International
Digital Sanskrit Library
Integration
The proposed project aims to enhance online access to the cultural heritage of India by creating a globally distributed, internet-based, digital library in Sanskrit, one of the world’s richest culture-bearing languages. The resulting integrated information system will enrich access to internet-based digital content in Sanskrit located at digital libraries worldwide and thus enable broad use of these collections for research and education. The project integrates currently independent projects to create Sanskrit digital archives, digital lexica, and linguistic software; and to establish text-encoding standards; to enhance ancient and medieval manuscript access; and to develop OCR technology, display software, and Unicode-compliant text-editing software for Devanāgarī script. The easy accessibility of web-based materials is especially significant for less commonly taught languages such as Sanskrit.

I Digital archives of Sanskrit texts

Sanskrit is one of the world’s richest knowledge-bearing and culture-bearing languages. Extant manuscripts in Sanskrit number over 30 million, one hundred times those in Greek and Latin combined. The Digital South Asia Library at the University of Chicago (DSAL)¹ and South Asia Resource Access on the Internet (SARAI)² at Columbia serve as central distribution points for materials related to South Asia including Sanskrit language and literature. Digitized Sanskrit textual materials are available online in two forms: as image files and as machine readable text files.

Machine-readable text

Besides the Sanskrit Library’s own small collection of machine-readable texts and the texts in the databank of Sanskrit grammatical texts that George Cardona has agreed to have the Sanskrit Library host, a significant number of Sanskrit texts have been made available in machine-readable form at other web-based libraries and smaller websites. Machine-readable texts now comprise the fundamental Vedic works and their ancillaries, the epics, and numerous classical works. The brief catalogue below of five of the richest sites enumerates 175 machine-readable texts. The managers of the three richest sites are our collaborators Gippert, Tokunaga, and Freund:

- 81 Vedic texts and 41 Epic and Classical Sanskrit texts, and 6 Buddhist Sanskrit texts at TITUS: Thesaurus Indogermanischer Text- und Sprachmaterialen, managed by Jost Gippert at the University of Frankfurt.³
- 30 texts at Kyoto University under the direction of Munem Tokunaga.⁴
- 29 texts amounting to about 29 megabytes available on CD from Vedic Engineering currently being extended in collaboration with Peter Freund, Librarian, Maharishi University of Management.
- 19 texts at the Indology site besides those from Kyoto and several links to other minor archives.⁵
- 16 texts, 18 stotras, and 13 short recordings by Dale Steinhauser.⁶

The utility of this growing on-line stock of Sanskrit texts will be greatly enhanced by integration with lexical sources and linguistic software.

Image files

By far the greater proportion of online Sanskrit texts is in the form of image files. The Million Book Project has scanned some 186 printed Sanskrit texts.⁷ Manuscript
libraries, such as the University of Pennsylvania, and the Wellcome Library for the History and Understanding of Medicine, are displaying Sanskrit manuscripts in their collections online. Large Sanskrit manuscript libraries in India, such as the Saraswati Mahal Library in Tanjavur, are investigating scanning as an alternative to microfilming manuscripts. The potential volume of scanned text is enormous. While it is advantageous to be able to view these texts online, scanned images lack the accessibility typical of machine-readable text. Digital images of texts are not searchable. They are not accessible to content extraction. They present no more advantages on the internet than they would in printed form except for distant access and scalability. The utility of scanned text is entirely dependent upon the successful development of Devanāgarī OCR technology, such as is being conducted by our collaborator Venugopal Govindaraju at the Center of Excellence for Document Analysis and Recognition at the University of Buffalo under grants from the NSF (see below).

