**Heße — Efficient**

<1 image>

**An Efficient Collaborative Web-Based Working Environment for the Creation of a Digital Sanskrit Dictionary**

Heße, S., Einicke, K. and Ritter, J.

We present an efficient web-based collaborative working environment for the creation of a new digital Sanskrit dictionary, making it possible to abstract from the structured representation of a dictionary entry and instead focus on content decisions when adding new entries through a convenient input mask.

\*\*\*

Sanskrit is, among other things, the liturgical language of Hinduism. With its textual transmission dating back to the 2nd millennium BC and written records in the form of inscriptions dating back as far as the 1st/2nd century AD, Sanskrit texts make up one of the richest cultural and intellectual archives of the pre-modern Asian world. At the end of the 19th century Böhtlingk and Roth published their groundbreaking Sanskrit dictionaries (Böhtlingk, 1855–1875; 1879), creating an indispensable tool in indological research and laying the foundation for modern lexicographical Sanskrit studies. They have since never been superseded. The only comprehensive effort to create an addendum based on the same scientific standards was made by Schmidt (1928). Since then a large number of significant scientific advances in Sanskrit lexicography have since been made. While the portfolio of dedicated large-scale dictionaries is manageable (Böhtlingk, 1855–1875; 1879; Apte, 1957–1959; Grassman, 1873; Monier-Williams, 1899) and largely accessible for research (Kapp and Malten, 1997), the exact opposite holds true for the enormous amount of individual lexicographical accomplishments that have been made over the course of generations. This cumulative knowledge, published in different forms and places (glossaries, specialized dictionaries, articles in journals and anthologies), evades targeted access by being scattered both temporally and editorially, making it nearly impossible to obtain an extensive overview of the current progress of studies of Sanskrit vocabulary and its semantics. In consequence, one has to largely rely on the state of knowledge from 1928 when translating Sanskrit texts. This deficit will most certainly cause indological research to qualitatively fall behind other philologies over time. To counteract this process, a systematic scientific revision of the accumulated advancements made in Sanskrit lexicography since 1928 is urgently needed.

Our project, Nachtragswörterbuch des Sanskrit (NWS), aims at creating a digital Sanskrit dictionary whose lemmatic content consists of about 150 publications, containing an estimated 11,000 pages worth of scientific collections of Sanskrit vocabulary. The goal is to make this content available in a single digital dictionary with uniform presentation and structure by systematically analyzing each lexicographic collection with academic expertise. This includes identifying, extracting, and transferring relevant content into a new and unified lemmatic structure without adding further information. A transcription is decidedly not part of our project. We identify the key challenges that we have to overcome as follows.

First and arguably the biggest challenge is the sheer amount of content. Due to the scientific assessment of the source material on a per-entry basis, an automated approach can be ruled out. Instead, our main priority is to make this workflow of identifying, extracting, and transferring lexicographic content into our dictionary as convenient and efficient as possible by providing a tailored working environment for the indologists on our team. The second challenge is due to the fact that the content of our dictionary consists of various different sources. In order to avoid duplication of information we include only the first (oldest) occurrence of duplicate content. Reviewing the source material by date of publication is one necessity for this; however, the challenge lies in being able to access the live state of the dictionary in order to look up previously entered content under the same headword, keeping in mind that this content may not necessarily have been entered by the same person but could have come from anyone on our team. This brings up the next challenge, which is the fact that multiple users will be working on the dictionary, quite possibly at the same time, needing access not only to the content they are working on themselves but to the whole dictionary in its current state.

