Report on Cologne Digital Sanskrit Lexicon Project

Mārcis Gasūns
Sanskrit Zealots' Society
Russia, Obninsk
gasyoun@gmail.com

Abstract

The Cologne Digital Sanskrit Lexicon Project (CDSL), initiated in 1994, represents a groundbreaking effort to create a one-of-a-kind digital repository of Sanskrit dictionaries. Over the past three decades, the project has evolved from a closed university initiative in 2013 into a collaborative, open-source space, hosted on the GitHub repo ecosystem with 6400+ issues by 2025, to integrate grammar, vocabulary, and textual corpora into a unified digital framework. The project began with the digitization of key 19th-century Sanskrit-English and Sanskrit-German dictionaries, such as Monier-Williams' Sanskrit-English Dictionary and Böhtlingk's Sanskrit-Wörterbuch, and has since expanded to include 42 dictionaries compiled in the last millennia and some major grammar reference works.

The CDSL project remains unique in its approach to linking dictionaries with a parallel text corpus and semi-digitized PDF scans, particularly the Sanskrit-Russian-English-German corpus of Vedas, which serves as a critical resource for cross-referencing and validating lexical entries. This integration of dictionaries and corpora allows for a more dynamic and accurate representation of Sanskrit lexicography, moving beyond static, printed dictionaries to an evolving digital resource. The total number of normalized entries is 392600+ and it still grows similar to the Himalayas at around 2 millimeters per year.

1 Introduction

At the 37th ICANAS, held in 2004 in Moscow, under the aegis of the Institute of Oriental Studies (IOS) of the Russian Academy of Sciences, was the first time I met Peter Scharf's book Ramopakhyana, printed in 2002. First, it occurred to me how a Sanskrit reader should be made, including traditional and detailed classification of compounds. Second, I met there with prominent Japanese and Indian Indologists, and sitting in the restaurant after one of the sessions a small talk burned the bridges. They said that both of them are making marginal notes on their copies of Monier-

Williams (MW) dictionaries. I understood that there exist many private lists with corrections and additions for Sanskrit dictionaries that will never see the light of day. Dust is what they will gather, not attention. And that work should be centralized at least partly, gathering such marginal notes. Now, after twenty years, I must admit we are not a single step closer. But we are getting ready. There is still no Wiki-format dictionary of Sanskrit, where one could add a meaning or quote and get verified and validated by the scholar community. That was the day Dr. Peter Scharf and Sir Monier-Williams changed the very way I think about Sanskrit studies. Changed once and forever.

The Cologne Digital Sanskrit Dictionaries project is a long-term project (1994-2025) with one major aim: to be the ultimate place for validated Sanskrit dictionaries. In contrast to the derived projects, it's updated regularly. None of the websites that are based on Cologne data maintain a connection to Cologne, not all of them even give full credit. It's a pity that corrections never reach outside Cologne, once the original data is scraped. None of them retain links to the corpus, deleting the tags, the most valuable asset of Cologne. Most of the links have been implemented only lately. In the ever ongoing correction of mistakes and additional hyperlinks Cologne digital dictionaries will remain superior to other digital Sanskrit dictionary projects and it's our responsibility to keep up the standard and spread the word.

Nowadays it's performed collaboratively without a complex organizational superstructure voluntarily, but lack of funding is slowing it drastically. As stated at the 10th International Sanskrit Conference in Bangalore (Kapp and Malten, 1997) almost 30 years ago:

> The Cologne Digital Sanskrit Lexicon (CDSL) project undertakes to digitize and merge the major bilingual Sanskrit dictionaries compiled in the 19th century. Its aim is to provide a basic lexical corpus to provide easy access to all available meanings of Sanskrit words and to allow the creation of a number of computer programs that will help to analyze Sanskrit texts. In the first stage Monier-William's Sanskrit-English dictionary (MW) has been digitized to be followed at a second stage by three other dictionaries (Cap, PW2 and Sch). All these have been structured and mostly unified to allow access to the meanings as developed by the different lexicographers. As a final goal it is hoped that a step can be taken towards an integrated Sanskrit word catalogue which codifies the distribution of lexical units in Sanskrit text corpora by linking them to the existing descriptions in dictionaries by a numeric system which functions as a placeholder for a word sense which can be expanded or changed.

2 Sanskrit Dictionaries

2.1 Sanskrit-German Dictionaries

2.1.1 Core Sanskrit-German Dictionaries

The initial plan contained the most frequently quoted Sanskrit-English Dictionary (Monier-Williams 1899) and three genealogically connected Sanskrit-German Dictionaries: Sanskrit Wörterbuch (Cappeller 1887), Sanskrit-Wörterbuch in kürzerer Fassung (Böhtlingk 1879) and Nachträge zum Sanskrit-Wörterbuch (Schmidt 1928).
Even the abbreviated IDs of dictionaries have changed over time quite a few times. IDs are important and stated only on the Cologne homepage. One has to remember them by heart, to be able to switch between the dictionaries quickly, for example, in the new 'simple' search mode[1], developed in 2019. It's the most advanced UI feature by far. IDs are all capitals now as well. Cap has become CCS, PW2 is PW, and only Sch remains as SCH. PW2 was never used outside Cologne, it's PW or PWK now. As noted in (Gasūns 2006):

> The Sanskrit dictionary which is quoted as the ultimate authority, the Sanskrit-Wörterbuch (W) printed in Petersburg (P) by Otto Böhtlingk (B) and Rudolph Roth (R), großes (G) and after that kleines (K) (full and concise versions of it). Depending on which item is seen as the most important one, there are at least five variants of referring to it by means of an abbreviation: PW(Bt, Th, Wb), Bö (Wh), PWG (MG), BR (Db, Ln), PW1 (Kö); pw (Th, Wb), KBR (?), PW2 (Kö); PWN[achtrage] (Wb for Sc).

Why exactly did these German dictionaries become the core? PWK continues PWG. CCS is an abbreviated version of PWK. SCH is an update to PWK. MW was seen as an English translation of PWK by some. And, as the final touch, it was updated in 2013-2016 as Nachtragswörterbuch zu den Petersburger Wörterbüchern[2] in Germany. So there are at least four levels of continuation. The Nachtragswörterbuch is the first big Sanskrit dictionary project after 1928, if not to count Apte 1957.
Even though the 'retrodigitization'[3] is still not over, other than 19th-century dictionaries have been added in the meanwhile, including Practical Sanskrit-English Dictionary (Apte 1957), An Encyclopedic Dictionary of Sanskrit on Historical Principles (Ghatage 1976). None of the three English-Sanskrit Dictionaries were

[1] https://sanskrit-lexicon.uni-koeln.de/simple

[2] https://nws.uzi.uni-halle.de

[3] Making a file as close to the original printed book as possible.

planned initially: English-Sanskrit Dictionary (Monier-Williams 1851), English-Sanskrit Dictionary (Borooah 1877), Student's English-Sanskrit Dictionary (Apte 1920) thanks for the combined efforts with Peter Scharf from Brown University.

2.1.2 Challenges in Merging Dictionaries

The merge of the major bilingual Sanskrit dictionaries did not go as well as planned. The final merge never happened, if ever should. One of the latest preliminary merges, finalized ten years ago, entitled 'Experimental displays for Böhtlingk/Schmidt dictionaries'[4], merged all of the major Sanskrit-German dictionaries (`pwg | pwk | pwkvn + sch`) with known at that time errors corrected, remains unpopular. Something that might make more sense, is not merging the dictionaries, but displaying the results of all the dictionaries at once, entitled 'dalglob1: experimental multi-dictionary display'[5]. That's a UI that we should focus on in light of the initial 'merge' idea. Cologne was never strong in UI, other than the 'simple' search mode. It's the ever-updated structured data that remains of value, but we can't ignore ordinary users as well.

2.2 Incomplete Tagging

2.2.1 Meaning Tagging

As for the '(incomplete) tagging of different meanings' not much has been done after to fix it. Meaning tagging was never a priority, because it's far from being a low-hanging fruit and the amount of work might be above our expectations. How to correlate Aptes' meanings to MW? We have never attempted. Meaning tagging is not in the final form. You do not try to correlate Webster and Collins for English, do you? It's unique.

