

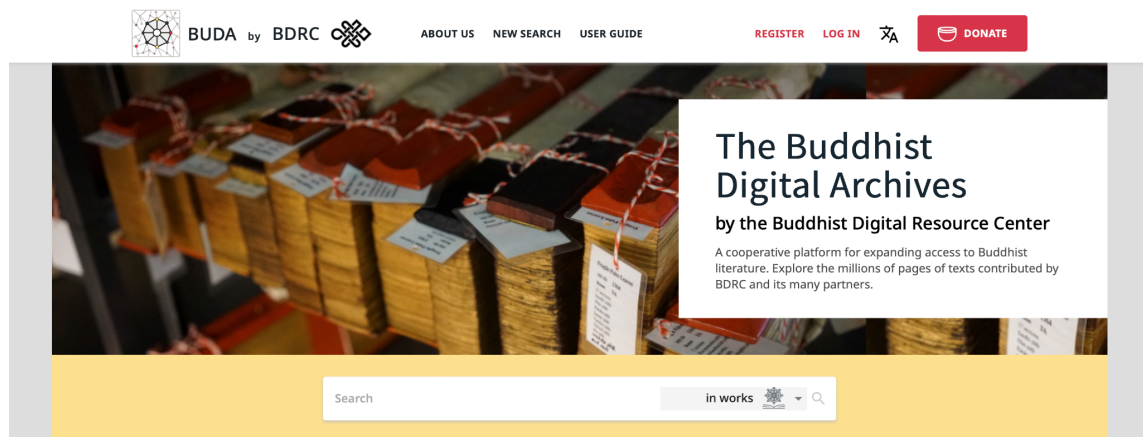
May 8, 2023

The Buddhist Digital Resource Center is honored to be considered for the prestigious Aming Tu Prize for an innovative and creative contribution to Digital Buddhist Studies. BDRC's nominator generously asserted that BDRC's new platform Buddhist Digital Archive (BUDA) is an outstanding recent addition to Buddhist Studies for both its practical benefits to researchers and its technological breakthroughs. Below we will introduce BUDA further to the judges of the Prize.

BDRC was founded as the Tibetan Buddhist Resource Center (TBRC) in 1999 with the goal of digitizing the personal library of our founder E. Gene Smith (1936-2010) and making the electronic texts available online for free. TBRC quickly established itself as the preeminent source of research materials for the field of Tibetan Buddhist Studies, with an unsurpassed database of information about Buddhist authors, places, subjects, and books. Its resources facilitated a major advance in our knowledge of Buddhism in Tibet and also helped redistribute Tibetan texts throughout the Tibetan population both inside and outside of the PRC. In 2015 the organization made the courageous and thoughtful decision to expand the mission of TBRC to digitize, catalog, and make accessible major collections of Buddhist texts in all of the canonical languages (i.e., Sanskrit, Chinese, and Pali). In particular, we wanted to do for other under-resourced living traditions of Buddhism what TBRC had accomplished for Tibetan Buddhism. To fulfill the expanded vision, TBRC became BDRC: Buddhist Digital Resource Center.

The [BUDA library platform](#), for which BDRC has been nominated for the Aming Tu Prize, is the embodiment of the BDRC vision to provide universal access to the Buddhist textual heritage. The BUDA platform was developed between 2018-2020 with generous funding from the Robert H. N. Ho Family Foundation and is still being enhanced by BDRC's technical lead Élie Roux. Researchers and practitioners from all fields of Buddhist Studies embrace BUDA because it provides them with a uniquely diverse and powerful resource. BUDA offers the most extensive collection of scans of Buddhist texts anywhere; over fifty-nine thousand volumes comprising more than 30 million images. BUDA's extensive collection of

texts in many languages, powerful search capabilities, free access to data and metadata to all texts in the public domain, and so forth, are the fruits of BUDA's skillful implementation of advanced technologies that lead the way in Buddhist Studies. In this short submission we will highlight the three central technologies behind BUDA: Linked Open Data, Computer Vision, and NLP.



BDRC was arguably the first major Buddhist resource to fully adopt Linked Open Data (LOD). It features an open-source implementation of a full LOD stack, including infrastructure and RDF editor. In developing BUDA as a Linked Data platform, significant efforts were made to transform BDRC's extensive legacy data, create a new data editor, evangelize LOD to the broader Buddhist Studies community, and expand the archive through data sharing. One of BUDA's first accomplishments with LOD was the formulation of the [Buddhist Digital Ontology](#), a flexible and shared ontology for representing Buddhist biographical, geographical, and bibliographical data that is interoperable with mainstream library conventions. It is the first implementation of an authoritative bibliographical model (BIBFRAME 2.0) in a Buddhist Studies context. One of the earliest partners to adopt BDRC's Buddhist Digital Ontology was Dharma Drum Institute of Liberal Studies.