II Text-encoding standards

The utility of digital Sanskrit texts is currently limited by inconsistencies in text-encoding standards. Standard text-encoding schemes for Sanskrit include various schemes based upon scholarly Roman transliteration of Sanskrit, based upon the standard computer keyboard, and based upon common principles in the use of modern Indian scripts. None of these is based upon a one-to-one correspondence between graph and phoneme which would be suitable for linguistic processing. As a result of these multiple bases for standards, over and above the variety of schemes using the same basis, the utility of digitized Sanskrit text suffers many fold from the faults of non-standardization. Some websites require the user to download proprietary fonts to view their digitized text correctly. Some display their text as graphic images sacrificing the benefits of digitized text. Due to the lack of sensible text-encoding standards for Sanskrit, so much of the little technical expertise dedicated to Sanskrit software development has been absorbed in the trivial tasks of (1) transliteration between specific text-encoding schemes and (2) alphabetization and collation for particular text-encoding schemes. The small amount of linguistic software that has been developed suffers from mutual incompatibility and incompatibility with texts and lexica.

Roman transliteration-based standards

Several scholars at the VIIIth World Sanskrit Conference in Vienna established a standard character-encoding scheme utilizing the so-called “upper ASCII” for the representation of classical Sanskrit (CS) and an extended version including characters needed for the Romanization of Vedic and Tamil (CSX). Several fonts, both freely available and commercial, are based upon the CSX character-encoding scheme. Character-encoding schemes based upon Roman transliteration of Sanskrit, besides being incompatible with those based upon Indian scripts, are ill-suited to linguistic processing of Sanskrit. First, they distribute Sanskrit characters randomly among “upper ASCII” slots in order to preserve all the characters used to represent European languages; thus they present no coherent ordering of the characters used for Sanskrit sounds. Secondly, they violate the principle of a one-to-one correspondence between Sanskrit sounds and the characters used to represent them. They represent aspirates and open diphthongs, which are unitary phonemes in Sanskrit, by digraphs (kh, gh, etc. ai, and au). While it is possible to disambiguate the spirant h from aspiration with reference to its phonological context, it is not possible to distinguish between the diphthongs ai, au and the sequences of simple vowels ai, aii.
Keyboard-based standards

A few text-encoding schemes designate a set of characters and character sequences exclusively for the representation of Sanskrit. These include the Harvard-Kyoto scheme which is fairly widely used for data-entry, and the Velthuis transliteration which is often used with TEX. A related system is the ITX standard. Bound as they are to conventions of Roman transliteration, these schemes suffer from the legacies of Roman script for representing aspirates and diphthongs discussed in connection with Roman transliteration-based standards. Velthuis’s scheme entails the non-intuitive measure of using a null string between consecutive vowels a and i or u to disambiguate these sequences of simple vowels from the diphthongs ai and au.

Modern Indian language-based standards

Unicode is burdened with the heritage of the idiosyncratic use of Indian scripts by modern Indian languages. It is based upon the Indian Standard Code for Information Interchange (ISCII) which was designed primarily to represent these modern languages.11 Both Unicode and ISCTI are ill-suited for linguistic processing of Sanskrit because they violate the principle of a one-to-one correspondence between graph and phoneme. They adopt the convention of Devanāgarī script of representing the absence of short a by a graph (virāma) rather than employing a graph to represent it. By extension of the principle of the virāma, the most sophisticated methods to date for forming Devanāgarī glyphs from underlying character sequences require a special character not directly correlated with a phonetic or morphological element of the Sanskrit language in their character-encoding schemes. Unicode and ISCTII also possess duplicate sets of characters for each of the vowels other than short a. This duplication is not linguistically motivated and is unnecessary to represent Sanskrit. While in modern Indian languages it is ambiguous whether or not a is included in the glyph of a preceding consonant, in Sanskrit this ambiguity is absent.