The display scripts do the separation of meanings. At the code level it's done only in a single dictionary, MW. The semicolon groups '`;`' separate different senses. We can't say, that the MW IDs are reserved only for meanings, they are used for both, not only headwords, but meanings as well (including etymologies). So there is still plenty to explore and research in regards to tagging, including meaning tagging.

2.2.2 Verb Form Tagging

Similarly with 'tagging of verb forms', although a lot of verb markup was added, it's still far from perfect, including the unstructured representation at the beginning of

[4] https://sanskrit-lexicon.uni-koeln.de/scans/csl-apidev/pwkvn

[5] https://sanskrit-lexicon.uni-koeln.de/scans/csl-apidev/sample/dalglob1.php

verb entries. We have some level of tag coverage, but it needs to be improved, it's neither complete nor accurate.

2.3 Computer Programs

2.3.1 Online Tools and Services

As per 'allow the creation of a number of computer programs that will help to analyze Sanskrit texts', that indeed has surpassed the plan, but mostly in the form of online services, some just a knockoff, still far more popular than the source. Tools, such as INRIA morphological analyzer/generator for Sanskrit[6], DCS (Digital Corpus of Sanskrit)[7], Ambuda[8], Kumulatives Nachtragswörterbuch[9] heavily rely on MW and VedaWeb Rigveda[10] on GRA and SRPC, Sanskrit-Russian Parallel Corpus. In Cologne's 'simple' search mode, frequency data based on DCS is used. Before that DCS based its lemma list on MW[11]. Now Cologne benefits itself from aggregated data from DCS. What has not been solved yet is that once derived, as said before, the websites that based on Cologne data, never implement the updates going on at Cologne. Even the critical ones at the headword level, not to mention the entry-level ones.

2.3.2 Offline Sanskrit Analysis Tool

An offline computer program deserves to be mentioned as well. "Saudamani"[12], developed by Vladimir Leonchenko since 2019 and not known outside Russia. It's 90,000 lines of code for Sanskrit quantitative analysis, nothing similar was done after Whitney, who introduced the approach for the first time in Indology. We have compared not only the macrostructure and microstructure of different Sanskrit-Russian dictionaries (Kretov and Leonchenko, 2021) but researched on totally new level the MW plagiarism question (Kretov and Leonchenko, 2022).

---

[6] https://sanskrit.inria.fr/MW/1.html has not updated since 2012.

[7] http://www.sanskrit-linguistics.org/dcs

[8] https://ambuda.org/texts/ramayanam/6.43

[9] Ein kumulatives Nachtragswörterbuch zu den Petersburger Wörterbüchern (pw)
von Otto Böhtlingk und den Nachträgen von Richard Schmidt https://nws.uzi.uni-halle.de

[10] https://vedaweb.uni-koeln.de/rigveda/view/id/1.1.1

[11] https://github.com/OliverHellwig/sanskrit/tree/master/dcs/data with the current Nov. 2018 dump of the DCS main database contains a list of known issues, but even the publically documented ones were never answered since 2019. Some are critical and validation of the raw data of DCS would be a timely task.
It's a pity that the initial order of the meanings has been lost at DCS, making it alphabetical for no obvious reason. Scraping mode https://github.com/sanskrit-coders/dcs-scraper means we can fix something derivated, but never the source.

[12] https://samskrtam.ru/saudamani-win

The issue of stealing words and meanings was first raised in the 19th century by Böhtlingk himself in one of the prefaces (Böhtlingk 2007) to the Petersburg Dictionaries. First, he wanted to keep silent, but Max Müller agitated him to speak out loud and so he did (Stache-Weiske 2015). Böhtlingk's arguments were then again weighted in the 20th century, without adding any new data (Zgusta 1988). 35 years after Zgusta we have used digital tools not to verify the 40+ old selected cases, but to compare all the 168633 MW lemmas to 106169 PWG lemmas (Gasūns 2024) as per the 21th century capabilities. Surprisingly only 87091 lemmas are in common. The entries in MW, a dictionary treated as a substitute for an English translation of PWG, are generally longer than entries in PWG, including more meanings.
Despite its smaller amount of lexica, PWG contains all the basic, most common vocabulary, which makes up 4/5 of its volume, while MW contains a large number of *samāsas*, making up half of its volume. The average size of a dictionary entry in characters, MW: 110 vs. PWG: 245. 30,000 entries in both dictionaries differ only by 15% in size, 819 entries are identical in size. Out of 485 citation sources in MW, 330 sources are identical with PWG. 41,6% of citations link to the same sources. We have yet to verify the bold statement that "MW has largely followed the order of meanings in PWG" (Kapp and Malten, 1997). It remains far from obvious.
The union of grammar, the text corpus, and the dictionary remains a dream far away.

2.3.3 High Priority Dictionaries

After 30 years in business, Cologne remains the origin of Sanskrit digital dictionaries, the industry standard. From the 42 Digital Lexicons at Cologne, 8 dictionaries are chosen for highest priority:

- WIL 1832 Wilson Sanskrit-English Dictionary
- MW 1899 Monier-Williams Sanskrit-English Dictionary
- AP 1957 Practical Sanskrit-English Dictionary
- PWG 1855 Böhtlingk and Roth Grosses Petersburger Wörterbuch
- PW 1879 Böhtlingk Sanskrit-Wörterbuch in Kürzerer Fassung
- SCH 1928 Schmidt Nachträge zum Sanskrit-Wörterbuch

+ Nachträge und Verbesserungen

- VCP 1873 Vacaspatyam
- SKD 1886 Sabda-kalpadruma

3 History of Digitization of Sanskrit Dictionaries

3.1 Cologne Digital Sanskrit Dictionaries

3.1.1 Early Digitization Efforts

So before we move forward, it's time for a review of the original report, made by the grandfather of Cologne Dictionaries (Kapp and Malten, 1997), when Indology was still alive in Cologne. There are five points to be looked at:

1) 'In the first stage Monier-William's Sanskrit-English dictionary (MW) has been digitized.' Indeed it was, nothing would be possible without it. 'An earlier project to digitize Indian lexical resources at the University of Chicago in 1985 (which failed for lack of funding) also included MW.' (Kapp and Malten, 1997).

2) 'to be followed at a second stage by three other dictionaries (Cap, PW2 and Sch).' Yes, and they remain the core dictionaries, although the Sanskrit-English dictionaries are the most wanted ones.

3.1.2 Evolution of Data Formats

3) 'All these will be structured and unified to allow access to the meanings as developed by the different lexicographers.' The movement to Unicode started before github. Thomas' approach was a numerical approach, which looks horrendous from today's perspective. If one opens `MONIER.ALL` one can see how ugly it is when open, it becomes non-readable. So Jim moved out from `ALL` format for all dictionaries, his Python code converted ASCII to UTF. Peter was still involved. First to get rid of the encoding, then off the number-letters. It was a long run 2013-2024. It was where the real development began. Sanskrit words are usually spelled with SLP1 transliteration in the backend nowadays and IAST in the frontend by default.

First different repos were made for each dictionary in 2014, but a lot of duplications happened. Keeping in mind the unity of the structure for most dictionaries, like the similarity of how to display MW or PWK, so `https://github.com/sanskrit-lexicon/csl-orig` is where the unification in terms of the infrastructure happened and where the main source files for each dictionary are kept nowadays. Separate repos are left for discussions and analytical research, not for the dictionary files.

The next big move was in 2018, it was moving to a L-based data format.

3.1.3 Integration of Dictionaries and Corpora

4) 'As a final goal it is hoped that a step can be taken towards an integrated Sanskrit word catalogue which codifies the distribution of lexical units in Sanskrit text corpora by linking them to the existing descriptions in dictionaries by a numeric system which functions as a placeholder for a word sense which can be expanded or changed.'

Placeholders are not shown, but integrated. Catalogue of senses[13]? No cross-reference of senses is available. For example one can have an advanced search in one dictionary for the English word 'night', but nothing of such kind is still available for all the dictionaries at once.

DalGlob[14] is not about meanings, but similar in some way. We have the headwords spelled as they are in the dictionaries, which makes life harder (because orthography has changed several times since they were compiled). In 'metaline' metadata we could make things searchable. Why not add an additional field of a standard spelling, a normalized spelling as part of the metadata? There will be issues with a few dictionaries, like WIL and two other dictionaries closely based on it, because of anubandhas in the verbal roots (would require manual edition). It would be a step in the direction of unifying headwords. Not sure we want to get to the unification of senses.