To address these challenges and meet the necessary requirements we opted for a web-based collaborative working environment, enabling us to provide an up-to-date version of the dictionary to all team members. The most important feature of this working environment is the ability to add new entries to the dictionary using a tailor-made web input mask (see Figure 1). This allows abstracting from the internal deep structured database representation of a dictionary entry and instead working with a specialized user interface. This allows the indologists on our team to focus on the content as well as the actual lemmatic structure instead of how to represent that structure internally. Furthermore we exploit the fact that publications are predominantly being reviewed one at a time from front to back. Using that knowledge we can prevent the need to input repetitive information by automatically filling in certain input elements in the web mask when entering consecutive entries. Whenever we can’t automatically deduce information with certainty, we try to anticipate the input by offering context-sensitive suggestions that update themselves while the user is typing. Using this feature is an invaluable resource when working with scribal abbreviations of Sanskrit texts, which in our case already amount to over 2,400 in total. By starting to type any part of an unabridged Sanskrit texts name one can effectively search for the corresponding abbreviation and accept the correct suggestion only by using a couple of keystrokes. The same feature is also used in different contexts, e.g., when choosing among more than 250 literary references or when entering cross-references. While convenience features are certainly the most notable when using our working environment on a daily basis, another important aspect is the added benefit of automatically validating new entries before they are added to the dictionary.1

The most difficult part during development proved to be making the input mask adaptable enough to support the full range of expressiveness of the underlying deep structure of a dictionary entry, while offering a clear, intuitive, and uncluttered user interface. We achieve this by using an approach where the user starts with a basic form, dynamically adding new structural elements whenever needed. This allows for arbitrarily complex entries when needed but provides a minimalistic view by default. While this approach works just fine for the overall structure of the dictionary entry, a different method is needed when it comes to tagging single words or even just characters of a text inside an input field. To this end we developed a feature that allows tagging parts of texts inside an HTML text input field in a way best described as analogous to using a highlighter. By making a text selection using the mouse or keyboard, the user can then tag the selection by using a hotkey combination or clicking a button in a floating context-sensitive toolbar. As Sanskrit is predominantly written in Devanagari, transliteration support including automatic Unicode normalization is another feature built into the input mask. Furthermore, by allowing each user to view dictionary entries in their textual representation and flag them as correct or give a written opinion on why they are incorrect, proofreading becomes a group effort.

The structured content of our dictionary is being stored in a relational database whose schema we derived from a subset of the TEI-P5 guidelines on encoding dictionaries.2 Exporting to TEI conform XML is therefore possible. Furthermore we use the Ruby on Rails framework to generate different textual representations from the structured database content. This allows us to use a singular data source but have different textual representations of our dictionary entries without duplication of content.

