The Journey of Digital Tibetan canonical collections

This short essay traces the history of creating Tibetan e-texts and explores the evolving methods, key outcomes, and profound implications for Tibetan Buddhist studies. This evolution can be understood through three main, often overlapping, phases.

Phase 1: Pioneering Input and Early Tools (Late 1980s – Early 2000s)

The late 1980s marked the dawn of Tibetan digital humanities, driven by a pressing need to preserve a rich literary heritage and the nascent potential of personal computing. The foremost challenge was representing the complex Tibetan script digitally. Early efforts, such as Ronald Schwartz's "Barkhang" programme (1983) and Pierre Robillard's "LTibetan" font system (1988), were crucial in enabling Tibetan script to be typed and displayed on computers.[1] These foundational font technologies, alongside others developed by pioneers like Tony Duff, paved the way for the creation of actual Tibetan e-texts.

The most significant undertaking in this initial phase was the **Asian Classics Input Project (ACIP)**, founded in 1987 by Geshe Michael Roach.[2] Inspired by projects like the Thesaurus Linguae Graecae, ACIP embarked on the monumental task of manually keyboarding vast swathes of Tibetan Buddhist literature, including much of the Kanjur and Tanjur. Computer centers were established in India, where Tibetan refugees were trained to input these texts, initially in a Romanized format (ASCII) and later transitioning to the Unicode standard for Tibetan script, which gained traction in the mid-1990s.[3] ACIP's first CD-ROMs, released in 1990, were met with great enthusiasm, demonstrating the immense demand for digital access to these scriptures. For over two decades, ACIP's work provided the largest publicly available archive of inputted Tibetan e-texts, forming the bedrock upon which many subsequent digital projects were built. Their focus was on comprehensive preservation and accessibility, creating a "first pass" digital capture of the canonical content.

[1] Hackett 2013: 93-95 details these early font developments.
[2] ACIP History (archived); Hackett 2013: 96.
[3] Hackett 2013: 97; Hackett 2003: 5.

Phase 2: Academic Engagement, OCR Development, and Resource Aggregation (Mid-1990s – Late 2000s)

As digital technologies matured, the focus expanded from basic input to broader academic utilization, the exploration of automated text capture through Optical Character Recognition (OCR), and the creation of platforms to organize and provide access to the growing digital resources.

Academic centers like **The Tibetan and Himalayan Library (THL)** at the University of Virginia became important nexuses for digital Tibetan studies, curating ACIP data for scholarly use and developing digital tools. Simultaneously, the **Tibetan Buddhist Resource Center (TBRC)**, founded by the visionary E. Gene Smith (now the Buddhist Digital Resource Center, BDRC), began its large-scale mission to scan and digitally preserve Tibetan texts as high-quality images.[4] While BDRC's primary output was these invaluable scanned facsimiles, they also experimented with OCR technology to convert these images into machine-readable text.

Developing reliable OCR for Tibetan script, with its stacked letters and cursive forms, was a significant hurdle. Early academic projects, such as one at Bell Labs, showed moderate success with typeset script.[5] Later, more specialized systems emerged, including the open-source C++OCRLib (RimeOCRLib) developed by V.A. Danilov and A.A. Stroganov, which was used to process hundreds of thousands of pages for TBRC and other institutions, including Kanjur texts.[6] Other systems like "Yakpo" and "Namsel" (developed at UC Berkeley) also contributed to the pool of OCR-generated text.[7] It is important to understand that OCR output, especially in its early stages, typically requires substantial manual proofreading to achieve scholarly accuracy. However, it provided a faster, if imperfect, way to generate initial inputted text from the vast image archives created by BDRC and others.

This period also saw focused efforts on digitizing specific important Kanjur editions. With the proliferation of digital images and nascent e-texts, tools for navigating this landscape became essential. Paul G. Hackett's **"Buddhist Canons Research Database"** hosted by the American Institute of Buddhist Studies (AIBS) at Columbia University, initiated in the late 2000s, aimed

[4] Hackett 2019: 98; Trinley et al. 2021: 2.
[5] Hackett 2019: 102.
[6] Danilov & Stroganov 2018.
[7] Hackett 2019: 102.

to be a comprehensive index to the Tibetan Buddhist canon. It provided cataloging information, cross-referencing different editions, and linking to BDRC's scanned images and ACIP's inputted e-texts, offering early full-text search capabilities across these resources.[8]

Phase 3: Critical Editions, Collaborative Platforms, Corpus Linguistics, and Advanced Research (Late 2000s – Present)

The current phase is characterized by a concerted effort to move beyond raw digital capture towards creating highly accurate, critically established, and richly annotated digital editions. This involves meticulous proofreading of earlier e-texts, the development of sophisticated linguistic corpora, and the creation of collaborative platforms and advanced lexicographical tools.

A pivotal moment in this phase was the "UVa-SOAS 2013 eKangyur" project. This initiative, a collaboration involving the University of Virginia and SOAS University of London, aimed to create a significantly improved digital version of the Derge Kanjur. It built upon earlier e-text work done at UVa and served as a crucial stepping stone. The outputs of this project became a key base layer for subsequent, more intensive proofreading and refinement efforts.

