White Paper Report

Report ID: 100485
Application Number: HG-50022-10
Project Director: Peter Scharf (peter.scharf@univ-paris-diderot.fr)
Institution: Maharishi University of Management Research Institute
Reporting Period: 7/1/2010-6/30/2014
Report Due: 9/30/2014
Date Submitted: 9/29/2014
White Paper
HG-50022-10
Sanskrit Lexical Sources
Digital Synthesis and Revision

Dr. Peter M. Scharf
Maharishi University of Management Research Institute

29 September 2014
Contents

1 Narrative 1

1.1 Personnel 1

1.1.1 Blaise Pascal research chair and IIT Bombay visiting professorship 1

2 Project activities 2

2.1 Data-entry 2

2.2 Lexical source integration 4

2.2.1 Data-cleansing, markup, and interface development 4

2.2.2 Catalogue 5

3 Accomplishments 6

4 Audiences 8

5 Evaluation 9

6 Continuation of the project 9

7 Long term impact 9

8 Grant products 10

Appendices 11

A Presentations 12

B Publications 15

C Screen shots 16
List of Tables

1.1 Sanskrit-English Dictionaries . . . . . . . . . . . . . . . . . . . . 7
1.2 English-Sanskrit Dictionaries . . . . . . . . . . . . . . . . . . . 7
1.3 Sanskrit-French Dictionaries . . . . . . . . . . . . . . . . . . . . 8
1.4 Sanskrit-German Dictionaries . . . . . . . . . . . . . . . . . . . 8
1.5 Sanskrit-Latin Dictionaries . . . . . . . . . . . . . . . . . . . . . 8
1.6 Sanskrit-Sanskrit Dictionaries . . . . . . . . . . . . . . . . . . . 8
1.7 Specialized Dictionaries . . . . . . . . . . . . . . . . . . . . . . 8
List of Figures

C.1 The Sanskrit Library integrated dictionary interface . . . . . . . . . 16
C.2 Cologne Digital Sanskrit Lexicon project home page . . . . . . . . . 17
Narrative

1 Introduction

The stated aim of the project was to synthesize, extend, revise, and improve the principal lexical reference works of Sanskrit, the primary culture-bearing language of India, and one of the world’s richest, and to provide access to them at the University of Cologne and in the Sanskrit Library. The project integrated existing digitized major Sanskrit lexical reference works and extended them by digitizing and integrating subsequently published additions and corrections, specialized dictionaries, indigenous Indian monolingual dictionaries, and traditional linguistic analyses. Furthermore, the project improved the utility and accessibility of the lexical resources by linking them with digitized Sanskrit texts.

1.1 Personnel

The collaborative German/U.S. project was managed by Dr. Peter Scharf, a Sanskrit scholar at Maharishi University of Management Research Institute (MUMRI) and Thomas Malten, the director of the Cologne Digital Sanskrit Lexicon project (CDSL) at the University of Cologne. The U.S. team included Ralph Bunker, a software engineer at MUMRI and Jim Funderburk, a retired mathematician and amateur Sanskritist who had worked with Scharf previously. He volunteered considerable time to the project including the development and management of the CDSL website.

1.1.1 Blaise Pascal research chair and IIT Bombay visiting professorship

During the period 1 February 2012 – 31 July 2013, Scharf spent a year as laureate of a prestigious Chaire Internationale de Recherche Blaise Pascal funded by the State of France and the Ile de France region. The award involved a year of research housed in the History of Linguistics laboratory at the Université Paris Diderot and provided a handsome research account. He was invited as a visiting professor to the Indian Institute of Technology Bombay (IIT Bombay) for a semester December 2012 – April 2013. While the award and invitation drew Scharf away from
MUMRI, they permitted him to engage unforeseen resources in the project. He negotiated an agreement with the Indian Institute of Technology Bombay to hire two Sanskrit post-doctoral research associates for the year and one computer science post-doctoral research associate for a few months to assist in the Sanskrit lexical sources project. The computer scientist, Pawan Goyal, was also engaged for three months at the French national computer science research center (INRIA) during October–November 2012 and May 2013 directly in France. A second post-doctoral associate in Sanskrit was hired in Europe for a month.