5) 'Last but not least connecting Sanskrit with Tamil vocabulary is envisaged. To this end, the major Tamil dictionaries have already been converted into digital form.'

Not sure if transcoding 'Tamil Dictionaries'[15] for SLP1 for Tamil is needed and if the link fulfilled the initial vision. The only Tamil expert on the team is Nāgabhūṣaṇarāvu and we have converted him to Sanskrit Dictionaries. In the field of Dravidian lexicography, less work is left to be done after Nāgabhūṣaṇarāvu arrived.

3.2 Copyright Infringement

3.2.1 Early Copyright Issues

Although in 1997 (Kapp and Malten, 1997) it was stated that 'to allow free access to the data compiled it is necessary to impose as little as possible restrictions on lexical databases', in fact, the words 'seems desirable to have high-quality Sanskrit databases freely available to all researchers' remained only a declaration at least until 2005.

In 2004 Thomas Malten sent me from Cologne to Moscow a DVD with the PWG and PWK scans, where every page was named containing the Sanskrit headword that started on it. No website was yet available in Cologne[16].

---

[13] https://sanskrit-lexicon.uni-koeln.de/scans/csl-santam/php/index.html seems to be related, but a different dataset.

[14] https://sanskrit-lexicon.uni-koeln.de/scans/csl-apidev/sample/dalglob1.php

[15] https://www.sanskrit-lexicon.uni-koeln.de/tamildictionaries/tltab/index.php

[16] Institute of Indology and Tamil Studies (IITS) does not exist any longer. Klaus Ludwig Janert was the first professor in Cologne, so the department must have started in 1963. Ater him was Dieter B. Kapp and Ulrike Niklas. After 2006, Indology was reduced to Tamil studies and the Indology program became part of a weird composite 'Cultures and languages of Asias' bachelor and masters

After uploading the dictionary PDF scans publicly on Gasūns' server, Thomas Malten, at that time a senior lecturer at the Institute of Indology and Tamil Studies in Cologne, wrote two letters. The one addressed not for me contained the words: 'including the scanning and processing of scanned images of Skt dictionaries. So far there are about 10000 double pages scanned, the names of which have been aligned with the page numbers of the originals and many of the pages have been supplied with the first sanskrit word on each page in the filename' and that 'the dictionaries as such are not copyright, but our work on them is protected under EU law as a database. After finishing the project everything will become available free for non-commercial use with all proper acknowledgments.'

3.2.2 Current Copyright Status

So at that stage in 2005, it was supposed I 'acted rashly and without evil intentions', but it was the first time I was treated like a criminal in the digital world of dead languages, smuggling dictionaries. In 2004 I started scanning books on Sanskrit and have scanned thousands of books in high-res quality with tens of thousands 'double pages' in the meanwhile, spreading them as far as I could reach. So I was not aware that there could be a different stand on it, but I deleted the data as per Malten's request. And now, after 20 years, we still have to fight for the freedom of books on Indology and dictionaries of rare languages. Times have changed. The time has come.

3.3 Computers in Lexicography

3.3.1 Evolution of Lexicography Tools

The mind-boggling changes that lexicography (and the computer world at large) underwent in just half a century are dramatic. From pre-personal computers with data punched into cards and paper tape to today's software in the cloud which independently prepares draft corpus-driven dictionary entries for lexicographers to review though remains a far future for Indian languages. But one thing we know for sure, is that users contribute via crowdsourcing like in the 'The Professor and the Madman' movie from 2019. And it is crowdsourcing which we would want to attract even wider from the perspective of planning for the next half a century in Sanskrit lexicography. AI alove will not make it, even Dharmamitra[17] will not suffice. Data verified by volunteers and multiplied by computer programs, including AI technologies, is where we are heading.

---

program. After Prof. Ulrike Niklas retired in 2022 what was still left folded into the Asian department and was swallowed by the Institute of Languages and Cultures of the Islamicate World.

[17] https://dharmamitra.org

3.3.2 Challenges in Digital Lexicography

In what ways are digital dictionaries already able to serve users better than hard-copy dictionaries? The use of large corpora during dictionary compilation, combined with software helps to analyze and synthesize huge amounts of data. From the perspective of the dictionary-maker, this ensures being certain that nothing important has been overlooked; or even: that what is frequent and common is effectively covered in the dictionary, while knowing that one has safely omitted what is truly less important.

The first dictionary to use computers extensively in its editorial preparation was Random House Unabridged. We in the field of Sanskrit lexicography, have not yet used computers extensively in editorial preparation. Sanskrit scholars use computers as typewriters. For many years since 2007, the existing Sanskrit digital dictionaries were merely printed dictionaries made available on an electronic platform. We are still bad at making them even a 1-to-1 copy, because data entry errors, samāsas, and encoding of dictionary structure markup are nowhere even close to perfect even for the most widely used Monier Williams Sanskrit-English dictionary, 1899 edition.

Tens of thousands of mistakes or markup issues are still ahead that can't be fixed in a single decade. It took the German-Russian genius 23 years to compile and print two of the biggest-ever Sanskrit dictionaries almost alone and we have not managed in 30 years to replicate them group-wise. But we still try to continue. Even if we fail.

3.4 Encoding Standard

3.4.1 SLP1 Encoding

What looked like kuJjakuTIra from `.{kuJjakuTIra}3{kuJja--kuTIra\}` in 1997 is in 2025 in SLP1 encoded as: `kuYjakuwIra`

```
<L>51478<pc>288,1<k1>kuYjakuwIra<k2>kuYja-kuwIra<e>3
<s>kuYja-kuwIra</s> ¦ <lex>m.</lex> a bower, arbour, <ls>Mālatīm.</ls>;
<ls>Gīt.</ls><info lex="m"/><LEND>
```

To understand how much of the backend work has been going on after the initial funding stopped, the same entry initially:

```
<H3>100{kuJjakuTIra}3{kuJja--kuTIra}¦ •m. <a…bower…,…arbour> <¯Ma1lati1m.>
<¯Gi1t.> MW034516
```

3.4.2 XML Structure

Thomas was not interested in XML. XML started with Malcolm, so the question was now how to make corrections to MW? XML header files need further exploration for names and dates to add details to the historiographical research that has only

started. Thomas was interested in typists from India, he was looking for minimizing mistakes in typing. Thomas developed transcoding principles, making them consistent. For example, they double-typed Burnouf: two different typists, investing to compare the differences, faithfully typing what was in the images. Something similar to that was done ten years later with Vacaspatyam, comparing two separate editions of the Sanskrit-Sanskrit dictionary, Tirupati vs. Cologne[18], with TextDiff.
So it took a decade to move from emails to github[19]. 'As regards methods of long-term corrections and improvements to CDSL, these should be seen in the light of the possibilities of electronic communications' (Kapp and Malten, 1997). Github made collaboration possible, otherwise, everyone had a collection of emails, and it was hard and inefficient.
Where is the code? Can other people see the code? The issues quickly piled up between 2014 and 2024 not so quickly to be solved (`csl-orig`, `CORRECTIONS`[20], `cologne-stardict`, `csl-apidev`, `csl-corrections`, `csl-pywork`[21], `csl-websanlexicon`, `MWS`[22], `PWK`, `hwnorm1` repos being the most active ones). Some of them will remain unsolved, but at least documented.
We have been made aware of hundreds and thousands of issues related to the eight main dictionaries, not to mention the dozens of others, the second-tier dictionaries. As the foremost authoritative resource for Sanskrit dictionaries, there is a pressing need to expand the team of volunteers dedicated to Sanskrit lexicography regularly.

