

6 Data Management Plan

The project assistants will be responsible for data management; an efficient division of responsibilities will be decided upon when the project begins. The project directors and the SARIT board will evaluate the organization, accessibility and security of the data at regular intervals during the grant period, concurrent with internal project review. A change of personnel, either during or after the grant period, will involve reassigning permissions to curate the data (i.e., to make changes to the texts or the bio-bibliographical databases); it should not affect access to the data for other project personnel or for the public.

This project will produce data of three kinds. The first is a web platform onto which the SARIT texts will be loaded and through which they can be publicly accessed (through a variety of search and display functions), along with the associated bio-bibliographical databases. A widely-used and well-documented open-source application like eXist-db, used for building and querying XML databases, can be adapted for this purpose. Project staff will produce tools specific to the display and transformation of Indic texts (for example, transliteration stylesheets). Any modifications that the SARIT project makes to this platform will be made available for free on a code-sharing site such as SourceForge after sufficient testing. Generic tools which result from the cooperation with the HRA will likewise be made available in this form.

Digital texts are the core of SARIT’s functionality, and most of the data produced by this project will be in this form. The texts are encoded in TEI-conformant XML; many will be accompanied by annotation layers, also in XML format. The encoding standards used by project staff will be published to facilitate third-party contributions. In addition to text-data, the XML files will include structured metadata. Production of these files will take place at both Columbia University and the University of Heidelberg: to ensure consistency and avoid duplication, the files will be stored at a single location in Heidelberg and accessed and edited through a version-control system. Various such systems are available; Git is among the most powerful and has been used by SARIT personnel in the past. When the XML files are completed, they will be loaded onto a database on the SARIT server and thereby “published” on SARIT. They can be both used through the web platform, as well as downloaded for offline use.

The funding and maintenance of the SARIT server will be guaranteed by the Cluster “Asia and Europe in a Global Context.” Regular backups of both the in-process and completed text data will be made to servers at Columbia and Heidelberg. Major revisions to the text data after the date of publication (for example, annotations on the SARIT website) will be noted in the revision history in the metadata for each file; minor revisions are accessible through the versioning system.

The XML files will be the source from which other kinds of data are generated. Human-readable texts (in HTML and PDF formats) will be uploaded at the time that the corresponding XML files are loaded onto the SARIT server; these texts will also be periodically backed up. The web application will generate a variety of data in response to user queries (keyword searches, word-frequency lists, morphological analysis); these data will not be permanently stored. Other kinds of data, such as images of manuscript folios, might be linked to the texts (and either accessible through SARIT or downloadable as an archive), although no provisions for such multimedia presentations are planned for the web application. These data would be stored in the same way as the texts.

Users will be able to sign in and annotate the SARIT texts through a simple interface; this is the only feature for which a registration and approval process is necessary. These will be “stand-off” annotations, stored in an XML layer separate from the text itself; as soon as they are added they will be publicly

visible. Part of the HRA's agenda is to develop the technology for this interactive annotation layer. The curators of the text data—that is, the Project Directors *ex officio* and those scholars to whom they grant curatorial privileges—will review the annotations regularly; each text will have its own review schedule.

The databases to which SARIT will be linked will be located on servers at their respective institutions (EAST at the Cluster “Asia and Europe in a Global Context” at Heidelberg, and SKSEC at Columbia University Libraries) and maintained by staff at those institutions. Bibliographical data in these databases will conform to the MODS standard, and is integrated within a prosopographical database in the case of EAST (PostgreSQL, web interface Django). Compatibility between EAST and SKSEC, and between each of them and SARIT, and with future digital texts projects in Indology requires the adoption of a standard system of reference, which will be produced and disseminated during the grant period.