Phonology-based text-encoding

Indigenous Sanskrit linguistics as well as modern linguistics has always been based upon phonology rather than upon the script. A standard text-encoding scheme should be clearly distinguished from fonts used to display Sanskrit and keyboard arrangements used to enter text. The text-encoding scheme developed at Brown University faithfully captures Sanskrit phonology and is thus ideally suited to linguistic applications.12 It improves upon the above schemes by designating character codes exclusively for Sanskrit sounds and sound qualities in the sound catalogues and sound quality descriptions found in classical Indian linguistics. The scheme is easy to convert into all other standard encodings, including Unicode. Such a simple scheme should be adopted as the standard for internal representation, such as text-encoding and linguistic processing software, while external interfaces, such as scanning and display, should be script-based. The Brown team assembled a panel to survey the state of the art in the subject of Sanskrit text-encoding at the joint conference of the ACH and ALLC in 2003, presented its results at the 24th Annual International Conference on South Asian Languages Analysis, November 19-21, 2004 at Stony Brook, and has prepared a book-length manuscript for publication entitled, "Linguistic Issues in Coding Sanskrit."
III Digital Sanskrit lexica

The Cologne Digital Sanskrit Lexicon (CDSL) project, under the direction of Thomas Malton at the University of Cologne, aims at producing a unified database comprising the materials in all major nineteenth-century bilingual Sanskrit dictionaries. At present, the CDSL database includes the complete contents of the large Monier-Williams dictionary of 1899 and Cappeller's lexicon of 1891. By next year, the CDSL plans to complete data-entry of Böhtlingk and Roth's seven-volume Sanskrit Wörterbuch, Böhtlingk's Sanskrit-Wörterbuch in kürzerer Fassung with Schmidt's additions, and Böhtlingk's six-volume collection of Sanskrit sayings. The CDSL further intends to investigate the possibility of digitizing the card files of the Sanskrit dictionary compiled at the Deccan College, Pune, containing some nine million citations.

With support from the Consortium for Language Teaching and Learning, the Sanskrit Library at Brown University has digitized and electronically published W. D. Whitney's tables of Sanskrit verbal stems and derivatives.

Several other digitized lexical sources are presently available or in preparation. Jun Takashima at Tokyo University of Foreign Studies' Research Institute for the Languages and Cultures of Asia and Africa has produced an abridged version of Apte's The Practical Sanskrit-English Dictionary. Jens Braarvig at the University of Oslo and Center for Advanced Study, Oslo is currently involved in the compilation of a digital lexicon for Buddhist Sanskrit texts. The Digital Dictionaries of South Asia project associated with the Digital South Asia Library announces the production of four digital lexica: Monier Williams', Cappeller', Apte's, and Macdonell's A Practical Sanskrit Dictionary of 1929.

While the digitized Sanskrit lexica are useful reference tools in themselves, their most promising attribute is that they allow the development of software applicable to the general corpus of Sanskrit texts.

IV Computational implementation of a Sanskrit production grammar and parser

The value of digitized text is greatly enhanced by integration with morphological and lexical tools. In order to link from words in texts to the relevant dictionary entries it is necessary to be able to refer the inflected form to the lemma. For the modern languages of Western Europe this problem is relatively easily solved. But in languages that possess a rich inflectional morphology, a single word may appear in many surface forms. Phenomena such as morphological suppletion and non-concatenative morphology complicate the problem; the relation between inflected form and lemma may be opaque. For instance, in ancient Greek ἐνενκα is the first person singular aorist active indicative of pherō 'carry'; in Sanskrit paśyati is the third person singular present active indicative of the verbal root drś, and kṣan is the third person plural aorist active indicative of ghas 'eat'.

