



Published in the Russian Federation
Oriental Studies (Previous Name: Bulletin of the Kalmyk Institute
for Humanities of the Russian Academy of Sciences)
Has been issued as a journal since 2008
ISSN: 2619-0990; E-ISSN: 2619-1008
Vol. 15, Is. 4, pp. 740–750, 2022
Journal homepage: <https://kigiran.elpub.ru>



УДК / UDC 930.2

DOI:10.22162/2619-0990-2022-62-4-740-750

Цифровая трансформация письменного наследия тибетского буддизма: состояние и перспективы

Борис Ванданович Базаров¹, Олег Сергеевич Ринчинов², Андрей Александрович Базаров³

¹ Институт монголоведения, буддологии и тибетологии СО РАН (д. 6, ул. Сахьяновой, 670047 Улан-Удэ, Российская Федерация)

доктор исторических наук, академик РАН, профессор, директор

 0000-0001-5326-1317. E-mail: bazarov60@mail.ru

² Институт монголоведения, буддологии и тибетологии СО РАН (д. 6, ул. Сахьяновой, 670047 Улан-Удэ, Российская Федерация)

доктор исторических наук, ведущий научный сотрудник

 0000-0003-1379-771X. E-mail: o.rinchin@gmail.com

³ Институт монголоведения, буддологии и тибетологии СО РАН (д. 6, ул. Сахьяновой, 670047 Улан-Удэ, Российская Федерация)

доктор философских наук, ведущий научный сотрудник

 0000-0002-6837-7280. E-mail: bazarow_andr@mail.ru

© КалмНЦ РАН, 2022

© Базаров Б. В., Ринчинов О. С., Базаров А. А., 2022

Аннотация. *Введение.* Освоение достижений буддийской культуры представляется важным и актуальным ввиду неуклонного роста глобального значения региона Восточной и Центральной Азии. Поэтому обращение к памятникам буддийского наследия, хранящимся в обширных российских коллекциях, на современном академическом и технологическом уровне с упором на активное применение цифровых технологий позволит найти решение многих задач социального и культурного развития России, укрепит историческую роль и место нашей страны в геополитическом пространстве Азии. *Целью* обсуждаемого исследования является анализ современного состояния цифровизации письменного наследия тибетского буддизма и оценка перспективных направлений ее дальнейшего развития, связанных с использованием технологий искусственного интеллекта. *Задачи и методы.* Изучение современных подходов в этой области выполнено на основе сравнительного анализа деятельности консорциума Buddhist Digital Resource Center (BDRC), одной из старейших и наиболее успешных инициатив, действующей с 1999 г. В качестве перспективного направления рассмотрено использование методов искусственного интеллекта для решения проблемы создания корпуса машиночитаемых тек-

стов буддийских сочинений и машинного перевода. *Результаты.* Раскрыты принципы оцифровки и представления письменных памятников, реализованные в новой цифровой платформе BUDA (Buddhist Digital Archives), запущенной в 2021 г. Показана эффективность лежащих в ее основе архитектур IIF (International Image Interoperability Framework) и LOD (Linked Open Data) для интеграции ранее разрозненных цифровых хранилищ в глобальное информационное пространство буддийского культурного наследия. Для оценки перспектив использования технологий искусственного интеллекта было выполнено пилотное исследование по применению нейросетевых методов «глубокое обучение» для создания экспериментальной модели оптического распознавания символов тибетской письменности, в ходе которого достигнут результат 94 % распознанных символов. Таким образом, заложена база для решения проблемы поточного декодирования сканированных письменных источников на тибетском языке и создания всеобъемлющего корпуса буддийских сочинений в машиночитаемом формате, что открывает перспективы создания новых эффективных цифровых текстологических инструментов. *Выводы.* Интеграция существующих и перспективных подходов позволяет осуществить цифровую трансформацию буддийских письменных традиций, что открывает для них новые формы функционирования и новые возможности развития в современном мире и предоставляет возможность в полной мере раскрыть культурный и интеллектуальный потенциал буддийской цивилизации для современного общества.

Ключевые слова: цифровизация, тибетский буддизм, письменное наследие, искусственный интеллект, глубокое обучение

Благодарность. Статья выполнена в рамках государственного задания № 121031000263-3 «Письменные традиции народов Байкальского региона в контексте историко-культурного наследия России и Внутренней Азии».