3.4.3 Data Structure

Initially, there was no concept of alternate headwords[23]. If one searches for akabara, he will find akabbara and akavara as well. But if one would search for akabbara and akavara in the old days, nothing would be found. It's a question of infrastructure and has been solved only lately. Before 2024 we had it originally:

```
<L>159.1<pc>1308,1<k1>akabara<k2>akabara,akabbara,akavara<e>1
<s>akabara</s> or <s>akabbara</s> or <s>akavara</s>, ¦ <lex>m.</lex> (emperor) Akbar, <ls>Inscr.</ls><info n="sup"/><info lex="m"/><LEND>
```

After 2024 it's:

```
<L>159.1<pc>1308,1<k1>akabara<k2>akabara<e>1
<s>akabara</s> or <s>akabbara</s> or <s>akavara</s>, ¦ <lex>m.</lex> (emperor) Akbar, <ls>Inscr.</ls><info n="sup"/><info lex="m"/>
```

[18] https://github.com/sanskrit-lexicon/VCP/tree/master/vcpte-vac
[19] https://github.com/funderburkjim/2024review/tree/main stats for 10 years a github.
[20] post-Peter `CORRECTIONS` are replaced by `csl-corrections` repositary now.
[21] The CSL prefix means it is related with the presentation.
[22] Repositaries with IDs of dictionaries are the place for corresponding analytical research.
[23] https://github.com/sanskrit-lexicon/alternateheadwords

```
<LEND>
<L>159.2<pc>1308,1<k1>akabbara<k2>akabbara<e>1
{{Lbody=159.1}}
<LEND>
<L>159.3<pc>1308,1<k1>akavara<k2>akavara<e>1
{{Lbody=159.1}}
<LEND>
```

Not all team members agree with the new format. There remains a difference of opinion. The new format could be made as well as a result of a Python script at the time of XML generation, leaving the original book as one file and the introduced alternate headwords as another file. From Thomas' format to Jim's format. The only major issue with the old approach, when generating an XML on the go, remains the assignment of random L numbers, for example, `ID=45270`. We are still not sure we have counted all the words correctly. Sanskrit is growing regularly. Deleting or adding supplements or corrections goes on and stable L numbers have a bonus of a trustable placeholder.

3.5 Team Members

3.5.1 Key Contributors

What started as a project led by Thomas Malten (originally from Germany, now in Cambodia) in Cologne in partnership with Peter Scharf (then USA, now in India) and coded by Malcolm D. Hyman (Germany), has been led by Jim Funderburk (USA). Since then, volunteers have joined, including Dr. Sampada Savardeka, and more sporadically, Felix Rau and Jonathan Migliori.

Since 2007, Dr. Mārcis Gasūns (Russia) has been communicating via emails, and in 2014 he founded the sanskrit-lexicon Github[24] organization. He was joined by new team members: Dr. Dhaval Patel, Dr. Usha Sanka, Nāgabhūṣaṇarāvu Kālepu (India), and Anna Rybakova (Russia).

In 2004 Jim went to Brown University to meet Peter Scharf and Malcolm D. Hyman. He was interested in Ramopakhyana initially[25], only then did dictionaries took him as a hostage. So Ramopakhyana was the turning point not only for Dr. Gasūns, but Funderburk as well, both in 2004. Jim realized that things still needed to be corrected. P. Scharf contacted Th. Malten, if they could collaborate. It took a year in email mode between Jim, Thomas, and Malcolm just to understand how to manage, it was very complicated what Thomas had. By 2008 the markup was decided[26].

[24] https://github.com/sanskrit-lexicon

[25] https://sanskritlibrary.org/catalogsText/rAmopAKyAna.html

[26] https://www.sanskrit-lexicon.uni-koeln.de/talkMay2008/mwtags.html

The switch to SLP1 was made in 2008. Sanskrit Library Phonetic ASCII encoding (SLP1) makes it impossible to lose any data (Scharf and Hyman 2009), which was not so in the era of Hardvard-Kyoto convention because of the small list of Vedic words with hiatus (hiatus-190-entries.txt).

Almost none of the initial team members are active now. The ship has lost its captain. The first stage with Thomas Malten was mainly over by 2004, the more by 2013 with some additional work being done even in 2018. Four initial dictionaries and Apte that Thomas did independently typing without Peter was where it all started. To have a lot more dictionaries and get funding was Peter's idea around 2007, but it actually got done in 2010-2013. Not only dictionaries but reference works were added, like Whitney's Roots[27] with the assistance of Susan J. Moore.

The original digital Monier Williams was created under the direction of Thomas Malten and entered by typists in Azhivaikkal in Thanjavur district, Tamil Nadu, with character codes partially marking divisions in the text. Scharf and Funderburk, with the assistance of Hyman and Chandrashekar, analyzed the existing markup, transformed it to XML, disambiguated it, and extended it. In the good old days corrections were evaluated by Malten, Scharf, or Chandrashekar and implemented by Funderburk.[28]

3.5.2 Major Cologne Contributors 1994-2025

1994-2013 Thomas Malten (Germany)
2004-2013 Peter Scharf (USA)
2004-2009 Malcolm D. Hyman[29] (Germany)
2004-2025 Jim Funderburk (USA)
2014-2023 Sampada Savardekar (France)
2014-2025 Mārcis Gasūns (Russia)
2014-2021 Dhaval Patel[30] (India)
2021-2025 Nāgabhūṣaṇarāvu Kālepu[31] (India)
2024-2025 Scott Rhodes[32] (USA)

[27] Based on https://sanskritlibrary.org/Sanskrit/whitney/index2.html

[28] https://sanskrit.inria.fr/MW/MWHeader.html

[29] Until his untimely passing.

[30] Nowadays to implement Scott and Nāgabhūṣaṇarāvu corrections.

[31] Strategic approach, hard to implement corrections.

[32] Tactical approach, easy to implement corrections, submitting 10 corrections a day, the most prolific correction submitter. But a lot of time is required to implement them. It would require 2 weeks for 1000 batches, 5-6 hours per day.

By coincidence, the initial dream of Thomas (1997) is partly fulfilled by Nāgabhūṣaṇarāvu: 'Last but not least connecting Sanskrit with Tamil vocabulary is envisaged. To this end, the major Tamil dictionaries have already been converted into digital form.'

3.5.3 Indian Volunteers

As per Nāgabhūṣaṇarāvu CDSL is changing very slowly. Most of the work can be done in no time, is his stand on most of the corrections needed. In 2006 he started a Tamil dictionary[33] section on his own website, developing it till 2016 as a retired engineer from Hyderabad together with his colleague Seshatalpa Sai. Now in 2025, it has 30 more dictionaries to come, tagged already. He alone is like a department. Hyderabad is the new capital of Sanskrit, because not many Sanskrit studies are left in St. Petersburg.

In a personal email in 2013, it was stated:

> 'We worked most of last year to rationalize the markup of the Tirupati version of Vacaspatyam. It was such a mess and full of errors that Thomas decided to do the data entry again from scratch. We have included what we did in SL multidictionary but expect that Thomas's version will make it obsolete.'

So that means that in the last 15 years, there have been several rounds of cleaning a single dictionary and still there is much left to do. Since early 2025 and for the next 1.5 years Nāgabhūṣaṇarāvu is at Vacaspatyam cleaning now the 5400+ pages. In 2022 Vacaspatyam and Shabdakalpadruma headword-only cleaning was done, now he is at the entries, after getting a good scan. Not always did we have a good scan in the beginning, several scans have been updated at a later stage.

As per Nāgabhūṣaṇarāvu Telugu consists 70% of Sanskrit lexica. So if we search for 'kamala' at andhrabharati.com we will see the '`saṃ.`' tag that stands for Sanskrit, '`vi.`' for *viśeṣaṇa*, adjective, and '`strī.`' for feminine with an '`ā`' ending:

```
saṃ. vi. ā. strī.
1. lakṣmi;
2. uttamurālu;
```

3.6 Digital Dictionaries of South Asia

3.6.1 Challenges in Maintaining Digital Resources

There were two websites in the early days of the Internet with Sanskrit dictionaries. The question is never to make a digitization once and forever. That is impossible.

[33] https://andhrabharati.com/dictionary

The question is who will be able to maintain the community of people who will keep the fire burning after the major work is over.

3.6.2 Comparison with Other Projects

Initially, Cologne Digital Sanskrit Dictionaries was not the only project dealing with Sanskrit dictionaries, and there still is `https://dsal.uchicago.edu/dictionaries`, although updated only cosmetically every now and then. The Chicago website has only two Sanskrit dictionaries.

Both of the DSAL Sanskrit dictionaries are on the Cologne website and with fixed print errors. No gaps in Arabic or Greek now; MW72 needs to be checked. So the Chicago-hosted copies of Sanskrit dictionaries are of no interest, and the UI is even less user-friendly than that at Cologne.

Macdonell's dictionary was never popular (Macdonell 1929). MD is a medium-sized dictionary, 20748 lemmas, added in 2006. Typographically compact and well printed, its transliteration in the original book is outdated, and etymologies are only sporadic (contrary to WIL, where etymologies are given in a systematic manner superior to all known dictionaries since 1832). MD at CDSL is an evolving copy of DSAL.

