Semiannual report on the NEH-DFG Project "SARIT: Enriching Digital Collections in Indology"

November 2013

Columbia University subproject (Sheldon Pollock and Andrew Ollett)

Introduction. In this first report we will discuss our progress towards three overarching goals:

1. the production of high-quality TEI versions of Sanskrit texts;
2. the production of an online program for interacting with these texts; and
3. the production of a prosopographical database for Sanskrit authors.

This project is funded through an NEH-DFG Bilaterial Digital Humanities Grant. There are thus two branches of the project: the Columbia University branch (CU), led by Sheldon Pollock, and the Heidelberg University branch (HD), led by Birgit Kellner. The NEH/DFG award was announced on May 8, 2013, and the priorities in the following months were hiring project personnel and coordinating development between CU and HD.

Personnel. Andrew Ollett began immediately as the Assistant Director for CU. On May 21, Jay Ramesh,a CU graduate student, was hired as a quarter-time Research Assistant for CU. At the same time, Shiva Subramaniam, another student, was hired to work on the project, but he has been paid from Pollock's Mellon grant and not from the NEH grant. Concurrently, HD began a search for an Assistant Director, and filled the position on October 4, 2013 with Liudmila Olalde.

Organization and communication. Ollett started a "Wikischolars" page at CU for all project participants on May 5. It contains guidelines for TEI encoding—representing the practices followed at CU and HD—as well as notes on specific projects, agendas and minutes for project meetings (see below), and resources for the project's technical, intellectual, and legal aspects. The Wikischolars page was meant to supplement and archive the discussion on the SARIT-development e-mail list, which Dominik Wujastyk had set up before the NEH/DFG project. Currently the Wikischolars page has 14 members, including CU and HD project personnel, members of SARIT's advisory committee, and participants in the "Abhinavabhāratī Online" subproject at CU.

CU and HD planned to have meetings every few months at which we would discuss the major issues facing the SARIT project at the time. The first meeting occurred on July 8, 2013, over Skype. Participants included Ollett (CU) and Kellner (HD), as well as members of SARIT's advisory committee (Wujastyk, Patrick Olivelle), SARIT's current webmaster (Patrick McAllister), and a representative of the Heidelberg Research Architecture (Jens Østergaard Petersen). The topic was "Coordinating SARIT development between New York and Heidelberg." The minutes are attached. The major points of discussion were:

- Development of a new online interface for SARIT. The Heidelberg Research Architecture (HRA) would, pending approval, begin development of an online interface for TEI-XML files using standoff markup. Some questions remained about standoff markup, the timeframe of development, and who would be responsible for the costs of development.

- Sharing data. It was agreed to use a combination of Dropbox and Git to share project data: the former because it is easy to use, and the latter because it has more sophisticated version control and conflict resolution.
- Schedule. We discussed the need for individual project schedules (for CU and HD) and the need to announce the SARIT project and its progress at scholarly conferences and in publications.
- Details regarding project communication and future meetings were also discussed.

Ollett will meet with the HD team in January 2014 to discuss the next set of issues, related to the encoding of SARIT texts in TEI-XML.

Text production. One of the top priorities for this project is the production of high-quality machine-readable versions of important Sanskrit texts. The focus will be on three areas, as per the proposal:

- CU will produce texts relating to the theory of Indian theater, and in particular the *Abhinavabhāratī* of Abhinavagupta. (See below.)
- CU will also produce texts from the philosophical school of *Mīmāṃsā*, which is concerned with the systematic interpretation of Vedic texts.
- HD will produce texts from Buddhist philosophy, and in particular the work of Dharmakīrti and his tradition of Buddhist epistemology.

Each branch of the project sends out materials to a vendor in India for double-keyboarding. These double-keyboarded texts have a limited amount of XML markup. CU and HD need to edit these texts such that they conform to the TEI standards and to SARIT's own standards for XML markup. When this editing is complete, the texts are made available to the public on SARIT. In broad outline text production works like this:

Double-keyboarding → TEI Encoding → Publication on SARIT

Since the beginning of the project in May 2013, CU has received the following double-keyboarded texts:

Name of text	*Date received*	*Page count*	*Size in Kb*
Kāvyalakṣaṇa	June 12	287	1229
Śivārkamaṇidīpikā	September 17	1094	8704
Bṛhatī and *Pañcikā* (in progress)	---	407	1624

The cost for the first two texts was $1,486.10. The last two texts will be included on the next invoice.

Since the project started, CU has been working on adding TEI encoding to texts that we have already received from the vendor. Ollett trained Subramaniam and Ramesh in the encoding procedures during two meetings in July 2013. Since then, Ollett, Subramaniam, and Ramesh have been adding encoding to two texts, the *Abhinavabhāratī* and the *Nāṭyaśāstra* (on which the *Abhinavabhāratī* is a commentary). Progress has been slow, in part because Ramesh and Subramaniam are completely new to TEI-XML, but approximately 7 chapters (out of 37) are completely finished, and many of the rest have undergone preliminary formatting. Project participants have access to a spreadsheet that records the progress made on the *Abhinavabhāratī*. Ollett has also begun work on the *Kāvyalakṣaṇa* and the *Tantravārttika*. Since the goals for SARIT's encoding have not yet been completely agreed upon—this is the topic of the upcoming project meeting in January—CU has pursued the following general goals, in order of importance:

> **Conformance with the TEI P5 Guidelines**. Conformance with standards is extremely important to the usability and sustainability of the data that the SARIT project produces. Thus all of the SARIT files are encoded according to the TEI standards. This means following the TEI guidelines in the process of adding the encoding, and it also means validating the file against a schema when the encoding is completed.
>
> **Exhaustive structural markup**. Each file is encoded in such a way as to reflect the structure of the text. This includes marking up textual divisions (books, chapters, sections, etc.) as well as paragraph elements and verse elements. The markup is "exhaustive" because the entire text must be accommodated within the XML structure. Adding structural markup in many cases also means assigning unique IDs to structural elements; these IDs will play an important role in a Canonical Reference System that will be designed and implemented at a later date.
>
> **Text-critical markup**. To the fullest possible extent, text-critical information in the source text is to be represented in the SARIT version using TEI's "Critical Apparatus" module.
>
> **Semantic markup**. For the *Abhinavabhāratī* in particular, CU has been putting semantic markup into the SARIT version of the text. This includes the names of persons, places, and works, and cross-references within and outside of the text.