Для цитирования: Базаров Б. В., Ринчинов О. С., Базаров А. А. Цифровая трансформация письменного наследия тибетского буддизма: состояние и перспективы // *Oriental Studies*. 2022. Т. 15. № 4. С. 740–750. DOI: 10.22162/2619-0990-2022-62-4-740-750

Digital Transformation of Written Heritage of Tibetan Buddhism: State of the Art and Prospects

Boris V. Bazarov¹, Oleg S. Rinchinov², Andrei A. Bazarov³

¹ Institute for Mongolian, Buddhist and Tibetan Studies of the Siberian Branch of the RAS (6, Sakhyanova St., 670047 Ulan-Ude, Russian Federation)

Dr. Sc. (History), Full Member of the RAS, Professor, Director

 0000-0001-5326-1317. E-mail: bazarov60@mail.ru

² Institute for Mongolian, Buddhist and Tibetan Studies of the Siberian Branch of the RAS (6, Sakhyanova St., 670047 Ulan-Ude, Russian Federation)

Dr. Sc. (History), Leading Research Associate

 0000-0003-1379-771X. E-mail: o.rinchin@gmail.com

³ Institute for Mongolian, Buddhist and Tibetan Studies of the Siberian Branch of the RAS (6, Sakhyanova St., 670047 Ulan-Ude, Russian Federation)

Dr. Sc. (Philosophy), Professor, Leading Research Associate

 0000-0002-6837-7280. E-mail: bazarow_andr@mail.ru

© KalmSC RAS, 2022

© Bazarov B. V., Rinchinov O. S., Bazarov A. A., 2022

Abstract. Introduction. Making public knowledge pertaining to Buddhist culture proves important and topical enough due to the steady growth of global significance attributed to East and Central Asia. Therefore, insights into monuments of Buddhist heritage represented in vast collections of Russia — at advanced academic and technological levels with an emphasis on the active use of

digital technologies — shall yield solutions to quite a number of sociocultural development problems, and strengthen positions of our nation in Asia's geopolitical environments. *Goals.* The study aims at analyzing the current state of the art in digitization of Tibetan Buddhism's written heritage and assessing some promising AI related technologies development trends. *Materials and methods.* The study provides a comparative analysis of activities by the Buddhist Digital Resource Center (BDRC) consortium, an oldest and most successful initiatives operating since 1999, for advanced approaches in this field. The use of artificial intelligence methods to solve the problem of creating a machine-readable text corpus of Buddhist writings and machine translation is considered as a promising direction. *Results.* The work discloses key principles of digitization and representation of written monuments implemented on the new digital platform BUDA (Buddhist Digital Archives), shows effectiveness of the underlying IIF (International Image Interoperability Framework) and LOD (Linked Open Data) architectures. The efforts to assess artificial intelligence technologies prospects included a pilot study on the use of 'deep learning' neural network methods to create an experimental model for optical recognition of Tibetan characters which yielded a result of 94 % of recognized characters. This lays a foundation for flow-line decoding of Tibetan-language scans and creating a comprehensive corpus of Buddhist writings in a machine-readable format, which offers the challenge of developing new efficient digital textual tools. *Conclusions.* The integration of existing and promising approaches allows for a digital transformation of Buddhist written traditions, the latter to open new functioning forms and development opportunities in the contemporary world backed with a possibility to reveal the cultural and intellectual potential of Buddhist civilization for modern society to the full.

Keywords: digitization, Tibetan Buddhism, written heritage, artificial intelligence, deep learning

Acknowledgements. The reported study was funded by government assignment, project no. 121031000263-3 'Scriptural Traditions of Baikal Peoples in the Context of Historical and Cultural Heritage of Russia and Inner Asia'.

For citation: Bazarov B. V., Rinchinov O. S., Bazarov A. A. Digital Transformation of Written Heritage of Tibetan Buddhism: State of the Art and Prospects. *Oriental Studies*. 2022; 15(4): 740–750. (In Russ.). DOI: 10.22162/2619-0990-2022-62-4-740-750



Введение

Научная значимость и актуальность изучения письменного наследия тибетского буддизма определяются огромным влиянием буддийской культуры в странах Восточной и Центральной Азии. С одной стороны, буддийская культура имеет важное структурообразующее значение для социально-политической жизни в азиатских обществах, обеспечивая непрерывность передачи духовных знаний и традиций, сохранение культурной и национальной идентичности. В то же время, благодаря универсальным идеям и ценностям, она способствовала сложению общего культурного пространства, развитию новых направлений межкультурного взаимодействия, технологических и торгово-экономических обменов, включая такие яркие исторические явления, как «Шелковый путь», «Чайный путь» и др. В связи с неуклонным ростом значения этого макрорегиона в современном глобальном мире освоение достижений буддийской

культуры представляется важным и актуальным.

Россия имеет длительный опыт проведения взвешенной политики межкультурного и межконфессионального взаимодействия, в том числе и с народами, исповедующими буддизм и имеющими многосторонние связи со странами буддийского ареала. В современных условиях использование такого подхода предоставляет России существенные преимущества при формировании стратегии «мягкой силы» в стратегически важном макрорегионе Восточной и Центральной Азии, учитывающие региональный историко-культурный контекст. В этой связи обращение к памятникам буддийского наследия, представленных в обширных российских коллекциях, на современном академическом и технологическом уровне с упором на активное применение цифровых технологий позволит найти решение многих задач культурного развития России, укрепит историческую роль и место нашей

страны в геополитическом пространстве Азии.