2 Project activities

2.1 Data-entry

The project activities were divided between the German and U.S. partners as follows: the German partner digitized the lexical sources; the U.S. team converted character data-entry to XML and produced HTML displays. Digitization of the lexical sources began with the production of digital images of the dictionaries. The images were sent to Auroracana Exports for double-keyed data-entry and double-blind error correction. Character markup was supplied for visually distinct textual properties such as boldface headwords, and italicized citations. Remaining doubts, where the scanned images or the printed text itself were unclear, and where headwords between dictionaries do not correspond, were examined and resolved by the U.S. team and its Sanskrit assistants at IIT Bombay. Funderburk, Goyal, and Scharf wrote computer programs that made rich use of regular expressions to identify headwords and their definitions, and to determine their corresponding positions in the digital images of the sources. From this information, Funderburk and Goyal constructed simple XML structures for each lexical source, with one record per headword. This XML form was read into a simple SQLite database table. Web displays were constructed as PHP programs to query the SQLite database by headword and to link entry displays to the scanned images.

Dr. Thomas Malten, our collaborator at the University of Cologne, was in charge of data-entry for the project. As stated in the proposal section IIB describing the scope of the project (p. 10), the German partner would complete the data-entry for the bilingual dictionaries for which only digital images are available listed in section IIA1 (p. 7) and provided with full bibliographic details and extents in Appendix B (pp. 59–60). The German partner would also produce digital images and machine-readable data files of the lexical sources listed in Appendix C (pp. 61–70).

Data-entry proceeded in two stages: (1) double-keying the text, (2) a repeated cycle of comparing the two versions, marking differences and returning the marked
version, with no indication of what the other version produced, to each data-entry clerk for correction. This double-blind error correction process adopted for the second stage is a revision over the previous error-correction process which allowed the data-entry clerk to see both versions of the text, her own and the one produced by the other clerk. The previous process resulted in each clerk repeatedly favoring her own error. The new process forced each clerk to reexamine the source independently to attempt to correct her error. Although the new process took longer, it produced cleaner, more certain results.

Data entry for the six volumes of Ghatage et al. included in the project was completed early in 2011. In the meantime two additional volumes of the work not originally included in the project were published. Data-entry of these works was completed by the end of October 2011. The Sanskrit Library contracted an agreement with the University of Hyderabad to share digital resources and thereby obtained a copy of the *Amarakosa*. Likewise, the Sanskrit Library contracted an agreement with Osmania University’s Sanskrit Akademy to obtain a copy of the *Vācaspatya*. While obtention of the latter lexical source allowed the project to integrate the dictionary, the copy was found to include systematic errors, so the project directors decided to proceed with data-entry anew, and an independent digital edition was produced.

Due to the extra time required to include the additional new volumes of the major lexicon Ghatage et al., the project was unable to digitize eight of the specialized sources listed in Appendix C, section IA2 (pp. 62–64). These consist of the following works:


We hope to digitize these in a future project.

### 2.2 Lexical source integration

#### 2.2.1 Data-cleansing, markup, and interface development

Bunker, the principal computer scientist engaged in the project located at MUMRI, created a local version of the Sanskrit Library website and developed an interactive HTML interface that allowed assistants, wherever they were located and whatever was the quality of their internet connection, to access dictionary images and write files from the HTML interface on their private computers. We created local workstations that provided all the facilities of the Sanskrit Library web server. Utilizing the environment created by Bunker, Goyal developed HTML pages to validate headword coordination between newly digitized dictionaries with the headwords in MW.

During 2012-2014, the Sanskrit post-doctoral associates at IIT Bombay utilized local workstations to integrate the supplement to Monier Williams’ *A Sanskrit-English Dictionary* (MW) with the main body of the text, and create an index of the headwords in the *Vācaspatya* to digital images of the work. During the past
year, one corrected some two thousand errors in headwords in Ghatage et al. and mapped them to headwords in MW. A post-doctoral assistant at IIT Bombay examined 785 unclear printed characters and corrected them in the digital edition. Funderburk created similar displays for the correction of doubts in other dictionaries using which a post-doctoral assistant similarly corrected 500 unclear printed characters in Apte’s English-Sanskrit dictionary, 700 in Apte’s Sanskrit-English dictionary, and 200 in Rādhākāntadeva’s Sabdakalpadruma. The other Sanskrit assistant mapped headwords in Bhaṭṭācārya’s Vācaspatya to headwords in Monier-Williams’s Sanskrit-English dictionary.

Funderburk and Scharf developed the list display under the current project in 2012. Funderburk and Bunker also developed displays for mobile devices.