An additional problem encountered in Sanskrit is that knowledge of phonology is required even to identify the boundaries between words. Contextual sound changes (sandhi) regularly eliminate word boundaries. In many combinations, the final vowel of a preceding word and the initial vowel of a following word are replaced by a single vowel. For example the sequence of the words māla and isyate appears as mālasyate. Verse texts require that sandhi be left unanalyzed in order not to hinder the prosody. Thus in order to link from words in Sanskrit texts to the relevant dictionary entry a knowledge of phonology is necessary in addition to a knowledge of inflectional morphology.
The Sanskrit Library at Brown University is currently producing a computational implementation of a realistic Sanskrit morphological and phonological production grammar and parser. Continuous Sanskrit text is generated from lexical items in a process-oriented style inspired by the Pāṇinian grammatical description of Sanskrit. Morphological and phonological rules are written using regular expressions in XML independent of any particular programming language and subsequently converted into executable code. The process of building continuous Sanskrit text from lexical items is segmented into three phases:

1. inflectional morphology
   a. of nominal stems
   b. of verbal stems
2. contextual phonological variation (sandhi)
3. construction of compounds and productive affixation

Phase 1a, completed with the assistance of the Consortium for Language Teaching and Learning in 2002-3, generates the inflected forms of any Sanskrit nominal stem. Phase 1b, completed in 2004, generates all the inflected forms of any Sanskrit verbal stem utilizing the stems in a database compiled from Whitney’s Roots. Scharf and Hyman plan to extend this database with information in Werba’s Verba Indoarica, Oberlies’ A Grammar of Epic Sanskrit, Goto’s “Materialien zu einer Liste altainischer Verbalformen,” and other scholarly articles. Phase 2, close to completion under a second grant by the Consortium for Language Teaching and Learning in 2005, generates sandhi between words. Phase 3 will generate syntactic agglutinative structures including nominal compounds and productive affixed forms. The inflectional morphology generator produced in phase 1 associates a morphological identifier with each form. The bidirectionality of this relation allows the implementation of a Sanskrit parser capable of analyzing Sanskrit sentences into their constituent words and identifying inflected forms. The morphological parser will be enriched with rudimentary syntactic analysis such as agreement phenomena.

Scharf is currently developing web-based interactive Sanskrit pedagogical tools under a grant from Brown University to work with Brown’s Scholarly Technology Group (STG).

In order to analyze forms in Sanskrit texts a parser must be combined with a database of lexical stems. Projects such as the Perseus Digital Library have succeeded in producing lexical data for morphological analyzer/generators by programmatic mining of machine-readable versions of traditional printed lexica. The lexical sources described above in section III should be sufficient for producing the lexical component of a basic morphological analyzer/generator for Sanskrit.

With a completed lexical database and morphological generator, it is possible to produce a full-form lexicon of Sanskrit, which maps every surface form onto a tuple \((L, M)\), where \(L\) is a lexical base and \(M\) is a set of morphosyntactic features. Morphosyntactic features are indicated in accordance with the morphological tagging scheme published by Scharf. Such a lexicon can be compiled into a trie or into a binary search tree for the purpose of performing morphological analysis. Alternatively, the full-form lexicon may be compiled into a finite-state transducer, thereby allowing both analysis and generation.

In reality, however, the Sanskrit lexicon is not closed because the language is characterized by agglutinative processes that are highly productive and are not captured in standard lexical sources. These processes include the formation of compounds and the addition of productive affixes to lexical bases. About half of a typical classical Sanskrit sentence consists of such structures. A computational implementation of a realistic morphological and phonological production grammar and parser therefore clearly must
account for such structures. In order for the parser to be able to identify the elements of
these structures, the grammar must produce the phonological varieties in which lexemes
appear in agglutinative structures (phase 3). These will be associated with the lexemes in
much the same manner as the phonological variations of inflected forms are associated with
their bases. The bidirectionality of this relation will allow the implementation of a Sanskrit
parser capable of analyzing agglutinative structures into their constituents.

V Challenges raised by Devanāgarī script

The most common script for writing and printing Sanskrit is Devanāgarī. This script
includes both syllabic and alphabetic features. Consonant graphs imply an inherent short
/a/ vowel, unless another vowel is explicitly indicated or the absence of a vowel is indicated
by the virāma sign. With the exception of word-initial vowels, for which there are
independent characters, vowels are written using dependent signs which may occur above,
below, before, or after the character(s) representing the preceding consonant sound(s).
Consonant sequences are written as ligatures; traditional Sanskrit orthography requires
glyphs for more than a thousand such sequences. Sequences involving r are especially
complex, involving both diacritics and ligatures.