В фондах Центра восточных рукописей и ксилографов (ЦВРК) Института монголоведения, буддологии и тибетологии Сибирского отделения Российской академии наук (ИМБТ СО РАН) хранится одно из крупнейших в мире собраний раритетных рукописей и ксилографов на тибетском языке, основу которого составляют многочисленные комплекты буддийского канона *Кангьюр* и *Тенгьюр*. Проводимые здесь исследования тибетоязычных памятников письменности с применением цифровых технологий вносят вклад в решение важных для науки и общества проблем освоения культурного наследия народов буддийской Азии, включающих анализ и реконструкцию философских, этических и естественнонаучных концепций, которые получили отражение в корпусе буддийской книжности и традиционной медицине.

Цифровизация колоссального собрания буддийских знаний в фондах ЦВРК требует внимательного изучения международного опыта, накопленного к настоящему времени. Необходима оценка текущего состояния в этой области, выявление наиболее успешных реализованных подходов. На основе анализа достижений и недостатков существующих инициатив по цифровизации тибетского книжного наследия определяются наиболее перспективные направления, обеспечивающие приоритет российской науки и устойчивое развитие тибетологических исследований в ИМБТ СО РАН.

Задачи и методы исследования

Целью исследования являются анализ современного состояния цифровизации тибетского письменного наследия и оценка перспективных направлений исследований в этой области.

Перед исследованием было поставлено две основные задачи. Первой являлось, используя сравнительный метод, изучение современных подходов к цифровизации тибетского книжного наследия на примере одной из наиболее успешных инициатив в этой области — Buddhist Digital Resource Center (BDRC), их тщательный анализ, выявление преимуществ и существующих проблем. Для этого были подробно рассмотрены информационные ресурсы, представленные

на веб-портале BDRC. Второй задачей являлась оценка перспектив использования технологий искусственного интеллекта для решения проблемы создания корпуса машиночитаемых текстов буддийских сочинений и машинного перевода. Для этого была налажена коллаборация между ИМБТ СО РАН и Новосибирским государственным университетом (НГУ), в составе которой были выполнены междисциплинарные исследования по применению нейросетевого метода «глубокого обучения» для создания экспериментальной модели оптического распознавания символов тибетской письменности. В качестве материала для реализации пилотного проекта были взяты памятники письменности из тибетского фонда Центра восточных рукописей и ксилографов ИМБТ СО РАН [ЦВРК].

Современное состояние цифровизации тибетской книжности

Цифровизация буддийского письменного наследия началась более 30 лет назад усилиями религиозных общин и академических кругов в разных странах. За прошедший период этот процесс стал настолько интенсивным и всеобъемлющим, что в настоящее время становится очевидным: подобно освоению письменности и печати в ранние периоды, цифровизация переформатирует буддийское информационное пространство, делая его более связным и открытым.

Крупные и технологически продвинутые инициативы в области цифровизации тибетской книжности осуществляются западными про-тибетскими академическими кругами с широким привлечением представителей коренных этнических и религиозных сообществ. Среди них необходимо упомянуть Asian Classics Input Project (ACIP), действующий с 1988 г., Tibetan and Himalayan Digital Library (THL, с 2000 г.), Buddhist Digital Resource Center (BDRC, с 1999 до 2016 гг. — Tibetan Buddhist Resource Center, TBRC). Эти инициативы различаются организационными и технологическими подходами. Например, консорциум ACIP проявляет большую социальную ориентированность, активно вовлекая в свою деятельность местные сообщества. BDRC нацелен в существенной степени на запросы религиозного сообщества, в то время как THL исходит преимущественно из академи-

ческих интересов. Общей чертой является определяющее значение академического фактора [Ринчинов 2020: 649].

На наш взгляд, наибольших успехов в цифровизации письменного наследия тибетского буддизма достиг консорциум BDRC [About us 1999], основанный американским буддологом Джинном Смитом (E. Gene Smith, 1936–2010) с целью сохранения тибетского литературного наследия и обеспечения его максимальной доступности. Дж. Смит полагал, что традиционные формы передачи находятся в критической опасности в основном из-за неблагоприятных политико-социальных факторов. Исходя из этого, BDRC приступил к поиску памятников буддийской письменной культуры, а для их сохранения, организации и распространения упор сделал на широкое использование цифровых инструментов. К деятельности консорциума присоединилась Служба цифрового хранения (DRS) Гарвардского университета, являющаяся крупнейшим хранилищем культурно и научно значимых цифровых ресурсов. Создаваемая совместными усилиями цифровая библиотека тибетских текстов со временем стала важным ресурсом для тибетологии и буддологии, и вокруг BDRC сложилось активное сообщество академических ученых, религиозных деятелей, издателей и заинтересованной широкой публики. Каждый год к ресурсам BDRC обращаются из 150 стран мира [Access 2012].