Apte was a Sanskrit genius and the 3rd edition of the dictionary (Apte 1957) is the only one that has moved forward as per MW. Digital Dictionaries of South Asia had long ago settled copyright with Prasad Prakashan. For that reason, AP was available publicly at DDSA, and only unofficially at CDSL. Concerning the copyright of Apte 1957 Peter Scharf wrote recently since P.K. Gode, the last surviving editor, passed away in 2021 and Y.G. Joshi, the publisher, in 1963, after 60 years of the decease of the latter, the copyright has ceased. Thomas Malten has confirmed and we are making that dictionary available at Cologne publicly for the first time in 2025.

4 Resources

4.1 Normalizing Headwords

How headwords are spelled? There are minor differences around dictionaries, especially Indian and non-Indian. Normalization of headwords is accomplishable[34]. We do not have a fixed standard as of now. Normalized headwords was a step in that direction (Patel 2016), but it's far from complete. With the right adjustment, the cross-dictionary headword issues can be solved.

For example, internally we know the word devī is a feminine from deva. But we can't search for devī, devī will not show us deva, even if we can find the feminine headword.

[34] https://github.com/sanskrit-lexicon/hwnorm1/tree/master/sanhw1

Feminine forms as search forms could be considered as alternative headwords. Quite easy, if we identify them with the L-body technique.

4.2 Correction of Dictionaries

4.2.1 Correction Changelog

As per Th. Malten vision of the 'continuing work of correcting typographical errors' never stopped and only got more systematic and productive, after Malten entrusted the Sanskrit Lexicons to the DCH (Data Center for the Humanities, University of Cologne) with the explicit mission to archive, preserve and make accessible the digital resources entrusted. So the Sanskrit dictionaries aren't affected by the decline and disappearance of the Indology department in 2013. As Thomas, a one-man band, had not perfected the typist error correction regime, he did not get rid of all typoes. There are some error-prone places, like 5-10% of citations in Apte have errors. It is our task to finish what he started, share, and give it to the world, as not everyone can study Sanskrit for 10,000 years alone. As our recent prolific typoe-catcher, Scott Rhodes stated on the yearly Cologne volunteer call: 'A mistake in Sanskrit is a placeholder for a correction, the only way to reach true knowledge.' Anyway, who would want a book on Sanskrit without mistakes?

Starting from 2014, when moved to github, the first submitted typoe was by Dr. Usha Sanka (India), leaving Sanskrit dictionaries afterward. These errors are submitted through the 'Correction' link on the internal pages of dictionaries, redirecting to 'Sanskrit-Lexicon Correction Form'. The TSV file does not contain all of the ever-made corrections. Later in the process, some changes were never documented, as there were too many of them at once (when structural changes involving regex were implemented in batch mode). So not everything is shown in `cfr.tsv`[35].

4.2.2 Print Changes

`mw_printchange.txt`[36], begun in 2008, documents changes made to the digitization of MW in which the digitization is intentionally different from the printed edition. The reasons for these changes are various. This list does not include 'scan' errors; only a few are thought to be errors of interpretation by Monier-Williams. These 700+ corrections are not visible as changes on the web, which might be confusing.

[35] https://github.com/sanskrit-lexicon/csl-corrections/blob/master/app/correction_response/cfr.tsv

[36] https://github.com/sanskrit-lexicon/csl-corrections/blob/master/dictionaries/mw/mw_printchange.txt

4.2.3 Types of Corrections

There are three types of changes at CDSL:

1) intentionally changed, print change means, for example: `The scan (print) has X, but CDSL has intentionally changed to Y in the digitization.`

The 'Additions and Corrections' to MW given on pp. 1308-1333 and containing ca. 4000 entries have been incorporated rather lately. Corrections given in reviews (e.g. Winternitz 1900) or private correction lists should also be considered for inclusion. The collection of private correction lists is still a task that we are only starting to work on.

2) unintentional web-only typos, left by Thomas typists.

The other type of change to Thomas's digitization is where the `print (scan) has X and the digitization HAD Y (a typo) and we change Y back to X (in agreement with print)`. These are called (informally) 'typos'. Most of them are not documented.

3) changes by Jim in the organization of data in XML.

Thomas was not initially XML, the original file was a TXT. Jim made several changes to XML, in that process some things where Thomas was in agreement with the print, the XML conversion introduced differences.

For example in PWG added spaces after numerals in a row, `zu P. 2, 4, 31.` instead of `zu P. 2,4,31.` Nāgabhūṣaṇarāvu disagreed and Jim removed the trivial difference, such changes were never documented. These changes are other than print changes, removal of (possibly) redundant elements when the digital copy was different to the printed text. Jim changed something after Thomas, Nāgabhūṣaṇarāvu insisted on having it as it was prior per print. Here is an example of a Jim-introduced difference to `mw_orig_utf8.txt` (Thomas version):

`<H4>100{aJjas}2{a4Jjas}¦ •n.| (#{as}) •ind. <quickly…,…instantly> <¯RV.> <¯BhP.>`

current CDSL in SLP1[37] = `aYjas`

`a/Yjas ind. quickly, instantly, RV.; BhP.`

Note: '`(#{as})`' missing in the current CDSL, was removed sometime by Jim.

5 Linking Dictionary and Corpus

By 2022 the batch headword correction was over and a new era started.

5.1 Digitization of Links Targets

[37] Humans usually can't read SLP1 easily, if you are not Jim Fundeburk. It's meant for the bots between us. Indian volunteers prefer Devanagari to IAST romanized transliteration. Russians and Westerners prefer IAST.

5.1.1 History of Sanskrit Lexicography

Sanskrit lexicography, as we know it, was founded in Saint-Petersburg, Russia. Saint-Petersburg remains the capital of Sanskrit studies until a bigger dictionary is printed, so it is not strange at all to see so many volunteers from Russia. The grandfathers of Petersburg Dictionaries had to be not only lexicographers and interpreters but also text editors. Many of the source texts were available to them only as manuscripts. Nobody knows what exactly MSS were available to them in the capital of Indology during the golden days of Sanskrit studies in St. Petersburg, Russian Empire.

If we take the date of the beginning of PWG and the ending of PWK from the years 1855-1889 (Brückner and Zeller 2007), there is more than a 34-year time span, and several new editions of texts were published in these years. And these works, to name a few like Kādambarī, Gobhilagṛhyasūtra, and Hāsyārṇava, were step by step introduced in the dictionaries.

Under the same abbreviation several editions in different volumes for different chapters might be used and one still has to document it, (Jachertz 1983) can be seen as the first attempt of its kind. The abbreviation lists in printed dictionaries are never full, even in the German ones. Some of the earlier editions were replaced with better editions at some stage: Rājataraṅgiṇī, Bhartṛhari and Pañcatantra. To include Bhartṛhari, for example, Böhtlingk had to prepare a collection of subhāṣita's in three volumes. He did it twice. The digitization of the 2nd edition of the voluminous work, Indische Sprüche, itself has been finished lately at Cologne[38] with the help of Thomas Malten, but still not yet widely announced as most of the work was done after 2013.

Several important Sanskrit works were available in MSS or printed only after the printing of PWG. Dramas of Bhāsa were found only in 1910 in South India as palm leaf manuscripts. Or Aśvaghoṣa's Buddhacarita, the first chapter of which was printed in 1892 by Sylvain Lévi (Jachertz 1983).

5.1.2 Critical Editions

Until 1933 no work on the preparation of critical editions for Mahābhārata and Rāmāyaṇa even started and only after 1966 were made available in Poona and Baroda. The critical edition of Mahābhārata itself was digitized at a very early time (Ingalls 1985) but remains unverified[39] still after 40 years and is not identical to the

[38] Digitization 2nd ed. of Indische Sprüche, 1870-1873
https://funderburkjim.github.io/boesp-prep/web1/boesp.html?6919

[39] As stated at the Computational Analyses of Mahābhārata website: 'Prof Muneo Tokunaga was the first person to key the whole text of BORI version into computer, which was later revised by

printed edition. No real work continues at BORI in regards to the The Critical Edition of Mahābhārata[40], the underway projects (Cultural Index) are stuck, and the future projects have no future (complete digitization of the Bibliography, Critical Apparatus, Cultural Index[41]) in reality, only on paper. Baroda is out of the game. What about Poona outside of BORI?