The use of Devanāgarī script complicates computer processing of Sanskrit. Devanāgarī is one of the most complicated writing systems to display on a computer screen
or print; editing Devanāgarī puts special demands on text-processing software; and optical
caracter recognition (OCR) of printed Devanāgarī materials poses a number of substantial
technical challenges.

Display options for Sanskrit

Provided that text is stored in a reasonable encoding, it can be dynamically displayed
in Devanāgarī characters using appropriate software. Fundamental research has been
undertaken on the algorithms needed to render Devanāgarī script dynamically with correct
ligature substitution and positioning of vowels and diacritics. In these approaches, the
sequence of underlying characters is distinct from the vector of glyphs used for rendering.
Real-time shaping (the computation of the glyph vector from the underlying character
stream) requires new font formats and operating system services such as OpenType, Apple
Advanced Typography (AAT), and Microsoft’s Uniscribe. Some work has been done to
bring related technologies to the Java platform and Gnome desktop system. Yet a
significant number of existing computer systems and software programs do not provide
adequate support for Indic character shaping. Furthermore, we are only aware of a single
freely distributed OpenType font for Devanāgarī which however was designed for Hindi.
Available fonts generally do not support all the ligatures used for Sanskrit nor accents used
for Vedic. A number of fonts exist which are adequate for Sanskrit yet are undesirable in
that they employ idiosyncratic, non-standard encodings. Widely available, freely
distributed, high-quality fonts are a prerequisite to a modern digital library. The design of
high-quality Devanāgarī fonts legible in screen display will require considerable investment
of time and expertise by professional type designers.

We plan to collaborate with Ralph Bunker of the Computer Science Department at
Maharishi University of Management and Stephan Baums at the University of Washington
in designing solutions to the Devanāgarī display problems outlined above. Because
Unicode separates rendering issues from character-encoding, ligature selection need not be
encoded in the text but can be left to the font. Bunker's research involves the development
of a database of all ligatures found in printed Devanāgarī texts and software that allows a
user to build a customized OpenType font automatically. Any application that supports
OpenType fonts will then be able to render Sanskrit text with the user's choice of ligatures. This solution leverages the emerging OpenType standard to allow high-quality Devanāgarī typography without cumbersome data-entry conventions or specialized software. Baums is spearheading an effort to develop a cross-platform high-quality Unicode Devanāgarī font.

**Unicode-compliant text-editing software for Devanāgarī script**

Sanskrit language users conceptualize text at the phonemic or character levels, not at the glyph level. Therefore, text-editing software must allow the user to operate on characters, not just glyphs. The less isomorphic the character level and glyph level are, the more complex standard text-editing operations become. Pressing the backspace key, for instance, ought to translate into the removal of a character, but that does not necessarily imply the removal of a glyph. Suppose one wishes to delete the final r of yatr ( यत्र् ) and type i to form yati ( यति ). Pressing the backspace key after यत्र् should yield यत्; typing i then gives यति. Modern editors allow the user to position the insertion point (caret) by clicking at the desired location. In ligatures, especially where the components are not stacked horizontally, it is not obvious how to project the cursor position onto the underlying character stream. Further work on the ergonomics of editors for complex scripts such as Devanāgarī is needed.

An additional problem arises in that word-boundaries are obscured in Sanskrit not only phonologically by sandhi but also graphically by the fact that in Devanāgarī, consonant sequences, even when they extend across word boundaries, are rendered with a single ligature. Work currently under way in Pune and Cambridge, under the rubric of the InterLinear Sanskrit Processor (ILSP) project, aims to create a system that can handle invisible word boundaries within an “intelligent” editor. Such an editor would greatly facilitate the editing of Sanskrit texts which is an important task in building a digital library. We intend to work closely with the ILSP group to mutual benefit.

