digitalorientalia
Цифровое востоковедение
DOI 10.31696/S278240120026084-5 VOL. 3 | No. 1-2 | 2023

Цифровые технологии | Digital Technologies

Автоматизированное распознавание рукописных текстов с помощью алгоритмов искусственного интеллекта: российский и зарубежный опыт

Automated handwriting recognition using artificial intelligence algorithms: Russian and foreign experience

Юмашева Юлия Юрьевна

д.и.н., заместитель генерального директора по научно-методической работе,
ООО "ДИМИ-ЦЕНТР", Россия, Москва

E-mail: Juliayu@yandex.ru

Julia Yumasheva

Deputy General Director for scientific and methodological work, DIMI-CENTER, Co., Ltd. Russian Federation, Moscow

Резюме. Обзор посвящен программным решениям и проектам автоматизированного распознавания рукописных текстов исторических источников, выполненным в Российской Федерации и за рубежом. Автор рассматривает наиболее известные OCR/HTR-сервисы, представленные онлайн, акцентирует внимание на интернет-ресурсах по европейской и восточной (японской, китайской) палеографии, описывает основные технологические принципы их реализации.

Ключевые слова: палеография, рукописный текст, средневековые исторические источники, оптическое распознавание, распознавание рукописных текстов, искусственный интеллект, наборы данных

Abstract. The review is devoted to software solutions and projects for automated recognition of handwritten texts of historical sources, made in the Russian Federation and abroad. The author considers the most famous OCR/HTR services presented online, focuses on Internet resources on European and Eastern (Japanese, Chinese) paleography, and describes the basic technological principles for their implementation.

Keywords: paleography, handwriting, medieval historical sources, optical recognition, handwriting text recognition, artificial intelligence, datasets

Возможности новых информационных технологий в вопросах автоматизированного распознавания рукописных текстов исторических источников в настоящее время являются одними из наиболее обсуждаемых тем в научных дискуссиях. Этой проблематике был посвящен и Круглый стол, проходивший в феврале 2023 г. в РАНХиГС, центральной темой которого было обсуждение опыта научных учреждений Российской Федерации по применению автоматизированного оптического распознавания текстов электронных копий архивных документов с помощью алгоритмов искусственного интеллекта [1, 2].

Участникам и слушателям Круглого стола были представлены два основных доклада по заявленной проблематике. Первый доклад о проекте «Digital Петр», осуществляемом специалистами Санкт-Петербургского института истории РАН и ПАО «СберБанк», хорошо известен. Проект неоднократно был представлен на различных конференциях [3, 4] и даже имеет собственный сайт в сети Интернет [5]. Его авторы применили методы искусственного интеллекта (Artificial Intelligence, AI) — комбинацию трех нейросетей для автоматизированного прочтения рукописей Петра I. В результате AI с уверенностью распознал рукописи 1709–1713 гг., уже прочитанные палеографами и даже изданные в «твердой» обложке [6].

Во втором докладе специалисты РАНХиГС представили проект, который находится на стадии разработки. Исследование посвящено изучению экономики России первой половины XIX в. на основе изучения отчетов губернаторов – источника хорошо известного в отечественной историографии [7, 8]. Использование AI в данном проекте носит утилитарный характер и является средством «извлечения данных» (data wrangling) из рукописного текста с целью их последующего анализа.

Расширяя границы Круглого стола, в контексте обзора отечественного опыта применения AI для распознавания рукописных текстов, необходимо также упомянуть проект Центра восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии СО РАН (Улан-Удэ) (http://imbtarchive.ru/index.php). Специалисты Центра смогли «прочитать» (дешифровать) с помощью AI 500 страниц тибетских рукописей с точностью в 94% распознаваемых символов, однако с учетом всех особенностей тибетской письменности правильность текстов в данный момент оценивается примерно в 80% [9, 10].