За более чем 20 лет BDRC оцифровал значительное число буддийских сочинений, включая редкие рукописные и ксилографические издания, полученные самыми разными способами. Круг партнеров BDRC включает Библиотеку Конгресса США, Библиотеку тибетских трудов и архивов (LTWA) в Дхарамсале (Индия), управляемую Центральной тибетской администрацией, монастырские и университетские библиотеки из разных стран, где хранятся памятники тибетской литературы.

Одной из характерных черт BDRC является постоянное развитие поддерживаемой им цифровой платформы «вширь» и «вглубь». Развитие «вширь» обеспечивается за счет отбора и оцифровки как можно большего количества буддийских книжных материалов, включая и современные издания классики. Помимо сканированных

книг, представленных аннотированными цифровыми изображениями, BDRC занимается обработкой полнотекстовых материалов, представленных в машиночитаемых форматах, которые поступают из нескольких источников. Во-первых, электронные тексты предоставляют архивы и библиотеки в разных странах, выступающие партнерами BDRC, отдельные тибетские авторы и издатели. Во-вторых, полнотекстовые материалы создаются в рамках программы Namsel OCR (оптического распознавания символов), реализуемой в Калифорнийском университете в Беркли [Rowinski, Keutzer 2016].

Оба типа материалов, как сканированные, так и полнотекстовые, снабжаются разветвленными индексами по разным категориям и взаимосвязаны. При этом сканированные источники наиболее близки к оригиналу и сохраняют историко-культурный контекст. Начиная с 2016 г. BDRC занимается также интеграцией на своей платформе буддийских памятников на других языках, помимо тибетского. Выполнена оцифровка значительного объема писаний на санскрите, пали, бирманском, кхмерском, китайском языках [BUDA].

Развитие «вглубь» происходит за счет постоянного обновления технологической основы цифровой платформы, внедрения все более совершенных поисковых и аналитических инструментов. Основными этапами оцифровки являются сканирование, аннотирование (каталогизация) и индексирование для эффективной работы поисковых инструментов. Сканированные изображения размещаются в безопасном архиве, при этом пользователям обеспечивается открытый многоуровневый доступ к оцифрованным текстам. При аннотировании тексты размечаются метаданными, включающими название, автора, предмет, информацию о месте и времени издания и т. д. Принципы систематизации, предложенные Дж. Смитом, позволяют структурировать буддийскую литературу с учетом культурного, исторического, географического контекста. Обработываемые материалы индексируются по нескольким категориям, в существенной степени соответствующим методологии авторитетного контроля, включая обширные предметные таксономии, персоны, места. Благодаря этому создается глубокое

представление сложного и многогранного мира тибетской литературы [Cabezón, Jackson 1996: 29–31].

Этому способствует также возможность интеграции ресурсов цифровой платформы BDRC с информационными системами, которые создаются другими лицами и организациями. Доступ к хранилищу оцифрованных текстов со сторонних веб-сайтов предоставляется через открытый интерфейс доступа к данным на основе XML, позволяющий выполнять внешние программные запросы к базе данных BDRC.

В 2021 г. BDRC запустил новую цифровую платформу Buddhist Digital Archives (BUDA). Построенная с применением самых современных технологий, BUDA значительно улучшает доступ к обширной коллекции произведений тибетского буддизма и разнообразным материалам на санскрите, китайском, пали, бирманском и кхмерском языках [BUDA].

BUDA интегрирует множество инструментов, которые способствуют глубокому изучению буддийских текстов. Наиболее значительные улучшения касаются средств просмотра изображений (Image Viewer) и электронного текста (E-text Viewer). Чтение электронного текста в E-text Viewer сопрягается с одновременным просмотром сканированных изображений оригинального издания. BUDA предлагает новый многоязычный поисковый интерфейс, позволяющий пользователям формировать сложные запросы и получать наиболее релевантную информацию из цифрового архива BDRC.

Новая цифровая платформа построена на принципах открытого исходного кода. В ее основе лежат две основные архитектуры — это International Image Interoperability Framework (IIIF) и Linked Open Data (связанные открытые данные, LOD). Стандарты IIIF применяются для управления цифровыми изображениями и их отображения. Пользователям BUDA предоставляется многофункциональный интерфейс, который позволяет загружать, публиковать, встраивать и комментировать изображения.

Инициатива International Image Interoperability Framework (IIIF), выдвинутая в 2015 г. международным консорциумом, организаторами которого выступили Оксфордский и Стэнфордский университеты и Британская библиотека, нацелена на

внедрение единообразных способов описания и отображения цифровых образов культурных артефактов, таких как книги, карты, свитки, рукописи, нотные записи, архивные материалы, и обеспечение их глобальной доступности [IIIF].