In addition to formulating an entirely new ontology, BDRC also built for itself an [RDF editor](#). Despite LOD being promoted by leading humanistic digital collections, we found that there was no general-use, off-the-shelf, modern RDF editor for small digital humanities organizations like ourselves. We therefore created a web-based editor for the BUDA platform from scratch and have made the code completely independent from the rest of the platform so that it can be used as a generic RDF editor. The RDF editor ensures that all

of BDRC's new data is RDF native and as it is open source it also enables partner Buddhist organizations to more easily adopt this new standard for open data.

The users of BUDA benefit immensely from LOD because it has enabled BDRC to greatly expand the archive through IIIF protocols. BUDA's acquisition of significant collections of Sanskrit and Chinese languages materials was only possible because of our embrace of LOD. These, in addition to thousands of volumes of Tibetan texts accessed from major European and Asian libraries, are all deeply integrated into the BDRC database and search capabilities. Leveraging the riches of Buddhism made available through IIIF, BUDA offers the first large-scale integration of parallels in Tibetan, Chinese, and Indic texts so that researchers can access particular sutras in all three languages on a single site. What is more, BDRC understands LOD to be a two-way street and we proactively help Tibetan monasteries use IIIF to curate collections of BDRC scans to display on their websites.

In the development of BUDA, BDRC has also made innovative use of computer vision technologies in ways that further Buddhist research and benefit the broader field. Here we would like to highlight our large-scale and collaborative use of OCR for Tibetan texts in our archive. Since 2019 BDRC has leveraged Google's Cloud Vision OCR to convert over 3.3 million scanned images into e-texts and this process is ongoing. Because BDRC works at such a large scale we had to develop a sophisticated pipeline involving file distribution and retrieval processes with Google, and also a new e-text repository-cum-editor for the OCR output called [OpenPecha](#). To ensure that OCR-generated texts published on BUDA are sufficient for research purposes, we collaborated with a Harvard graduate student to develop a language model that reduces background noise and corrects OCR errors. About a third of the OCR'd images have been cleaned up using this model so far. BDRC also worked with Google to correct the unique errors that Cloud Vision made when processing Tibetan texts in the traditional pecha format, such as moving marginalia into the main text. Our mid-range plan is to OCR our entire Tibetan corpus, and to carry out the developments of language models and HTR that are needed to do so. BUDA enables scholars to search across its massive and growing e-text corpus with a single query, which is the most powerful research tool in Tibetan Buddhist Studies since the advent of TBRC itself.

BUDA has robust, multi-language search across its resources, which is one outcome of BDRC's work in NLP. BDRC has deployed a semantic query engine with Lucene language analyzers for [Tibetan](#), [Chinese](#), [Sanskrit](#), and [Khmer](#). The Sanskrit analyzer contains functions of syllable breaking, normalization and various transliterations (IAST, ALA-LC, Devanagari, SLP1). It also has a "lenient mode" in which the search ignores diacritics in the transliteration and normalizes geminates, vocalic r and anusvara. This is a rare feature and it makes the analyzer very attractive to other projects; [84000](#) have adopted it. The Chinese analyzer is finished and can convert from/to Simplified and Traditional Chinese characters (using data from Unicode), and to Pinyin. BUDA's Tibetan analyzer is particularly powerful. Additionally, BDRC added [Tibetan Collation Rules](#) for alphabetical sort order to the Unicode Common Locale Data Repository.

Research outputs and publications are a significant criteria for the Aming Tu Prize. BUDA not only enables a tremendous amount of research and publications. BUDA makes available millions of pages of rare Tibetan texts of great research value that are only available on BDRC (because BDRC digitized them on-site at local temples and they remain otherwise unpublished), in addition to providing convenient access to over a dozen editions of the Kangyur and Tengyur. BDRC's scans of palm leaf manuscripts in Burmese and Khmer scripts are likewise central to scholars of these fields of Buddhist Studies. In this guest [blog post](#) on BDRC, the scholar of Burmese Buddhism Petra Kieffer-Pülz discusses a recent discovery she made amongst BDRC's scans of Burmese texts and the resulting published research. Tibetan Translation projects large and small are also highly dependent upon, and greatly improved by, BUDA's archive.

BUDA is an ambitious and groundbreaking new platform that unites multiple advanced technologies to create a research resource that is equally vast and technically sophisticated. In developing BUDA, BDRC's team pushed the boundaries of Digital Buddhist Studies in several regards and it continues to evolve with the pace of technology. The Buddhist Studies community embraces it and the many collaborations we are engaged in are having a broad effect on the field at large. Receiving the Aming Tu Prize would be a great honor for BDRC and we hope that it will deepen the partnership between BDRC and Dharma Drum Institute for Liberal Arts.