5.1.3 Digitization from India

Poona, the home for several Sanskrit dictionaries, decided to stay off the digital world[42] as a *kūpamaṃḍūka*. An Encyclopaedic Dictionary of Sanskrit on Historical Principles (reached the middle of the letter `a`), is made at Deccan College in an environment that has only an Intranet, and no Internet connection on intention. Dr. Prasad Joshi, the general editor since 2017, has a plan to prepare 200 pages for print per year so that it will not be published even around 3000 years from now. Infinity is the limit. Ending a dictionary is not a goal, it's a process. Something a German lexicographer will never understand. Mayrhofer, who accomplished the printing of two Sanskrit etymological dictionaries in the 20th century, as once Böhtlingk did, introduced a term that in English would sound like 'finishibality of a dictionary'. It's easy to start. It's hard to put an end to it.

But not every German dictionary has ever been finished[43]. If we look not only at the history of Sanskrit lexicography, The Thesaurus Linguae Latinae was one of many big, scholarly projects taken on by the German government in the late 19th century, it's been going on since 1894. But 20 researchers in Munich, Germany does not equal the 4 researchers out of the 40 wanted compared to Poona, India. In 2019 I visited Deccan College in person. And the Latin dictionary reached the letter 'N', at the same time the Sanskrit dictionary does not hurry to leave the letter 'A'.

Prof John Dargavel Smith. Prof Oliver Hellwig encoded the Mahabharata corpus that is based on BORI e-recension with resolved Sandhis and full morphological and lexical analysis.'

[40] https://bori.ac.in/department/mahabharata

[41] MCI, Mehendale Mahabharata Cultural Index, was digitized in 2013, uniting the efforts of University of Cologne and Brown University with NEH-DFG support. Converted to UTF-8 encoding in 2018. Line break removal in 2021.

[42] The Quark desktop publishing software, which is used since 1996 for one of the two big dictionaries, that will never be published in Poona, The Critical and Comprehensive Dictionary of the Prakrit Languages on Historical Principles (reached the letter 'ka'), still uses non-unicode fonts. Although Indian languages were first introduced into the Unicode standard around the time of the initial release in 1991, they did not get popular before 2000 and never reached Sanskrit institutes in most of India.

[43] But most have, see A Concise Etymological Sanskrit Dictionary, 1953-1980 (Mayrhofer, Manfred), a semi-digitized dictionary at https://samskrtam.ru/sanskrit-lexicon/KEWA/ and although I have Mayrhofer's permission to post it online, I'm not sure the printer would like to see it hosted on a German website anytime soon.

So there is no support for digitization to come from Poona. Nor from Manipal. The 'Computational Analyses of Mahābhārata' at Manipal Academy of Higher Education[44] has added Kumbhakonam, Sastri-Vavilla, and the smaller Tatparyanirnaya to BORI edition. The website has not been updated regularly, and the project, like most digital Sanskrit projects from India, is discontinued. But I'm sure a new day will come and India's digital Sanskrit resources have a bright perspective ahead.

5.1.4 Digitization for Dictionaries

But none of these mentioned editions of Mahābhārata were quoted in PWG in the mid-19th century. The Bombay edition of Mahābhārata, 1863, has been scanned only lately, 4500+ pages, which would be what we want from cross-references. To be able to add the 67138 links from dictionaries as PWG to Mahābhārata, we lack an index to the semi-digitised PDF scan. Same with the Bombay edition of Rāmāyaṇa, 1888. An elementary task for volunteers. Such deprecated, semi-digitized editions died in Cologne mostly until 2013 (Sanskrit and Tamil Dictionaries, 2005; Wilson Sanskrit-English Dictionary, semi-digitized edition, 2008; Monier-Williams Sanskrit-English Dictionary, 2008 and 2012 displays), but are revived once again. Volunteers from Russia have been adding indexes to such PDF files since 2020 (`@AnnaRybakovaT`), but one wonders if Sanskrit still might be of interest as a topic of research in India? What if there are other volunteers as productive as Nāgabhūṣaṇarāvu?

As of now, we suppose there are not more than 2800 literary sources scattered around seven volumes of PWG. The list, as with every other one, needs cleaning itself. Preparing lists is a great way of finding typoes with fuzzy search, referencing to the previous step, that is error correction in dictionaries, which took around ten years only for the list of headwords. 116 of the abbreviations lead us nowhere, they are only mentioned in one of the lists but never cited. 128 major books, that are quoted 500+ times in PWG sum to 682898 quotations. To name only the most widely quoted ones:

```
67138 MBH.   MAHĀBHĀRATA, ed. Calc. (GILD. Bibl. 93).
56037 ṚV.    ṚGVEDA. Es wird nach Maṇḍala, Sūkta und
37762 R.     RĀMĀYAṆA. Ohne eine nähere Angabe ist be
25299 P.     PĀṆINI'S acht Bücher grammatischer Regel
25005 KATHĀS.      KATHĀSARITSĀGARA, ed. BROCKHAUS (GILD. B
22563 M.     MANU'S Gesetzbuch in der Ausg. von LOISE
22119 BHĀG. P.     BHĀGAVATAPURĀṆA, nach Anführungen im VP.
20121 ŚKDR.  ŚABDAKALPADRUMA (GILD. Bibl. 371).
```

[44] https://mahabharata.manipal.edu

16204 AV. ATHARVAVEDASAM HITĀ, herausg. von R. ROT
16151 AK. AMARAKOṢA nach der Ausgabe von COLEBROOK
16148 H. HEMACANDRA'S ABHIDHĀNACINTĀMAṆI, ein sys
15702 HARIV. HARIVAṂŚA im 4ten Bande des MBH.
&#1
15431 ŚAT. BR. The ŚATAPATHABRĀHMAṆA in the Mādhyandina
14738 Verz. d. Oxf. H. AUFRECHT, Verzeichniss der Oxforder
12990 MED. MEDINĪKOṢA, ed. Calc. (GILD. Bibl. 258).[45]

5.1.5 Types of Link Targets

Link target typology and selected linking target samples:
1) to external PDF image files (precompiled index required)
1.1) with bookmarks (representing book structure)
1.1.1) https://sanskrit-lexicon-scans.github.io/ramayanagorr/?6,72,57
1.2) without bookmarks
1.2.1) https://sanskrit-lexicon-scans.github.io/manu/index.html?9,321
1.2.2) https://sanskrit-lexicon-scans.github.io/kss/index.html?20,216
1.2.3) https://sanskrit-lexicon-scans.github.io/mbhcalc/?13.1729 [46]
2) to external digital corpus collections
2.1) Sanskrit-Russian Parallel Corpus: Rigveda, Atharhaveda
2.1.1) https://gasyoun.github.io and lately https://samskrtam.ru/parallel-corpus/rigveda.html, https://samskrtam.ru/parallel-corpus/atharvaveda.html
2.1.2) Panini, P. 4,2,126.
https://ashtadhyayi.com/sutraani/4/2/126
2.2) DCS based, based on GRETIL (not yet tested)
3) to internal Sanskrit dictionaries and reference works:
3.1) koshas in shlokas
3.1.1) headwords level tagging done:
- AMARAKOṢA nach der Ausgabe von COLEBROOK, 16151 references
- HEMACANDRA'S ABHIDHĀNACINTĀMAṆI, 16148 references
- HALĀYUDHA, 5159 references

3.1.2) lacks manual tagging of headwords:
- MEDINĪKOṢA, ed. Calc., 12990 references
- VAIJAYANTĪ, ein Wörterbuch. Excerpte dar, Vaijayantīkośa №199 in the list with 266 references, is being tagged at Karnataka Samskrit University for the last century, but the work seems to never end.

3.2) modern bilingual dictionaries, SKD; reference works, Whitney Roots.

[45] https://github.com/sanskrit-lexicon/PWG/blob/master/pwgissues/issue74/lsextract_all.txt
[46] https://github.com/sanskrit-lexicon-scans/mbhcalc/tree/main/python

5.1.6 Semi-digitization Index

What is an index for semi-digitization?

Vol.	Page	Parva	Start	End	count
1		1	I	1	15	15
1		2	I	16	45	30
1		3	I	46	75	30
1		4	I	76	104	29
1		5	I	105	134	30
1		6	I	135	164	30
1		7	I	165	194	30
1		8	I	195	219	25

Every single PDF page get the needed metadata, so it can be linked with the help of regex.