In addition to developing HTML interfaces for the Sanskrit assistants to carry out the above tasks, Goyal completed a mapping of Huet’s Sanskrit-French dictionary with the MW headwords and created initial HTML displays of several dictionaries. These include the following:


Scharf and Goyal designed the integrated dictionary interface webpage and revised the list display, and Goyal incorporated the new displays with the old in this interface. Part of the revision of the list display included refining the designation of homophonous headwords in MW to capture homophones that resulted from the addition of feminine suffixes to adjectives and to nouns that occur in all three genders. Bunker incorporated the new developments into the Sanskrit Library’s development server at http://sanskrit1d.ccv.brown.edu under Reference, Integrated Sanskrit dictionary, and its new server at http://sanskritlibrary.com.

2.2.2 Catalogue

A principle of the current project is to allow access to original sources rather than to merge diverse sources in a reedited synthesis. In line with this principle, proper
identification of lexical sources is necessary. Scharf developed a TEI-conformant XML template for cataloguing the digital lexical sources developed in the project and created catalogue entries for the dictionaries currently available on the Sanskrit Library website. Funderburk provides access to digitizations of the title pages and front matter under documentation at the Cologne site.

3 Accomplishments

The project now incorporates digitizations of the following classes of lexical resources:

- 11 bilingual Sanskrit-English dictionaries
- 3 English-Sanskrit dictionaries
- 2 Sanskrit-French dictionaries
- 5 Sanskrit-German dictionaries
- 1 Sanskrit-Latin dictionary
- 2 Sanskrit-Sanskrit dictionaries
- 10 specialized dictionaries.

These resources are listed in tables 1.1–1.7.

Each of these resources has four displays on the University of Cologne website: basic, list, advanced, and mobile friendly, each with a choice of input and display options. Input can be entered in one of three Romanizations: Sanskrit Library phonetic, Kyoto-Harvard, and ITrans. One can choose to display output in Devanagari or Roman Unicode as well as in one of these three. The basic display presents the headword and its definition alone. The list display presents a list of headwords at the left which shows the headword in context. For dictionaries such as MW which present headwords in a hierarchy under their derivational root, one can choose between a hierarchical list and an alphabetical list. The advanced display permits searching for words that begin with, end with, or contain a string as well as match it exactly. The advanced display also permits one to search the body of the text and thus serves as a reverse dictionary, permitting one to find all the Sanskrit headwords that contain an English word, for example. Finally the Cologne site offers a display amenable to mobile devices. The Cologne site also makes a collection of materials available for download. These materials include the original data-entry files, their transcoding to UTF8, a corrected text file, scanned images of the work, an XML conversion of the corrected data-entry with its document type definition (DTD), software to create the XML file from the corrected text, and materials to construct an offline version.
Lexical resources are also incorporated into the integrated display on the Sanskrit Library website. Here one can enter the headword once and look it up in any lexical resource just by clicking the resource in the list. Pressing the carriage return looks it up in the dictionary selected to serve as the default. One can also select from a broader range of input options, including several Romanizations as well as Roman and Devanagari Unicode, and display options including seven major Indic scripts. Lexical resources are presented in the list display. The list of headwords can be browsed in either direction. Clicking another headword in the list fills that word in the entry field and looks it up in the last used dictionary. At the Sanskrit Library, the integrated dictionary is linked to a morphological analyzer linked to more than a hundred Sanskrit texts. When one clicks a word in texts that have analyzed interword phonetic changes, the analyzer presents possible stems and corresponding grammatical identifications of the form. Clicking the stem, opens the integrated dictionary page, fills the stem in the integrated dictionary entry field and looks it up in the default dictionary.

Table 1.1: Sanskrit-English Dictionaries

<table>
<thead>
<tr>
<th>ID</th>
<th>date</th>
<th>Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIL</td>
<td>1832</td>
<td>Wilson Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>YAT</td>
<td>1846</td>
<td>Yates Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>GST</td>
<td>1856</td>
<td>Goldstücker Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>BEN</td>
<td>1866</td>
<td>Benfey Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>MW72</td>
<td>1872</td>
<td>Monier-Williams Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>AP90</td>
<td>1890</td>
<td>Apte Practical Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>CAE</td>
<td>1891</td>
<td>Cappeller Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>MD</td>
<td>1893</td>
<td>Macdonell Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>MW</td>
<td>1899</td>
<td>Monier-Williams Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>SHS</td>
<td>1900</td>
<td>Shabda-Sagara Sanskrit-English Dictionary</td>
</tr>
<tr>
<td>BHS</td>
<td>1953</td>
<td>Edgerton Buddhist Hybrid Sanskrit Dictionary</td>
</tr>
</tbody>
</table>