Это достигается путем разработки и внедрения общих стандартов и инструментов представления изображений и других мультимедийных материалов, предназначенных для организации M2M взаимодействия (Machine-to-Machine) между репозиториями (цифровыми хранилищами) изображений через Интернет. Важным направлением деятельности консорциума является развитие совместно используемой распределенной инфраструктуры, включающей специализированные серверы изображений и веб-клиенты [Emanuel 2018: 127].

В настоящее время в рамках технологии IIIF представлены несколько интерфейсов прикладного программирования [API]. Они предоставляют пользователям богатые возможности поиска, просмотра, сравнения и аннотирования цифровых изображений и их коллекций, организации совместной работы. Например, IIIF Presentation API позволяет описать и представить составной цифровой объект, состоящий из нескольких изображений или мультимедийных материалов, разнесенных физически. Эта функция чрезвычайно полезна для часто встречающихся случаев, когда фрагменты одного документа оказываются в разных хранилищах. Это могут быть отдельные листы многостраничной книги, фрагментированные листы древней рукописи и т. д. Такая ситуация характерна, например, для гандхарских рукописных коллекций [Salomon 1997: 354].

IIIF позволяет исследователям обращаться с составными документами как с единым целым вне зависимости от физического местоположения их частей. Другие API предназначены для организации межмашинного взаимодействия и интеграции распределенных ресурсов и баз данных.

Архитектура связанных открытых данных (LOD) предназначена для структурирования машиночитаемых данных, что облегчает объединение, извлечение и обновление сложноорганизованных массивов информации на семантическом уровне. Например, архитектура LOD, реализованная на платформе BUDA, позволяет представить ос-

новые аспекты буддийского письменного наследия в контекстах актора (кто?), объекта (что?), времени (когда?) и места (где?) [BUDA].

Универсальная модель описания текстов предоставляет возможность идентифицировать и совмещать их разноязычные версии. Вместо привычного связывания веб-страниц LOD интегрирует ресурсы на уровне особым образом организованных и атрибутированных данных (метаданных). Таким образом, LOD является одной из реализаций авторитетного контроля. Например, через механизмы LOD налажен обмен данными между BUDA и базой данных проекта Sakya Research Centre (SRC) Австрийской академии наук, в которой собран большой объем авторитетных данных, связанных с тибетской культурой [Sakya 2011].

Обращение цифровых гуманитарных проектов и библиотек к технологиям, подобным PIF и LOD, обусловлено тем, что их широкое внедрение позволяет интегрировать отдельные распределенные базы данных друг с другом, создавая обширную сеть семантически взаимосвязанных ресурсов, формируя глобальное информационное пространство на основе ранее разрозненных цифровых хранилищ материалов, связанных с культурным наследием. Поэтому следование указанным подходам в деятельности по цифровизации письменного наследия позволит информационным ресурсам, создаваемым усилиями специалистов ИМБТ СО РАН, стать важной интегральной частью глобального цифрового пространства тибетского буддизма.

Технологии искусственного интеллекта как перспективное направление цифровизации тибетского письменного наследия

Несмотря на значительные усилия, предпринимаемые в основном в рамках инициатив ACIP и BDRC, объем полнотекстовых материалов, связанных с наследием тибетского буддизма, и переводов письменных памятников на европейские языки продолжает оставаться незначительными. Острыми остаются проблемы критического исследования, перевода и интерпретации первоисточников буддийской классики, например, разновременных изданий канонических собраний *Кангьюр* и *Тенгьюр*. Для

того чтобы представить наследие тибетского буддизма во всей его полноте, возникает настоятельная необходимость во внедрении новых методов в цифровизации буддийского письменного наследия, его изучении и освоении, разработке эффективных методов анализа и систематизации информации, вводе в научный оборот новых источников.

На наш взгляд, дальнейшее развитие в этой области связано с развитием методов оптического распознавания символов (OCR) тибетской письменности и машинного перевода. Однако существующие на сегодня решения, например, упомянутый выше Namsel OCR, способные достаточно результативно обрабатывать печатные тексты (книги современного формата), остаются малоприспособленными для рукописей и ксилографов, которые отличаются неоднородным качеством воспроизведения текста, плохой сохранностью бумаги и т. д.

Для успешного и эффективного решения задач представляется актуальным внедрение технологий искусственного интеллекта, таких как нейросетевые алгоритмы «глубокого обучения» для распознавания символов тибетской письменности и машинного перевода, развитие корпусных методов текстологического анализа на основе подходов big data и других современных цифровых инструментов.

Под термином «искусственный интеллект» понимается ряд методик, позволяющих компьютерам имитировать человеческий интеллект, к числу которых относится машинное обучение, основанное на применении алгоритмов накопления опыта для решения разнообразных задач.