Building directly on this foundation, **Esukhia** has emerged as a leading organization, often working in partnership with BDRC. Esukhia took the e-texts from the UVA-SOAS 2013 eKangyur project and other sources (like BDRC's OCR output and ACIP texts) as starting points for their flagship projects: creating highly accurate digital editions of the Derge Kanjur and Tanjur. They continue to expanded this effort, focusing on meticulous, multi-pass proofreading and correction of Tibetan canonical collections texts.[9]

The development of advanced OCR and HTR (Handwritten Text Recognition) tools using machine learning has been a significant breakthrough. Platforms like Transkribus now host models specifically trained for Tibetan scripts, achieving remarkable accuracy in transcribing both printed texts and manuscripts. For example, rKTs' Namgyal collection has been processed and its model made available.[10] BDRC itself has continued to innovate in this area, developing

[8] Hackett 2013; http://aibs.columbia.edu/about.html.
[9] Esukhia-Barom Team 2019: 1; Trinley et al. (2021).
[10] Werner and Viehbeck 2024.

an **open-source Tibetan OCR application** that leverages deep learning to further improve recognition accuracy and accessibility of their vast image archive. While these AI-driven tools dramatically accelerate the initial conversion of scanned images into inputted text, though scholarly proofreading remains essential.

Crucially, these modern digital endeavors stand on the shoulders of earlier, traditional preservation efforts. The 16th Karmapa, Rangjung Rigpé Dorjé, played a pivotal role in the 20th century not only by sponsoring the reprinting of the Litang Kanjur and Tanjur in India, which ensured their physical survival and accessibility, but also by initiating a project to create a newly inputted digital e-text of the Kanjur at Rumtek Monastery. This project, started in the 1980s, involved manually typing the entire Kanjur, making it one of the earliest significant efforts by a traditional Tibetan institution to engage directly with digital text production for preservation and dissemination.[11]

Large-scale, state-level scholarly projects also contribute significantly. The **China Tibetology Research Center (CTRC) in Beijing**, through its Kangyur-Tengyur Collation Office, undertook the monumental task of producing the "Comparative Edition" (*dpe bsdur ma*) of the Kanjur (published 2006-2009) and Tenjur (published 1994-2005).[12] For the Kanjur, the CTRC collected eight different extant and accessible woodblock print editions: Derge, Yongle, Lithang, Beijing (Kangxi), Narthang, Cone, Outer Mongolian Khüree, and Zhol (Lhasa). Using the renowned Derge Kanjur as the primary base text, it was meticulously compared with the other seven editions. Discrepancies such as omissions, additions, differences, repetitions, or disordered sequences were recorded as collation notes and published, aiming for an edition of high quality, authoritativeness, and research utility. While the primary output of this project has been these invaluable scholarly *print* editions, it is certain that internal digital (inputted) versions were created for the complex typesetting and collation process. The public availability of these underlying inputted e-texts in open and manipulable formats has been limited, with some CTRC-affiliated portals offering highly controlled, view-only digital access to certain texts.[13] Nevertheless, the scholarly work embodied in the *Dpe bsdur ma* print editions provides an unparalleled reference for any future high-quality digital critical edition of the Tibetan canon.

[11] Tomlin 2020; Adarshah Lijiang Kangyur (etext).
[12] The Tibetan Tripitaka Collation Bureau of the China Tibetology Research Center, ed. 2006-2009.
[13] Trinley et al. 2021: 5.

Bibliography

Budapesti, Ilona. 2019. ‘Past, Present, and Future of Digital Buddhology’. In *Digital Humanities and Buddhism*, edited by Daniel Veidlinger, 25–40. De Gruyter. https://doi.org/10.1515/9783110519082-002.

Danilov, V. A., and A. A. Stroganov. 2018. “Optical Character Recognition (OCR) in C++: OCRLib”. Gomde OCR. https://github.com/RimeOCRLIB/OCRLib [Software documentation].

Esukhia-Barom Team. 2019. ‘Instruction Manual for the Electronic Database of Derge Edition Tibetan Tripitaka [德格版电子大藏经数据库使用说明]”. Esukhia-Barom. https://buddha.now/wp-content/uploads/2019/06/文件六-德格版电子大藏经数据库使用说明_简体版.pdf.

Hackett, Paul. 2003. ‘An Entropy-Based Assessment of the Unicode Encoding for Tibetan’. *Tibetan Information Technology Panel*.

———. 2013. “Digital Resources for Research and Translation of the Tibetan Buddhist Canon”. Preprint.

———. 2019. ‘Digital Encoding, Preservation, Translation, and Research for Tibetan Buddhist Texts’. In *Digital Humanities and Buddhism*, edited by Daniel Veidlinger, 91–110. De Gruyter. https://doi.org/10.1515/9783110519082-006.

Tomlin, Adele. 2020. ‘The Kangyur and the Gyalwang Karmapas’ Role in Their Publication and Preservation’. *Dakini Translations and Publications* མཁའ་འགྲོ་མའི་ལོ་ཙྪ་བའི་འགྱུར་དང་འགེམས་སྐེལ། (blog). 2020. https://dakinitranslations.com/2020/10/24/the-kangyur-and-the-karmapas-important-role-in-their-publication-and-preservation/.

The Tibetan Tripitaka Collation Bureau of the China Tibetology Research Center, ed. 2006-2009. བཀའ་འགྱུར་དཔེ་བསྡུར་མའི་གསལ་བཤད་སྒྲན་ལེལ་སྒྲོན་མེ། Vol. 1. 109 vols. བཀའ་འགྱུར་དཔེ་བསྡུར་མ། Beijing: Krung go’i bod rig pa zhib ’jug ste gnas kyi bka’ bstan dpe sdur khang.

Trinley, Ngawang, Tenzin, Dirk Schmidt, Helios Hildt, and Tenzin Kaldan. 2021. ‘Taming the Wild Etext: Managing, Annotating, and Sharing Tibetan Corpora in Open Spaces’. *ACM Transactions on Asian and Low-Resource Language Information Processing* 20 (2): 1–23. https://doi.org/10.1145/3418060.

Werner, Eric, and Markus Viehbeck. 2024. ‘Namgyal Manuscript Collection Datasets’. Zenodo. https://doi.org/10.5281/ZENODO.14247730.