Why it is important to master the citations starting from PWG? If done, it will automatically cover 90% of cases in MW as per Nāgabhūṣaṇarāvu intuitive guess. If counted, it's 41,6% of cases (see 2.3.2 Offline Sanskrit Analysis Tool). So it is not only 'Grundlage der Sanskritlexikologie im deutschsprachigen Raum' (Jachertz 1983), but it will remain the standard Sanskrit dictionary for the next 200 years and not only in German-speaking countries.

5.2.2 Unique Tags

Unique tags, 67 enlisted: <ab>, <arab>, <bio>, <bot>, <chg>, <cl>, , <ed>, <edit>, <etym>, <fr>, <ger>, <gk>, <heb>, <hom>, <i>, <info>, <is>, <iw>, <lang>, <lat>, <lbinfo>, <lex>, <ls>, <mong>, <ms>, <new>, <note>, <ns>, <nsi>, <old>, <pb>, <pcol>, <per>, <pic>, <rus>, <s1>, <s>, <sic/>, <srs/>, <sup>, <symbol>, <tib>, <toch>, <vlex>, <zoo>, <C1>, <C2>, <C3>, <C4>, <C5>, <C6>, <C7>, <C8>, <C9>, <C10>, <C11>, <C12>, <C>, <F>, <H>, <HI1>, <HI>, <P>, <Picture>, <Poem>, <VN>[47].

5.2.1 Importance of Citations

Some link dictionaries have link targets (= ls tags), and some doen't. Dictionaries ready for linking (seven dictionaries have markup pertaining to literary sources): AP90, BEN, BHS, GRA, MW, PWK, PWG. For example, Apte has many LS, but tags are not implemented. So there is a potential, but the work is yet to be done, if ever.

[47] https://github.com/sanskrit-lexicon/COLOGNE/blob/master/xmltag/all_xmltags.txt these are the tags as we have them now, per the dictionary.

6 Future Plans

We are still living by mid-19th century dictionary standards in the field of Indian lexicography. The corpus revolution in lexicography has not yet reached the field of Sanskrit and Cologne Digital Lexicons have to be the frontier when it will. If we are to make a two-century progress in the field of Sanskrit lexicography, evolution will not help, only a revolution can. The worst one can do when compiling dictionaries nowadays is to use introspection only. We need data (as found in corpora), and a linguistic theory to analyze and synthesize that data.

Since 2014 we have been getting ready for the transition to the orphan no-Jim mode. It started with questions related to Cologne server maintenance, a guide to sanskrit-lexicon.uni-koeln.de file system, and an intro to the email history of sanskrit-lexicon project. Until that final transition happens, we can still think of fixing bugs like broken links and internal inconsistencies, adding new link targets, adding new dictionaries, and developing sub-headword research. No one ever comes close to Jim Funderburk.

6.1 Additional Dictionaries

In 2025 the beginning of addition of 5 Sanskrit-Russian Dictionaries has been planned at Cologne, additionally to the already available Sanskrit-English, English-Sanskrit, Sanskrit-French, Sanskrit-German, Sanskrit-Latin Dictionaries. They have to be modified to comply with the DTD of XML at Cologne. The DTD or a schema is kept separate from both the unique article-specific data on the one hand and the repetitive metadata on the other, so truly instantaneous tailoring is effectively made possible for digital dictionaries.

The HTML results generated by an offline Russia-born Sanskrit search engine 'The Churning of the Ocean', based on the open corpora, could be integrated as well. Once the Digital Clay Sanskrit Library in 2006 planned for such search features. It never went open and was closed even before finalized, and the funding ceased. So, the German-based Cologne dictionary experience, together with Russian and Indian students crowdsourcing could become the safe and clean island of a verified Sanskrit text dataset, the golden standard.

6.1.1 Sanskrit-Russian Digital Lexicons

- K.A. Kossovich. Sanskritsko-russkiy slovar' (1854)

`https://samskrtam.ru/sanskrit-lexicon/kossowich/`

- D.N. Kudryavskiy. Sanskritsko-russkiy slovar' (1903)

`https://samskrtam.ru/sanskrit-lexicon/kudriavskiy/`

- F.I. Knauer. Slovar' from «Uchebnik sanskritskogo yazyka» (1908)

https://samskrtam.ru/sanskrit-lexicon/small/knauer_sm.html

- O. Frish. Sanskritskaya hrestomatiya. T. II. Slovar' (1956)

https://samskrtam.ru/sanskrit-lexicon/small/frish_sm.html

- V.A. Kochergina. Sanskritsko-russkiy slovar' (1987)

https://samskrtam.ru/sanskrit-lexicon/small/kochergina_sm.html

6.2 Abbreviation Markup

Of the more important ones for years ahead will remain the abbreviation markup, to literary sources and abbreviations for saving space in printed space. There is more to be done. We have not tagged all such places in the text, we have to change the scope of some of the tags, we have to revive the list of tags, and we have to understand what some of the abbreviations actually mean.

6.3 Compilation of New Sanskrit Dictionaries

Not every single digital library equals to a corpus, so we could raise the issues connected with the development of the biggest Sanskrit parallel text corpus, the Sanskit-Russian-English-German[48] Corpus has become the first and remains the main source for cross-reference links in the digital dictionaries at Cologne, linking dictionary back to corpora.

6.3.1 Reverse Sanskrit Dictionary

There are no new Sanskrit dictionaries planned, other than the Reverse Sanskrit Dictionary Gasūns plans as soon as the headword cleaning is over. And it never seems to be over. Not anytime soon, at least in the upcoming decade. So the 'duplication of effort in the compilation of new Sanskrit dictionaries be avoided' (Kapp and Malten, 1997) is something yet from a galaxy far away. 'Instead of including citations in the dictionary base of CDSL itself it will be sufficient to add pointers to external digital Sanskrit texts, where quotations can then be used in a much larger context' (Kapp and Malten, 1997) is ahead of our everyday needs.

6.4 Website Development

Lately, we have seen a massive drop in website visitors. Reaching a peak of 482,400 visitors in 2019, going down every year ever since. Until 2022 most visitors seem to have bookmarked it, then came searches, then links from websites.

[48] https://samskrtam.ru/parallel-corpus/

Search terms such as ‘monier williams’, ‘monier williams sanskrit dictionary’, ‘monier williams dictionary’, ‘moniel williams’ and ‘sanskrit dictionary’, ‘sanskrit english dictionary’ have almost disappeared.
Referrals from websites matter as well. Links from `spokensanskrit.de` and `spokensanskrit.org` disappeared years ago and together they were above other backlinks: `lexilogos.com`, `learnsanskrit.cc`, `sanskrit-sanscrito.com`, `en.wikipedia.org`.

6.4.1 Improving User Experience

One thing that would be important is URLs for each dictionary entry, which would make it possible to link an entry from say Wikipedia. Link format like `https://sanskrit-lexicon.uni-koeln.de/simple/mw/buddha` would increase the high-profile incoming links. So however ongoing is the correction of mistakes, however fast is the API, we still need to develop the functionality of the website from time to time. Cologne does not support Elastic search, so we should consider implementing C-SALT APIs[49]. Frontend and backend developers badly wanted. We still believe that voluntary crowdsourcing is the only future for Cologne and Sanskrit research in the upcoming millennia.

Acknowledgments

Jim Funderburk, the Cologne webmaster since 2004. Dhaval Patel, member of the core team since 2014, as of now correction validation. Nāgabhūṣaṇarāvu Kālepu from Hyderabad, an electronic engineer after retirement and his colleague Seshatalpa Sai, since 2021. Felix Rau from Data Center for the Humanities, University of Cologne, a previous student of Thomas Malten, supporting the server infrastructure, is our only link to the University of Cologne, which is otherwise nowadays uninvolved.