Одним из наиболее актуальных ныне является метод «глубокого обучения», основанный на применении программируемых нейронных сетей, имеющих несколько последовательных слоев преобразований и самообучения. Благодаря такой структуре нейронной сети глубокое обучение позволяет решать задачи выявления закономерностей в неструктурированных данных, таких как изображения, звук, видео и текст. По этой причине этот метод широко применяется во многих областях, в том числе для декодирования сканированных текстов, машинного перевода, содержательного анализа текстов.

Применение метода глубокого обучения подразумевает последовательное выполнение ряда этапов. Сначала необходимо подготовить определенное количество обучающих данных, размеченных в специализированных форматах, так называемых датасетов. Затем осуществляется выбор основного алгоритма, на основе которого строится нейросеть. Так, для задач распознавания символов наиболее результативно применение сверточных нейронных сетей, а для машинного перевода и анализа текстов — модели преобразователей (трансформеров). Следующим этапом производятся обучение, т. е. загрузка обучающих данных в нейронную сеть, и тестирование обученной модели; этот процесс повторяется многократно с новыми наборами обучающих данными для достижения необходимой точности решения основной задачи. После этого модель готова для работы с произвольными данными.

Общей чертой методов глубокого обучения является то, что создание работоспособных моделей требует большого количества тщательно подготовленных обучающих данных (датасетов), значительных вычислительных ресурсов (с использованием графических или тензорных процессоров) и существенных временных затрат. Вместе с тем данная методика позволяет получать отличные результаты для решения многих задач, недостижимые при использовании иных способов.

В июне 2021 г. ЦВРК ИМБТ СО РАН посетили Президент Российской академии наук академик А. С. Сергеев и Председатель Сибирского отделения РАН академик В. Н. Пармон. Они предложили внедрить методы искусственного интеллекта для изучения памятников буддийского письменного наследия. С их помощью была составлена коллаборация с участием ИМБТ СО РАН, в которой индустриальным партнером выступило структурное подразделение корпорации Мобильные телесистемы (МТС) «Центр искусственного интеллекта МТС» (МТС-ИИ), технологическим партнером стал Новосибирский государственный университет.

В составе этой коллаборации в октябре–декабре 2021 г. был выполнен пилотный проект «Подготовка датасета для обучения baseline моделей оптического распознава-

ния символов тибетского языка», в рамках которого впервые применена технология искусственного интеллекта «глубокое обучение» для создания работающей модели оптического распознавания тибетской письменности. Со стороны НГУ координаторами проекта выступили кандидат химических наук, доцент А. Г. Окунев и К. Бродт.

В ходе выполнения пилотного проекта сотрудники ИМБТ СО РАН выполнили высококачественное сканирование ряда редких изданий на тибетском языке, соответствующее спецификациям программы Endangered Archives Program Британской библиотеки [[Guidelines](#)].

В качестве основного источника было взято ксилографическое Чонэское издание *Кангьюра*, относящееся к первой трети XVIII в. Оно отличается неоднородной четкостью печати и относится к ксилографам среднего качества. После этого специалистами-тибетологами была выполнена наиболее трудоемкая часть проекта — ручной ввод тибетского текста в объеме 500 страниц в латинской транслитерации.

На основе отобранных текстовых материалов были сформированы 500 обучающих датасетов, каждый из которых включает размеченное изображение книжной страницы на тибетской письменности и связанные с разметкой текстовые аннотации в латинской транслитерации. Подготовка датасетов осуществлялась в специализированной программе Label Studio (<https://labelstudio.io/>), развернутой на сервере НГУ. В ходе разметки каждая строка, расположенная на изображении страницы ксилографа, выделялась полигональной фигурой из набора инструментов рисования, представленных в Label Studio [[Get Started](#)].

Затем в соответствующем текстовом поле размещалась аннотация, состоящая из транслитерации текста, который содержит выделенная полигоном строка. Специалистами НГУ был разработан нейросетевой алгоритм, позволяющий выполнять программное детектирование отдельных графем тибетской письменности из выделенной полигоном строки, что существенно облегчило задачу разметки, выполняемую сотрудниками ЦВРК. Для удобства работы размечаемые страницы *Кангьюра* группировались в проекты, включающие по 60 страниц. Экспорт датасетов для ввода в

нейронную сеть осуществлялся в формате JSON. На основе 450 сформированных указанных выше способом датасетов специалисты НГУ произвели обучение модели оптического распознавания символов, созданной на основе нейросетевых технологий. Остальные 50 датасетов использовались для проверки работоспособности OCR-алгоритма.

Полученный в результате машинного обучения алгоритм достиг точности распознавания графем тибетской письменности 94 %, что с учетом специфики последней (слоговое письмо) дает примерно 80 % общей правильности декодирования текста [Brodth et al. 2022].

Дальнейшее обучение OCR-алгоритма с целью повышения точности распознавания до приемлемого уровня (97–99 %) требует кратного увеличения количества обучающих материалов (не менее 1 500–2 000 датасетов).