**OCR technology**

The complexity of the Devanāgarī script makes optical character recognition (OCR) applications more complicated than is the case for many other writing systems. OCR for Sanskrit texts in Devanāgarī is important to the development of digital libraries because of the high cost of manual data entry. The quality of OCR can be improved by integration with morphological and lexical knowledge systems such as those being developed at Brown University and Cologne. Two years ago the Brown team initiated cooperation with Venugopal Govindaraju, Associate Director of CEDAR, the Center of Excellence for Document Analysis and Recognition at the University of Buffalo, who is pursuing innovative techniques for Devanāgarī OCR under NSF support. In January, Brown hosted a Devanāgarī OCR workshop with Govindaraju and Malten to further integrate our collaboration in the proposed project.

Govindaraju is currently engaged in a project to expand access to ancient Indian manuscripts in the Saraswati Mahal Library in Tanjavur by making advances in Devanāgarī script OCR and cross-lingual retrieval. His project entails technological advances in four major areas:

1. Digital restoration and enhancement of old manuscript images
2. Automatic recognition of the Devanāgarī script
3. Annotation of words using machine translation techniques

Govindaraju enhances OCR by first undertaking an initial document image analysis. This analysis determines the zone of text (as opposed to images, etc.), determines line layout, margin lines, skew of text in the margin, discoloration, ink blots, and type-size. A character analysis then determines horizontal separation of orthographic syllables, the vertical profile of ascenders and descenders, and classifies characters according to structural features such as whether the horizontal and vertical bars are interrupted or not, and the location of vertical bar(s): absent, left, right, center, or multiple. Identification of the character(s) in the image proceeds by narrowing the range of choices of possible matches utilizing various levels of verification. Truthing, as such verification is called, proceeds at the following levels of text units: character component, syllable, word, and passage. Truthing requires assembling large databases of previously identified images at each level. The identification of text in digital images proceeds by constraining the possibilities until a match is determined.

Access to machine-readable texts, lexical sources, a morphological production grammar, and phonotactic and phonetic rules allows the automated production of large databases of truthing data to provide constraints for optical text recognition at all levels. Whole manuscript images may be automatically identified by matching them against known texts in machine-readable form. Passages in images may be identified by matching them against known passages. Words in images may be identified by matching them against morphologically tagged lexical items. Phonotactic rules constrain the possible range of collocated characters.

Govindaraju proposes to utilize the Brown's linguistic software, Cologne's lexical data, and Frankfurt's machine-readable text archive to build databases to enhance truthing in OCR technology.

VI Integration of the digital library system

Many of the building blocks for an international library of Sanskrit texts are already available. Yet in the absence of a developed infrastructure for technology and human cooperation, such a digital library cannot be realized. We propose to bring together existing resources to this end. Most critical is the integration of texts, linguistic tools, lexical resources, and OCR technology. In addition, a flexible user interface is needed, which will provide powerful search functionality and allow for display of text in a variety of formats including Devanāgari and Romanization. Adoption of emerging technologies and open standards will allow for interaction with a wide variety of other digital resources.

Integration of the linguistic tools being developed at Brown, lexical sources from CDSL, and the TITUS machine-readable text base will allow us to provide an integrated educational and research environment for Sanskrit analogous to what is now provided by the Perseus Digital Library for classical Greek and Latin literature. Enhancement with the Saraswati Mahal Library's digital images and OCR technology under development at CEDAR Buffalo will the most advanced intelligent information retrieval systems access to manuscript images. Such an integrated resource can provide extraordinarily wide access and high utilization. For example, a recent report from the Perseus Digital Library states:

The audience for these texts and tools in an integrated digital library does exist and it is larger than one might expect. The past four years have seen remarkable increase in the use of the Perseus web site. Usage has grown from just under three thousand hits on its first day in July of 1996 to two hundred fifty thousand hits on peak days during the current academic
In 2004, the main Perseus site logged more than 8,000,000 page hits a month.