[49] https://cceh.github.io/c-salt_sanskrit_data and https://kosh.uni-koeln.de/cdsd/acc/restful

References

7.1 Key References

Report on the Cologne Sanskrit Dictionary Project, 1997
https://www.sanskrit-lexicon.uni-koeln.de/CDSL.pdf
Normalizing headwords of Cologne digital dictionaries, 2016
https://cse.iitkgp.ac.in/resgrp/cnerg/sclws/papers/patel.pdf
Latin Terms in Sanskrit Dictionaries, 2006
https://euralex.org/publications/latin-terms-in-sanskrit-dictionaries/
The Digital South Asia Library: Electronic Access to Seminal South Asian Resources, 1999
https://www.lib.uchicago.edu/e/su/southasia/dsal2.pdf
Apte, Vaman Shivaram. Revised and enlarged edition of Prin. V. S. Apte's The practical Sanskrit-English dictionary. Poona: Prasad Prakashan, 1957-1959. 3v.
Macdonell, Arthur Anthony. A practical Sanskrit dictionary with transliteration, accentuation, and etymological analysis throughout. London: Oxford University Press, 1929. "The original edition of this dictionary was published by Messrs. Longmans, Green & co., and has been reproduced photographically with their consent."
Otto Böhtlingk an Rudolf Roth: Briefe zum Petersburger Wörterbuch 1852–1885. Herausgegeben von Heidrun Brückner und Gabriele Zeller. Bearbeitet von Agnes Stache. Wiesbaden: Harrassowitz 2007.
Ingalls, Daniel H.H. and Daniel H.H. Ingalls 1985: The MahAbhArata: Stylistic study, computer analysis and concordance. In: Journal of South Asian Literature 20:17-46.
Mārcis Gasūns. Two hundred years of Sanskrit lexicography (1819-2019) / Zograf Readings "Problems of Traditional Indian Text", St. Petersburg, Russia, 2019
Mārcis Gasūns. Possibilities of understanding and orthography of the term “Sanskrit” / Roerich Readings, Institute of Oriental Studies of the Russian Academy of Sciences, Oriental Studies Readings, Moscow, 2019
Mārcis Gasūns. History of Cologne Digital Lexicons, 6th International Sanskrit Computational Linguistics Symposium, IIT Kharagpur, India, 2019
Mārcis Gasūns. Parallel Sanskrit-Russian Corpus / Zograf Readings "Problems of Traditional Indian Text", St. Petersburg, Russia, 2020
Alexey Kretov, Mārcis Gasūns, Vladimir Leonchenko. Parametric Analysis of the “Sanskrit-Russian Dictionary” by V.A. Kochergina / Institute of Oriental Studies of the Russian Academy of Sciences, Oriental Studies Readings, Moscow, 2021
Mārcis Gasūns. State of the Art Sanskrit Computational Linguistics. Institute of Oriental Studies of the Russian Academy of Sciences, Oriental Studies Readings, Moscow, 2021
Mārcis Gasūns. Blue lotuses of Sanskrit lexicography / Zograf Readings "Problems of Traditional Indian Text", St. Petersburg, Russia, 2021
Mārcis Gasūns. Sanskrit Triptych by V.A. Kochergina: Grammar, Reader, Dictionary / Indological Conference “The Dubyanskiy Readings”, Institute of Oriental and Classical Studies, HSE University, Moscow, 2021

Alexey Kretov, Mārcis Gasūns, Vladimir Leonchenko. Core Sanskrit Lexicon (according to Sanskrit-Russian dictionaries) / Tronsky readings, St. Petersburg, Russia, 2022
Mārcis Gasūns. Ancient Indian onomastics in the Vishnusahasranama / Zograf Readings "Problems of Traditional Indian Text", St. Petersburg, Russia, 2023
Mārcis Gasūns. The Development of Cologne Digital Sanskrit Lexicons. How voluntary crowdsourcing is changing the Sanskrit Internet / University of Vienna, Austria, 2024
Mārcis Gasūns. Future of Cologne Digital Lexicons, 7th International Sanskrit Computational Linguistics Symposium, Auroville, Puducherry, India, 2024
Mārcis Gasūns. Otto von Böhtlingk's correspondence about the Petersburg Dictionaries and the fate of the lexicographer / Zograf Readings "Problems of Traditional Indian Text", St. Petersburg, Russia, 2024
Mārcis Gasūns. Computational Linguistics Tools for Russian Sanskrit Scholars / "The Third International Conference Towards a Russian Language Buddhist Canon: Text Translation as a Dialogue of Cultures". Institute of Oriental Studies of the Russian Academy of Sciences, Oriental Studies Readings, Moscow, 2024
Mārcis Gasūns. Principles of construction of the garlands of ancient Indian names in “Vishnusahasranama” (Mahabharata XII.135) / Roerich Readings, Institute of Oriental Studies of the Russian Academy of Sciences, Oriental Studies Readings, Moscow, 2024
An Encyclopaedic Dictionary of Sanskrit on historical principles / general editor, A. M. Ghatage. Date: 1976-<1982>.
Peter Scharf and Malcolm Hyman. 2009. Linguistic Issues in Encoding Sanskrit. Motilal Banarsidass, Delhi.
Ladislav ZGUSTA, "Copying in Lexicography: Monier-Williams's Sanskrit Dictionary and Other Cases (Dvaikośyam)". Lexicographica 4 (1988): 145–164.
Agnes Stache-Weiske. 2015. "In ihrer rechten Hand hielt sie ein silbernes Messer mit Glöckchen ...": Studien zur indischen Kultur und Literatur. „Man muß zuweilen Insekten mit Kanonen schießen.“ Max Müllers Rolle im Streit zwischen Böhtlingk und Monier-Williams (pp. 323-336). Harrassowitz Verlag, Wiesbaden.
Winternitz, M. 1900: Sir M. Monier-Williams: {A Sanskrit-English Dictionary.} New Edition Oxford 1899. In: WZKM 14:353-360.

7.2 List of Dictionaries

WIL	1832	Wilson Sanskrit-English Dictionary
YAT *	1846	Yates Sanskrit-English Dictionary
GST *	1856	Goldstücker Sanskrit-English Dictionary
BEN *	1866	Benfey Sanskrit-English Dictionary
MW72 *	1872	Monier-Williams Sanskrit-English Dictionary
LAN	1884	Lanman's Sanskrit Reader Vocabulary
LRV	1889	Vaidya Sanskrit-English Dictionary
AP90 *	1890	Apte Practical Sanskrit-English Dictionary
CAE	1891	Cappeller Sanskrit-English Dictionary

MD	1893	Macdonell Sanskrit-English Dictionary
MW	1899	Monier-Williams Sanskrit-English Dictionary
SHS	1900	Shabda-Sagara Sanskrit-English Dictionary
AP	1957	Practical Sanskrit-English Dictionary, revised edition
PD *	1976	An Encyclopedic Dictionary of Sanskrit on Historical Principles
MWE *	1851	Monier-Williams English-Sanskrit Dictionary
BOR *	1877	Borooah English-Sanskrit Dictionary
AE	1920	Apte Student's English-Sanskrit Dictionary
BUR *	1866	Burnouf Dictionnaire Sanscrit-Français
STC *	1932	Stchoupak Dictionnaire Sanscrit-Français
PWG	1855	Böhtlingk and Roth Grosses Petersburger Wörterbuch
GRA *	1873	Grassmann Wörterbuch zum Rig Veda
PW	1879	Böhtlingk Sanskrit-Wörterbuch in kürzerer Fassung
CCS *	1887	Cappeller Sanskrit Wörterbuch
SCH	1928	Schmidt Nachträge zum Sanskrit-Wörterbuch
BOP *	1847	Bopp Glossarium Sanscritum
ARMH	1861	Abhidhānaratnamālā of Halāyudha
VCP *	1873	Vacaspatyam
SKD *	1886	Sabda-kalpadruma
ABCH	1896	Abhidhānacintāmaṇi of Hemacandrācārya
ACPH	1896	Abhidhānacintāmaṇipariśiṣṭa of Hemacandrācārya
ACSJ	1896	Abhidhānacintāmaṇiśiloñcha of Jinadeva
INM *	1904	Index to the Names in the Mahabharata
VEI *	1912	The Vedic Index of Names and Subjects
PUI *	1951	The Purana Index
BHS *	1953	Edgerton Buddhist Hybrid Sanskrit Dictionary
ACC *	1962	Aufrecht's Catalogus Catalogorum
KRM *	1965	Kṛdantarūpamālā
IEG *	1966	Indian Epigraphical Glossary
SNP *	1974	Meulenbeld's Sanskrit Names of Plants
PE *	1975	Puranic Encyclopedia
PGN *	1978	Personal and Geographical Names in the Gupta Inscriptions
MCI *	1993	Mahabharata Cultural Index