup from optional display formats and permits interlinking of text with lexical, linguistic, related textual and bibliographic resources on the one hand, and with images on the other. The integration of image analysis software, such as OCR software, with textual and bibliographic information permits the development of approximate image finding

aids by automated methods, and precise finding aids by including human supervision.

The Sanskrit Library (http://sanskritlibrary.org) has developed linguistically precise phonetic encodings for Sanskrit, revised the Unicode Standard to include Vedic characters, and developed comprehensive transcoding software for interchange between linguistic processing, data-entry encodings, and standard Romanization or Indic script Unicode display. Texts are linked to morphological analysis software, and digital lexical sources, and in exemplary distributed collaboration, with the Sanskrit Heritage parser. An integrated dictionary interface permits lookup in some forty different lexical sources including the major bi-lingual and monolingual dictionaries as well as little known and under-utilized specialized dictionaries. The Sanskrit Library has also created a pipeline for digital cataloguing of Sanskrit manuscripts including a template that incorporates the standards of the American Committee for South Asian Manuscripts (ACSAM) and conforms to TEI manuscript guidelines. Manuscript passages are optionally displayed in text-structure or manuscript format and are linked to corresponding searchable digital texts and to manuscript images. A comprehensive catalogue currently contains 160 entries for Sanskrit manuscripts at Brown University and corresponding manuscripts at the University of Pennsylvania and 1800 draft entries for Sanskrit manuscripts at Harvard University. A manual interface allowed association of passages in manuscript images with corresponding digital texts for the Brown and Penn mss. Newly developed text-image alignment software utilizes dirty OCR along with textual and bibliographic parameters supplied by catalogue entries to estimate the location of passages automatically with surprising accuracy.

3.18 Shimoda, Masahiro

- Indian Philosophy and Buddhist Studies // Center for Evolving Humanities
- Graduate School of Humanities and Sociology
- Tokyo University

Prof. Masahiro Shimoda will chair the panel History through Indic Texts.

3.19 Tomabechi, Toru

- International Institute for Digital Humanities, Tokyo
- toru.tomabechi@nifty.com

TEI Markup of Abhayākaragupta's Āmnāyamañjarī: An Attempt to Create an "Open Research Note" for the Study of Late Indian Buddhism

A text markup project may be conceived and designed in two different directions. The one direction is to create a sizable repository of multiple texts with basic (mainly structural) markup. The other is a type of "deep markup" project which aims at accumulating knowledge related to a particular work using e-text as information container. In this paper, we will report on the concept, objectives, and current status of our text markup project to create an "e-text-cum-philological-database" through deep TEI markup.

Since 2010, the "Vikramaśīla Project" (VP), funded with a JSPS grant and led by Prof. Taiken Kyuma (Mie University), has been putting together efforts of specialists in late Indian Buddhism to shed light on the relationship between Tantric and Non-Tantric Buddhist doctrines by laying focus on the works of authors associated with the Vikramaśīla monastery. One of those authors of special importance for the VP is Abhayākaragupta (11-12th c.), whose works are known for their richness in information. His magnum opus, the Āmnāyamañjarī (ĀM), is a commentary on the Sampuṭodbhavatantra and may be qualified "encyclopedic" in many respects. The VP has chosen the ĀM as platform to create an electronic research note on late Indian Buddhism. The research note has been designed as a deep-marked up TEI document, which is to be not only shared among the project members, but also made publicly available online. For our purposes, the ĀM is an excellent choice as material for several reasons. The text contains a large number of quotations from wide range of textual sources. In addition to explicit quotations, Abhayākaragupta also adopts, rather freely and often silently, many passages from works by his forerunners. Furthermore, he frequently refers to other texts of his own composition. Such references to external sources make the ĀM a very rich repository of historico-philological information. Reflecting Abhayākaragupta's wide and deep erudition, the contents of the text also cover a wide range of subjects well beyond the boundary of a mere Tantric commentary. This latter character of the ĀM renders the text itself a vast subject- and terminology-inventory quite useful for the study of late Indian Buddhism.

By encoding the philological, doctrinal and lexical information contained in the ĀM, we have been trying to explore the potential of the TEI-compliant textual markup for the study of Indian Buddhism and to, at the same time, probe into both the strength and the limitation of the current TEI guidelines. In the course of the collaborative work among the project members, we encountered a number of questions to be addressed: What is the best way to structuralize the document? – Which TEI element is appropriate for a particular information in the text? – How the collaboration should be organized? – How the result is to be shared? – and so forth. This paper is a (pre-)interim

report of our attempt to create an electronic research note which is "open" in several senses – by open collaboration, following open standards, for open access and designed as a platform of open-end accumulation of knowledge. Critical comments, advices, and other inputs from the experienced participants of the Workshop are sincerely welcome.