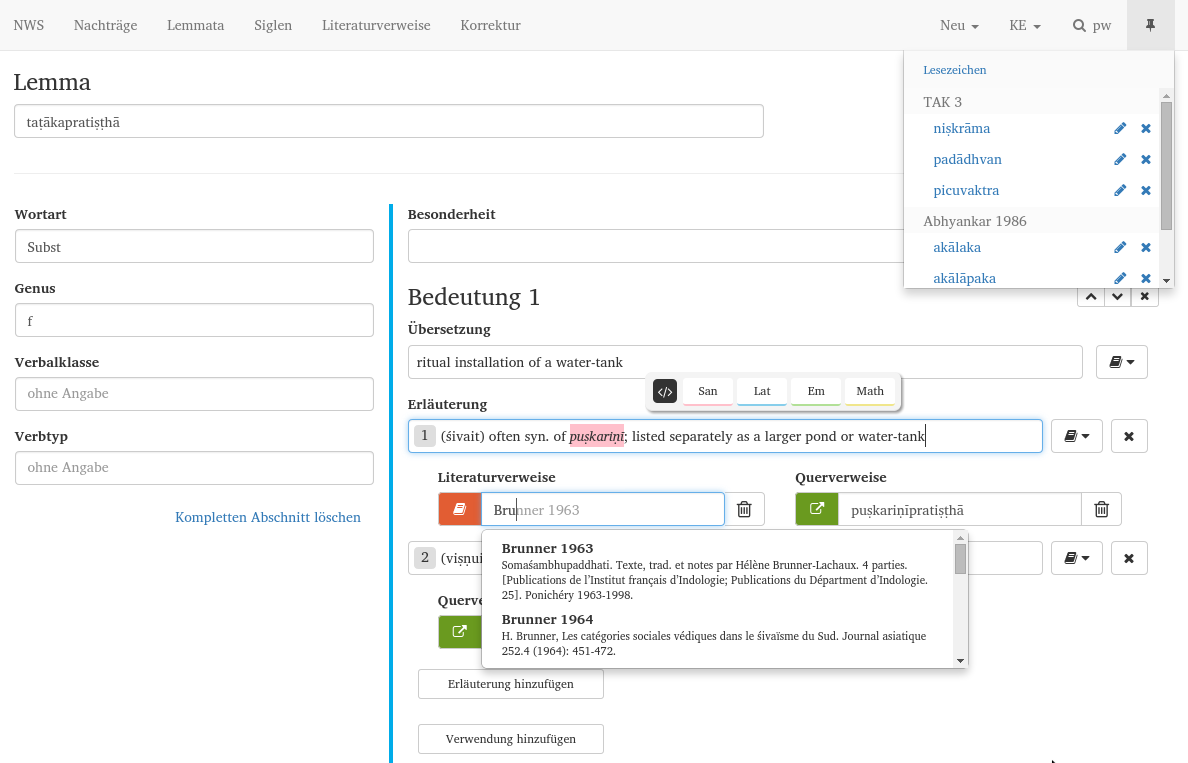


Figure 1. A screenshot of the input mask for new entries shows the autocompletion feature, the floating highlighter toolbar, as well as the personalized bookmarks menu in the top right.

**Results**

Development of this working environment initially took six months. We started testing our environment under live circumstances in June 2014. What was initially supposed to be a test phase ended up being the beginning of the use of our working environment in production. We have since iteratively integrated new and improved existing features. At the time of writing this, our team of four indologists has been successfully using this environment to review more than 30 sources, creating a total of nearly 10,000 dictionary entries. Looking at the time frame of our project, we are more than confident that this modern working environment will contribute to the timely completion of our digital Sanskrit dictionary.

**Future Work**

In the future we plan on integrating the content of the Cologne Digital Sanskrit Dictionaries3 into our dictionary, creating a unified digital Sanskrit dictionary. Furthermore, additional Sanskrit publications could be reviewed and integrated into the NWS as required, creating an even more comprehensive lexicographic Sanskrit resource. The possibility to realize such a project has been established with the development of this working environment.

**Funding**

This research was funded by the Deutsche Forschungsgemeinschaft (DFG) as part of the Kumulatives Nachtragswörterbuch des Sanskrit project, a cooperation between the Martin-Luther-University Halle and the Philipps-University Marburg under the direction of Prof. Dr. Walter Slaje, Prof. Dr. Jürgen Hanneder, Prof. Dr. Paul Molitor, and Dr. Jörg Ritter.

**Notes**

1. All literary references and citations are being checked for consistency. We’re currently not using a routine to check the whole entry for typing errors.

2. Text Encoding Initiative. P5: Guidelines for Electronic Text Encoding and Interchange, Section 9–Dictionaries.

3. Cologne Digital Sanskrit Dictionaries: http://www.sanskrit-lexicon.uni-koeln.de.

**References**

**Apte, V. S.** (1957–1959). The Practical Sanskrit-English Dictionary. Prasad Prakashan, Poona.

**Böhtlingk, O.** (1855–1875). Sanskrit-Wörterbuch. Kaiserliche Akademie der Wissenschaften, St. Petersburg.

**Böhtlingk, O.** (1879–1889). Sanskrit-Wörterbuch in kürzerer Fassung. Kaiserliche Akademie der Wissenschaften, St. Petersburg.

**Grassmann, H.** (1873). Wörterbuch zum Rig-Veda. Brockhaus, Leipzig.

**Kapp, D. B. and Malten, T.** (1997). Report on the Cologne Sanskrit Dictionary Project. *10th International Sanskrit Conference*, Bangalore, India.

**Monier-Williams, M.** (1899). *A Sanskrit-English Dictionary: Etymologically and Philologically Arranged with Special Reference to Cognate Indo-European languages.* Clarendon Press, Oxford.

**Schmidt, R.** (1928). Nachträge zum Sanskrit-Wörterbuch In Kürzerer Fassung von Otto Böhtlingk. Harrassowitz Verlag, Leipzig.