As of July 2004, the main TITUS server averages 270,000 hits a day. According to the computer center of the University of Cologne the average daily access to the Monier-Williams digital lexicon is 1300 and to Cappeller 700 queries. In the three and a half years since our own digital Sanskrit Library was officially launched, more than 4,000 users have registered.

The digital library that we envision differs from Perseus in that it will distribute resources among the international partner institutions. Current technology makes possible a distributed network where components interact seamlessly and transparently.

The research group at Brown University is committed to delivering two applications that will be of immediate utility to a Sanskrit digital library. These applications will allow for lemmatized searching of Sanskrit texts and for parsing inflected forms in the context of continuous Sanskrit text. The current project will develop a distributed framework that will allow these tools to be used with texts and lexica located on the servers of the European partner institutions.

Our system will be able to generate dynamic links from inflected forms in Sanskrit texts to the relevant dictionary entries in lexica supplied by the CDSL and public sources. Since the linguistic tools being developed at Brown preserve all stages in the derivation of a form, it will be possible to present the digital library user with a considerable amount of grammatical information relevant for education and research. Since the Brown tools are based on XML, it will possible to apply innovative applications involving linking. For instance, we will be able to provide links to the Pāṇiniāna śūtras that describe all the rules that apply in the derivation of a particular form. This constitutes a novel and extremely powerful application of technology that will be useful to the communities of Sanskrit scholars and students.

While some users will want to view texts in Devanāgarī script, others will prefer the traditional transliteration or another Romanization. We will implement software that allows for dynamic switching of text display in accordance with a user’s preferences. Such a system will also be able to accommodate the variety of fonts and encodings that may be available on an end-user’s system. A prototype of this software is currently employed for the display of Whitney's Roots and inflectional software on the digital Sanskrit Library at Brown University.

The International Digital Sanskrit Library Integration project intends to cooperate closely with the Archimedes Project at Harvard University and the Max Planck Institute for the History of Science in Berlin, which is developing standardized tools to serve as a unified front-end to diverse sources of linguistic and lexical data in any language. These tools serve as middleware between editing and browsing applications on the one hand and linguistic and reference services on the other. The Donatus system provides a unified interface to morphological analysis software and databases. The Pollux system provides a unified means of access to dictionaries, or to any other reference work that is organized by alphabetized headwords, in any natural language. These tools will enrich the digital Sanskrit Library, shorten the development cycle for our digital library applications, and provide interoperability with other digital library resources.
The potential impact of allowing wide-audience access to the large number of Sanskrit texts is enormous. These texts constitute an enormous body of knowledge in diverse domains that is grossly underrepresented in the Western academic community. The Digital Sanskrit Library Integration project will dramatically increase the accessibility of these texts which are currently accessible only to highly-trained specialists. This access will be profoundly valuable to students, scholars, and the wider public concerned with such fields as historical and general linguistics, philosophy and religious studies, pharmacology and medicine, history of science and mathematics, and general history and literature of South Asia.

VII Potential for future development

The current project will provide the basis for a globally distributed, internet-based, international digital Sanskrit library. The framework instituted in the project will allow for the future development of even more powerful electronic access technologies. Techniques for automatic content analysis add considerable value to collections of digitized texts. Useful techniques include keyword extraction, technical term discovery, automatic summarization, and searching based on latent semantic indexing. For highly-inflected languages such as Sanskrit, it has been shown that the performance of such techniques is greatly improved if the text has been lemmatized. Tools and data developed in the proposed project will in the future be able to be leveraged for the application of these techniques.
Plan of Work and Management Plan

The proposed project aims to create a globally distributed, internet-based, international digital Sanskrit library in order to permit broad access to large numbers of texts in many fields for scholars, students and the wider public. The project integrates currently independent projects to create Sanskrit digital archives, digital lexica, and linguistic software; and to establish text-encoding standards; to enhance ancient and medieval manuscript access; and to develop OCR technology, display software, and Unicode-compliant text-editing software for Devanāgari script.