Таким образом, в ходе реализации пилотного проекта в ЦВРК ИМБТ СО РАН были освоены технологии и внедрен рабочий процесс подготовки обучающих датасетов на основе текстовых материалов на тибетском языке для нейросетевых моделей оптического распознавания символов. Была показана принципиальная возможность разработки нейросетевой OCR для книжных памятников на тибетской письменности.

В результате реализации пилотного проекта заложена база для создания полнофункциональной системы оптического распознавания символов тибетской письменности, на основе которой будет выполняться поточное декодирование сканированных письменных источников на тибетском языке. Дальнейшее развитие этого направления исследований связано с созданием цифрового текстового корпуса тибетоязычных буддийских сочинений и сопутствующих инструментов текстологического анализа, в частности, тезаурусов и конкордансов для разных тематических областей, которые обеспечивают быстрый доступ к нужным текстам. Важным является также разработка текстологической основы для разработки системы машинного перевода с тибетского языка на основе технологий искусственного интеллекта, которая включает параллельный корпус и словарные базы данных. Эти работы также инициированы в ЦВРК ИМБТ СО РАН.

Выводы

Необходимо подчеркнуть глобальный характер цифровой трансформации информационного пространства тибетского буддизма. Именно цифровые технологии обеспечивают целостное представление о содержании, истории и современном состоянии тибетской письменной культуры. Действующие в этой области консорциумы и инициативы, такие как Asian Classics Input Project, Tibetan and Himalayan Digital Library, Buddhist Digital Resource Center, Sakya Research Centre, обеспечивают фундаментальную методическую и технологическую поддержку для исследований в этой области.

Вместе с тем даже их активная многолетняя деятельность не позволяет охватить полной цифровизацией письменное наследие тибетского буддизма ввиду его колоссального объема и разнообразия форм, в том числе региональных. Поэтому остро стоит проблема применения в этой области таких технологий искусственного интеллекта, как метод «глубокого обучения», которые позволят существенно интенсифицировать и автоматизировать работы по цифровизации памятников тибетской книжности. Выполненное нами в сотрудничестве со специалистами НГУ пилотное исследование «Подготовка датасета для обучения baseline моделей оптического распознавания символов тибетского языка» показало перспективность применения технологий искусственного интеллекта. С учетом полученных результатов происходит формирование нового приоритетного направления исследований в Центре восточных рукописей и ксилографов ИМБТ СО РАН.

Цифровая трансформация буддийских письменных и литературных традиций открывает для них новые формы функционирования и новые возможности развития в современном мире, давая возможность в полной мере раскрыть культурный и интеллектуальный потенциал буддийской цивилизации для глобального общества. Для России она важна тем, что открывает еще одно направление культурной, ценностной интеграции со стремительно растущим Востоком.

Источники

ЦВРК — Центр восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии Сибирского отделения РАН.

Литература

Ринчинов 2020 — Ринчинов О. С. Тибетское наследие в цифровом пространстве: подходы и реализации // *Oriental Studies*. 2020. Т. 13. № 3. С. 640–651. DOI: 10.22162/2619-0990-2020-49-3-640–651

About us 1999 — About us. Buddhist Digital Resource Center. Connecting Buddhist literature with the world [электронный ресурс] // Buddhist Digital Resource Center. URL: <https://www.bdrc.io/about-us/> (дата обращения: 15.07.2022).

API — API Specifications – International Image Interoperability Framework™ [электронный ресурс] // International Image Interoperability Framework. URL: <https://iiif.io/api/> (дата обращения: 15.07.2022).

Access 2012 — BDRC's Access Policies [электронный ресурс] // Buddhist Digital Resource Center. URL: <https://www.bdrc.io/access-policies/> (дата обращения: 15.07.2022).

Brodt et al. 2022 — Brodt K., Rinchinov O., Bazarov A., Okunev A. Deep Learning for the Development of an OCR for Old Tibetan Books // *Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022)*. Abstracts the Thirteenth International Multi-conference. Novosibirsk: Institute of Cytology and Genetics, SB RAS, 2022. Pp. 1086.

BUDA — BUDA: Buddhist Digital Archives. A New Platform for Preservation and Research [электронный ресурс] // Buddhist Digital Resource Center. URL: <https://www.bdrc.io/buda-archive/> (дата обращения: 15.07.2022).

Cabezón, Jackson 1996 — Cabezón José Ignacio, Jackson Roger R. Editors' Introduction // *Tibetan literature: studies in genre* / ed. by J. I. Cabezón and R. R. Jackson. Ithaca, N. Y.: Snow Lion, 1996. Pp. 11–37.

References

About us. Buddhist Digital Resource Center. Connecting Buddhist literature with the world. On: Buddhist Digital Resource Center. Available at: <https://www.bdrc.io/about-us/> (accessed: 15 July 2022). (In Eng.)