The completed chapters of the *Abhinavabhāratī* exhibit all of these features, and hence they have been submitted to the HRA as model texts. Many of these encoding procedures will need to be revised in light of future conversations about the new SARIT interface (see below); in particular, text-critical markup and semantic markup will probably be moved from the "base text" to a separate "annotation layer." However, CU will continue to add inline markup until the new interface becomes available.

Abhinavabhāratī Online. The *Abhinavabhāratī*, one of the most important texts on the theory and practice of the theater and on the philosophy of aesthetics, is a major focus of the CU project. One of the goals for the period of NEH-DFG funding is to make the *Abhinavabhāratī* available online, and to give researchers the opportunity to add text-critical and other kinds of annotations, resulting in a collaboratively-edited text that will be useful to a wide range of students and scholars. Several important steps have been made in this first six-month period.

(1) Assembling a special interest group. Pollock invited several scholars whose research involves the *Abhinavabhāratī* to form a "special interest group" on the "Abhinavabhāratī Online." They are Ashok Aklujkar (Vancouver), Lyne Bansat-Boudon (Paris), Daniele Cuneo (Cambridge), Elisa Ganser (Paris), and Gary Tubb (Chicago). The purpose of this group is to provide feedback on the project over the course of development, and to participate in the collaborative reedition of the *Abhinavabhāratī* when the new online interface becomes available. This group was formed in the beginning of July 2013.

(2) Obtaining appropriate licenses. In October 2013 Pollock met with the vice chancellor and registrar M S University, Baroda, the institute that currently holds the copyright to the text on which the “Abhinavabhāratī Online” is based, and began the process of getting the permissions necessary to make the text of the *Abhinavabhāratī* available online under a Creative Commons License. Pollock also raised the possibility of further collaboration between CU and the Oriental Institute Baroda, the relevant academic unit at the university, on the project.

(3) Progress towards a TEI text. As noted previously, Ollett, Ramesh, and Subramaniam are adding TEI encoding to the double-keyboarded text of the *Nāṭyaśāstra* and *Abhinavabhāratī*. For seven chapters so far, we have encoded the critical apparatus, references within the text (including references from the commentary to the root-text, called *pratīka*s in Sanskrit), references to other texts, and proper names, and we have introduced a stable reference system based on XML IDs. The finished chapters validate against a standard TEI P5 schema. Ollett has also produced a number of stylesheets to display the TEI text in a web browser. This work will continue throughout the next reporting period, that is, until May 2014.

(4) Progress toward a collaborative online editor. HRA has started to develop an online editor using the sample TEI files of the *Abhinavabhāratī* provided by CU. Collaborative editing is envisioned as a general feature of SARIT under the new interface (see below), but it is especially important for the “Abhinavabhāratī Online” subproject.

SARIT Interface Development. HRA has taken on the development of several features which are crucial to the new online platform that HD and CU both envision for SARIT. This is an undertaking of HRA which will therefore *not* require NEH or DFG funds: access to the HRA is one of the advantages that HD brings to the joint project. One of these features is the aforementioned collaborative editing software. CU is not directly involved with the development of this software, but Ollett has been in conversation with Peterson about its advantages and limitations and its impact on how CU will produce TEI texts in the future. Currently HRA is working on a proof-of-concept for moving from inline markup (the way in which CU and HD currently mark up their texts) to standoff markup (according to which the base text, text-critical markup, and semantic markup all stored separatedly).

Database development. One of the goals included in the project proposal was making the database produced for the “Sanskrit Knowledge Systems on the Eve of Colonialism” project publicly accessible. (The database is currently on a password-protected server at King’s College, London.) Pollock and Ollett retrieved the data from the King’s College server and fixed a number of errors introduced during an encoding conversion in the 2000s. The data—comprising more than 1000 persons and 1300 texts—has been exported and awaits a more modern and more interactive framework.

After the submission of the proposal, Pollock and Ollett became aware of an initiative led by Yigal Bronner (Jerusalem) to build a new prosopographical database focused on the life and works of Appayya Dīkṣita, an important intellectual of the 16th century (called “Appayya Dīkṣita’s Works”). Conversations between Bronner and Ollett in August 2013, and between Bronner and Pollock in October 2013, made it clear that Bronner’s database was similar in design and envisioned content to the “Sanskrit Knowledge Systems” database (a project in which Bronner participated). Moreover, the data from the “Sanskrit Knowledge Systems” was extensive, but it is tied up in legacy database software that is not being maintained; “Appayya Dīkṣita’s Works” does not have any data yet, but it has an intuitive and accessible design and benefits from a flexible content management system. It was therefore decided to merge the projects by populating “Appayya Dīkṣita’s Works” (which will be renamed in consideration of its widened scope) with data from the “Sanskrit Knowledge Systems” database. CU will cover the costs of importing the “Sanskrit Knowledge Systems” data from NEH funds.

The merger will provide for the integration of the “Sanskrit Knowledge Systems” data with SARIT, as envisioned in the proposal: once the merged database is populated with entries for persons and works, SARIT will be able to direct references to persons and works within its TEI texts to the prosopographical database.