В целом же, следует отметить, что тема автоматизированного распознавания рукописного текста не нова и разрабатывается специалистами разных стран более 30 лет с момента проведения первой международной конференции International Conference on Document Analysis and Recognition (ICDAR) в 1991 г. [11]. На сегодняшний день ICDAR — ведущее международное событие для ученых и практиков, занимающихся автоматическим распознаванием и анализом текстов документов. За десятилетия проведения этого мероприятия его участники — ученые разных стран — представили более тысячи докладов, посвященных различным аспектам осуществления проектов автоматизированного распознавания исторических текстов, содержащихся на разных носителях: от каменных блоков и глиняных табличек до машинописных документов и современных газет.

Разработка систем автоматического распознавания (Optical Character Recognition, OCR) получила дополнительный импульс в начале 2000-х гг. в связи с массированным проникновением в различные научные дисциплины методов Data Science (DS), в том числе машинного (глубокого) обучения (Machine Learning, ML; Deep Learning, DL) и искусственного интеллекта. Применение подходов, методов и алгоритмов DS, а также совершенствование аппаратно-программных решений в области сканирования исторических артефактов, позволили создать технологию распознавания рукописного текста

(Handwritten Text Recognition, HTR) и инструменты, нацеленные на решение задач автоматизированного распознавания текстов рукописных документов, «извлечение» данных и их индексацию, и даже реконструкцию утраченных фрагментов текстов.

К числу наиболее известных программных продуктов Handwritten Text Recognition, предназначенных для работы историков и филологов, относятся:

Transcribus (https://readcoop.eu/transkribus/?sc=Transkribus) – программное обеспечение, которое объединяет модели распознавания изображений и текста для облегчения распознавания рукописных символов. Программа доступна через графический интерфейс или через API и разработана, чтобы «аккуратно вписаться в архивный рабочий процесс, напрямую используя растущие репозитории оцифрованных изображений исторических текстов». Transkribus весьма популярен и часто используется европейскими архивами и другими учреждениями, играя решающую роль в расширении использования AI для извлечения содержания рукописных документов. Этот инструмент обладает дополнительными опциями, в частности, он позволяет использовать (интегрировать) в программу распознавания архивные «исторические индексы» (персональные, географические, предметные и т.п. указатели) [12], что в свою очередь, дает возможность создавать более гибкие механизмы поиска в архивных информационных системах, а также значительно ускорить процессы индексации [13] и создания связанных данных, в том числе с ресурсами музеев и библиотек [14].

Transkribus нашел применение для распознавания книг на бенгальском языке, которые отсканированы в рамках проекта Британской библиотеки «Два столетия индийской печати», осуществлявшегося в 2016–2019 гг. (https://www.bl.uk/projects/two-centuries-of-indian-print). Целью проекта была каталогизация и оцифровка более 1600 печатных книг на разных языках (бенгальский, ассамский, силхети и урду) из коллекций Южной Азии, датируемых 1713–1914 гг. (https://www.bl.uk/early-indian-printed-books), создание интерактивной карты книгопечатания, разработка методов автоматического распознавания и формирование наборов данных для обучения систем автоматического распознавания (https://bl.iro.bl.uk/collections/d4b2009d-b28d-4518-b219-fc0cd53007e7?locale=en).

eScriptorium (https://escriptorium.openiti.org/) – инструмент для распознавания и транскрибирования текста из печатных и рукописных документов, созданный с использованием методов машинного обучения. Одной из отличительных черт этого программного обеспечения является опция по сегментации и созданию метаданных (описанию) фрагментов изображений, основанная на применении формата изображений Международной платформы взаимодействия изображений (International Image Interoperability Framework, IIIF – https://iiif.io/get-started/why-iiif/), что дает возможность исследователям работать не только с текстом, но и с иллюминированными элементами рукописей.

eScriptorium тесно связан с еще одной системой распознавания текста **Kraken** (https://kraken.re/main/index.html), которая оптимизирована для исторических источников и рукописей, написанных нелатинским шрифтом. Kraken является программным обеспечением с открытым исходным кодом и активно используется в качестве основы для создания и развития различных систем HTR (https://github.com/mittagessen/kraken).