API Specifications – International Image Interoperability Framework™. On: International Image Interoperability Framework. Available at:

Sources

Institute for Mongolian, Buddhist and Tibetan Studies (SB RAS), Center of Oriental Manuscripts and Xylographs.

Emanuel 2018 — Emanuel, Jeffrey P. *Stitching Together Technology for the Digital Humanities with the International Image Interoperability Framework (IIIF)* // *Digital Humanities, Libraries, and Partnerships: A Critical Examination of Labor, Networks, and Community* / ed. by R. Kear, K. Joranson. Cambridge, MA: Chandos Elsevier, 2018. Pp. 125–135.

Guidelines — Endangered Archives Programme: Guidelines for photographing and scanning archival material [электронный ресурс] // Endangered Archives Programme. British Library. URL: <https://eap.bl.uk/sites/default/files/Guidelines%20for%20photographing%20and%20scanning%20archival%20material%20%202017.pdf> (дата обращения: 15.07.2022).

Get started — Get started with Label Studio [электронный ресурс] // Label Studio. URL: <https://labelstud.io/guide/> (дата обращения: 15.07.2022)

Rowinski, Keutzer 2016 — Rowinski Z., Keutzer K. Namsel: An Optical Character Recognition System for Tibetan Text // *Himalayan Linguistics*. Vol. 15. No. 1. 2016. Pp. 12–30.

Sakya 2011 — Sakya Research Centre – About [электронный ресурс] // Sakya Research Centre. URL: <https://sakyaresearch.org/about> (дата обращения: 15.07.2022).

Salomon 1997 — Salomon R. A Preliminary Survey of Some Early Buddhist Manuscripts Recently Acquired by the British Library // *Journal of the American Oriental Society*. 1997. Vol. 117. Pp. 353–358.

IIIF — Why IIIF? IIIF provides benefits to collection-hosting institutions and end users alike [электронный ресурс] // International Image Interoperability Framework. URL: <https://iiif.io/get-started/why-iiif/> (дата обращения: 15.07.2022).

<https://iiif.io/api/> (accessed: 15 July 2022). (In Eng.)

BDRC's Access Policies. On: Buddhist Digital Resource Center. Available at: <https://www.bdrc.io/access-policies/> (accessed: 15 July 2022). (In Eng.)

Brodt K., Rinchinov O., Bazarov A., Okunev A. Deep learning for the development of an OCR for old Tibetan books. In: *Bioinformat-*

- ics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022). Conference abstracts. Novosibirsk: Institute of Cytology and Genetics (SB RAS), 2022. P. 1086. (In Eng.)
- BUDA: Buddhist Digital Archives. A New Platform for Preservation and Research. On: Buddhist Digital Resource Center. Available at: <https://www.bdrce.io/buda-archive/> (accessed: 15 July 2022). (In Eng.)
- Cabezón J. I., Jackson R. R. Editors' Introduction. In: Cabezón J. I., Jackson R. R. (eds.) *Tibetan Literature: Studies in Genre*. Ithaca, N. Y.: Snow Lion, 1996. Pp. 11–37. (In Eng.)
- Emanuel J. P. Stitching together technology for the digital humanities with the International Image Interoperability Framework (IIIF). In: Kear R., Joranson K. (eds.) *Digital Humanities, Libraries, and Partnerships: A Critical Examination of Labor, Networks, and Community*. Cambridge, MA: Chandos Elsevier, 2018. Pp. 125–135. (In Eng.)
- Endangered Archives Programme: Guidelines for photographing and scanning archival material. On: British Library. Endangered Archives Programme. Available at: <https://eap.bl.uk/sites/default/files/Guidelines%20for%20photographing%20and%20scanning%20archival%20material%20202017.pdf> (accessed: 15 July 2022). (In Eng.)
- Get started with Label Studio. On: Label Studio. Available at: <https://labelstud.io/guide/> (accessed: 15 July 2022). (In Eng.)
- Rinchinov O. S. Tibetan heritage in digital space: Approaches and implementation. *Oriental Studies*. 2020. Vol. 13. No. 3. Pp. 640–651. (In Russ.) DOI: 10.22162/2619-0990-2020-49-3-640–651
- Rowinski Z., Keutzer K. Namsel: An optical character recognition system for Tibetan text. *Himalayan Linguistics*. 2016. Vol. 15. No. 1. Pp. 12–30. (In Eng.)
- Sakya Research Centre – About. On: Sakya Research Centre. Available at: <https://sakyaresearch.org/about> (accessed: 15 July 2022). (In Eng.)
- Salomon R. A preliminary survey of some early Buddhist manuscripts recently acquired by the British Library. *Journal of the American Oriental Society*. 1997. Vol. 117. Pp. 353–358. (In Eng.)
- Why IIIF? IIIF provides benefits to collection-hosting institutions and end users alike. On: International Image Interoperability Framework. Available at: <https://iiif.io/get-started/why-iiif/> (accessed: 15 July 2022). (In Eng.)