Table 1.2: English-Sanskrit Dictionaries

<table>
<thead>
<tr>
<th>ID</th>
<th>date</th>
<th>Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>MWE</td>
<td>1851</td>
<td>Monier-Williams English-Sanskrit Dictionary</td>
</tr>
<tr>
<td>BOR</td>
<td>1877</td>
<td>Borooah English-Sanskrit Dictionary</td>
</tr>
<tr>
<td>AE</td>
<td>1884</td>
<td>Apte Student’s English-Sanskrit Dictionary</td>
</tr>
</tbody>
</table>
4 Audiences

Sanskrit texts constitute an enormous body of knowledge in diverse domains that is grossly underrepresented in the Western academic community. Access to this body of knowledge, which is currently accessible only to highly-trained specialists, will
be profoundly valuable to students, scholars, and the wider public concerned with such fields as historical and general linguistics, philosophy and religious studies, history of science and mathematics, pharmacology and medicine, and general history and literature of South Asia.

5 Evaluation

Each of the webpages that displays the digital dictionaries contains a link to a form that permits users to submit corrections. Corrections are vetted by Funderburk, Scharf, and other volunteers and incorporated into data-revisions. Comments by users regarding presentation are accepted by Funderburk and Scharf and incorporated into design revisions. Such comments on early versions of the sites already led to the development of the various displays, and input and display options.

6 Continuation of the project

The Sanksrit Library plans to continue the project to complete the digitization of the specialized dictionaries that were not able to be included in the present project listed above in §2.1 and to include traditional thesauri. These plans were included in a proposal to digitize Sanskrit lexical and grammatical sources just submitted to the NEH and DFG.

The same software at the Sanksrit Library website that allows access to lexical resources from texts simultaneously provides the tools and environment for incremental production of a comprehensive digital Sanskrit lexicon linked to textual instances. While morphological and prosodic variants of headwords in existing lexical sources are identified in texts, linguistic software can flag word forms missing from existing lexical resources, suggest parses, and compile lists of suggested lexical bases linked to their contexts. These lists will allow researchers subsequently to locate terms to be defined in their contexts easily. In the future, qualified scholars will be invited to compose and submit lexical entries via a wiki to the Sanksrit Library board of editors for publication in the gradually expanding lexicon.

7 Long term impact

We expect that the website will provide the most current and most comprehensive lexical information about Sanksrit in the world to enrich knowledge and understanding of the many fields to which Sanskrit scholarship has contributed over several millennia.
8 Grant products

The principal products of the project are the lexical resources listed in tables 1.1–1.7 in §3 on the websites of the CDSL at the University of Cologne shown in Figure C.2 (http://www.sanskrit-lexicon.uni-koeln.de) and of the Sanskrit Library shown in Figure C.1 (sanskritlibrary.org). The project is described on the Sanskrit Library website in a page linked to the project title, ‘Sanskrit lexical sources: digital synthesis and revision,’ under ‘Projects.’ The Cologne CDSL home page provides documentation under the link so named. In addition Scharf delivered the project-related invited lectures and conference presentations listed in Appendix A and produced the project-related publications listed in Appendix B.
Appendices
Presentations

1. “Preserving knowledge through media transitions: ushering the heritage of India into the digital age.” Year of India series, Brown University, 11 November 2009.


5. “Developing computational resources to conduct linguistic research on Sanskrit and provide digital access to Sanskrit texts and manuscripts.” Joint seminar with Amba Kulkarni, Śrī Śaṅkarācārya University of Sanskrit, Kalady, Kerala, 27 January 2010.

6. “Preserving knowledge through media transitions: ushering the heritage of India into the digital age.” University of Madras, 1 February 2010.


2010, University of Massachusetts, Dartmouth, Mass.


14. “Sanskrit computational linguistics.” Center for Advanced Study in Sanskrit (CASS), Pune University, Pune, 15 January 2011.


21–22 mai 2013, Centre d’Études Supérieures de la Renaissance (CESR), Université François-Rabelais, Tours, 21 May 2013
25. “Non-arbitrary lexical tagging.” International Blaise Pascal Research Chair, lecture 8. Université Paris Diderot, 30 May 2013
Publications


Figure C.1: The Sanskrit Library integrated dictionary interface
Figure C.2: Cologne Digital Sanskrit Lexicon project home page