3.20 Wujastyk, Dominik

- wujastyk@gmail.com
- Singhmar Chair in Classical Indian Society and Polity, Department of History and Classics, University of Alberta, Canada

What do users want from SARIT in future?

This presentation will discuss the place of e-texts in contemporary scholarship, and the possibilities that e-texts may afford our successors. It will also interrogate the place of SARIT in this scheme, and why encoding standards are of central importance.

The SARIT library has matured substantially in the last five years. There are more texts, more highly developed guidelines for how texts are encoded, more sophisticated search facilities and a reimplementation of the whole platform on a more robust software basis. The direction of these developments has been driven principally by the views and requirements of the SARIT team themselves. This is good, and as it should be. Yet there are some consequences. SARIT has not yet emerged as the go-to service for Indic e-texts. That position is still held by the GRETIL service, in spite of its technical and qualitative inferiority. The SARIT project has not been as successful as it could be in changing the public perception amongst Indologists regarding the virtues of textual integrity, reliability, version control, and responsibility. Most scholars are aware of the value of a well-edited printed edition of an Indic text, but that same attitude has not yet become widespread where electronic texts are concerned. Furthermore, SARIT has not been successful in broadcasting updates to either its growing content or its developing technical features. SARIT has not consulted widely about the features that the general indological public most want from a library of electronic texts.

This presentation will address the strengths of the SARIT library as it exists in spring 2017, and present the results of a public user-consultation about directions for future development.

3.21 Zotter, Christof and Manik Bajracharya

- christof.zotter@adw.uni-heidelberg.de
- Documents on the History of Religion and Law of Pre-modern Nepal, Heidelberg Academy of Science and Humanities

See Bajracharya, Manik and Christof Zotter for the abstract.

3.22 Software demonstrations

This module is intended for the presentation of practical aspects mentioned in the talks.

1. Yigal Bronner: Hands-on presentation of the Panditproject.org database
2. Amba Kulkarni: Demonstration of Saṁsādhanī: A platform of Sanskrit Computational Tools
3. Stefan Baums: Research Environment for Ancient Documents (READ)
4. Peter M. Scharf: Presentation of the Sanskrit Library meter identification tool and the integrated dictionary interface
5. Patrick McAllister: Presentation of SARIT's workflow for editing and displaying Indic XML documents

4 Links

- https://asiabeyondboundaries.org/about/: Homepage 'Beyond Boundaries: Religion, Region, Language and the State' (see Dániel Balogh)
- http://east.uni-hd.de: Epistemology and Argumentation in South Asia and Tibet (see Birgit Kellner)
- http://gandhari.org: Gāndhārī Language and Literature (see Stefan Baums)
- http://github.com/chchch/upama: *upama*, library to compare Sanskrit TEI XML files and generate an apparatus (see Charles Li)
- http://historyandclassics.ualberta.ca/: Department of History and Classics, University of Alberta, Canada (see Dominik Wujastyk)
- http://panditproject.org: Prosopographical Database for Indic texts (see Yigal Bronner)
- http://saktumiva.org/wiki:dravyasamuddesa:start: Edition "The Dravyasamuddeśa of Bhartṛhari" (see Charles Li)
- http://sanskritlibrary.org: The Sanskrit Library (see Peter M. Scharf)
- http://sarit.indology.info: SARIT — Search and Retrieval of Indic Texts
- http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/nepal/index.de.html: Documents on the History of Religion and Law of Pre-modern Nepal (see Bajracharya, Manik and Christof Zotter)
- http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/nepal/publ_docs.en.html: Published documents on the history of religion and law of premodern Nepal (see Bajracharya, Manik and Christof Zotter)
- http://www.ikga.oeaw.ac.at/Mainpage: Homepage of the Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences
- http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00239: Deep learning (see Oliver Hellwig)

- http://www.r2d3.us/visual-intro-to-machine-learning-part-1/: Introduction to machine learning (see Oliver Hellwig)
- http://www.stockholmuniversitypress.se/site/books/10.16993/baj/: P. Robinson, "The Digital Revolution in Scholarly Editing." (see Philipp Maas)
- https://gandhari.org/blog/?p=251: Blog post about the "Research Environment for Ancient Documents" (see Stefan Baums)
- https://github.com/readsoftware/read: Software repository for the "Research Environment for Ancient Documents" (see Stefan Baums)
- https://projecteuclid.org/download/pdfview_1/euclid.ss/1089808279: Introduction to graphical models (see Oliver Hellwig)