The Brown team will create the computational implementation of a realistic lexically based morphological and phonological production grammar and parser for Sanskrit. The Cologne team will contribute the digitized lexical database, combining data from the major Sanskrit lexicons, which will provide the lexical data needed for real-world use of the linguistic software. The TITUS team at Frankfurt will provide their extensive corpus of Sanskrit texts and contribute to the adaptation of the linguistic software to corpora of special dialects such as Vedic and Buddhist Sanskrit. The natural language processing software and lexical database developed at Brown and Cologne will assist the CEDAR Buffalo team in developing intelligent OCR applications that will enhance access to digital images of texts. The project will integrate existing TITUS digital-library web-infrastructure and the Brown digital Sanskrit Library. This combined framework will serve as the primary distribution-platform for the project.

In addition, the project aims to integrate further with other existing digital Sanskrit resources. The project will bring together scholars engaged in the production of OCR technology, data entry, text encoding, XML format standardization, lexical source compilation, text-archive management, linguistic technology, and display software. The following scholars have expressed interest in collaboration:

1. Peter Freund, Librarian, Maharishi University of Management.
2. Muneto Tokunaga, Professor of Sanskrit, Kyoto University.
   Freund and Tokunaga have compiled significant collections of digitized Sanskrit texts and are actively engaged in data entry.
3. Stephan Baums, Department of Asian Languages and Literatures, University of Washington.
4. Jens Braarvig, Professor, Department of Culture Studies, University of Oslo; Center for Advanced Study, Oslo.
   Braarvig is currently involved in the compilation of a digital lexicon for Buddhist Sanskrit texts.
5. John Smith, Faculty of Oriental Studies, Cambridge University.
   Smith is engaged in a cooperative project to develop Unicode-compliant text-editing software for Sanskrit and Avestan (ILSP, InterLinear Sanskrit Processor), funded in part by the University of Cambridge.
6. Ralph Bunker, Professor of Computer Science, Maharishi University of Management.
   Bunker is engaged in the development of innovative OpenType display technology for Devanāgari script and Sanskrit language educational software.
The project will include a workshop and conference in each of three years, two in the U.S. and one in Europe. Week-long workshops will include the core project partners at Brown, CEDAR Buffalo, Cologne, and Frankfurt and will provide the primary forum for the evaluation of the on-going digital library integration project. Conferences will allow for interaction with a wider community of researchers including especially those enumerated above.

**Timetable**

**Year 1**
- **Workshop 1:** Determination of project working protocols and focus on standardization of text-encoding and other interoperability issues.
- **Conference 1:** Sanskrit text-encoding, display, and optical character recognition.

**Year 2**
- **Workshop 2:** Preliminary integration of linguistic tools with Sanskrit language data sources and OCR software.
- **Conference 2:** Current issues in natural language processing techniques applicable to Sanskrit.

**Year 3**
- **Workshop 3:** Completion of Sanskrit digital library integration.
- **Conference 3:** Special issues for Vedic and Buddhist dialects: linguistic tools, lexica, and texts.
References

1http://dsal.uchicago.edu/
2http://www.columbia.edu/cu/lweb/indiv/southasia/cuv1/
3http://titus.uni-frankfurt.de/
4ftp://ccftp.kyoto-su.ac.jp/pub/doc/sanskrit/
5http://www.ucl.ac.uk/~ucgadkw/indology.html
6http://sanskrit.safire.com/
8http://oldsite.library.upenn.edu/etext/sasia/skt-mss/
9http://www.ucl.ac.uk/~ucgadkw/yogasutras.html


17http://dsal.uchicago.edu/dictionaries/list.html#sanskrit.

18http://sanskritlibrary.org/morph/


23http://www.perseus.tufts.edu/.


30http://www.perseus.tufts.edu/