На основе Kraken и eScriptorium индийским филологом Роханом Чаунаном в инициативном порядке разработана и активно развивается система автоматизированного распознавания рукописного текста, написанного на языках урду, хинди и бенгали [15, 16].

Специалистам давно известно, что успешность применения автоматизированного распознавания рукописных текстов во многом зависит от наличия наборов палеографических данных и словарей – чем больше и репрезентативнее наборы, чем больше примеров почерка (начертания букв, иероглифов) они в себя включают, чем больше словари топонимов, вариантов написания имен, терминов и т. п., тем лучше будет работать распознавание.

Именно поэтому в основе большинства HTR-сервисов лежат наборы палеографических данных, созданные и опубликованные в сети Интернет в виде самостоятельных проектов. Среди них особой известностью пользуются проекты:

DigiPal (http://www.digipal.eu) — сайт, предназначенный для изучения средневековых европейских почерков XI–XII вв.;

Italian Paleography (https://italian.newberry.t-pen.org/) — онлайн-учебник итальянской палеографии для рукописей, написанных между 1300 и 1700 гг., с инструментом T-Pen для расшифровки и транскрибирования текста;

Digital Analysis of Syriac Handwriting (DASH, http://dash.stanford.edu/) — проект цифровой палеографии, который представляет электронные копии листов из 90% сохранившихся сирийских рукописей, написанных до XII в. включительно, и позволяет проводить палеографический анализ с последующим обучением HTR-программ;

MultiPal (https://www.multipal.fr/en/welcome/) — интерактивный онлайн-учебник по палеографии, который помогает научиться расшифровывать оригинальные рукописи, документы и надписи на различных древних и средневековых языках, шрифтах и почерках, в том числе: на латыни, греческом, египетском, коптском, арабском, иврите, арамейском, сирийском, китайском, санскрите и кириллице и др.

Формирование палеографических баз (наборов) данных и словарей — задача чрезвычайно сложная и трудоемкая, к решению которой целесообразно привлекать большое количество исследователей или волонтеров, работающих с письменными источниками. Этот подход активно развивается в проектах создания палеографических ресурсов и систем распознавания рукописных текстов, которые развиваются в странах Дальнего Востока и/или научных центрах изучения исторических источников по истории Японии и Китая.

Сравнительно недавно чтение средневековой японской письменности (как и в большинстве стран Западной Европы и США) осуществлялись с помощью краудсорсинговых платформ и активного участия волонтеров. В рамках этого направления наиболее известен проект **«Расшифровываем все вместе»** (Minna de Honkoku, Minna de Reprint, みんなで翻刻, https://honkoku.org), начатый в 2017 г. палеоэйсмической исследовательской группой Киотского университета. Целью проекта была расшифровка и перевод в машиночитаемый вид японских исторических материалов о землетрясениях. К 2021 г. с помощью 5 тыс. волонтеров удалось расшифровать и перевести в печатный вид более 600 млн знаков японских исторических документов, что, с одной стороны, позволило проводить исследования по истории природных катаклизмов, а с другой — сформировать необходимые для развития технологии автоматизированного распознавания наборы данных.

Одним из первых автоматизированных программных приложений для чтения японской исторической слоговой скорописи с древнейших времен до революции Мэйдзи, является проект **Hentaigana** (https://alcvps.cdh.ucla.edu/support/), разработанный Калифорнийским университетом, Инициативой Yanai и Университетом Васэда, в рамках которого создана база данных исторических иероглифов-слогов хентайгана («разновидностей кана»).

Фактически Hentaigana — японский аналог европейских палеографических баз данных, в которой каждый скорописный иероглиф сопровождается вариантами написания, указанием на «родительский» иероглиф (jibo) и полным («раскрытым») написанием хейтагана, позволяющими проследить его эволюцию.

Heitagana — отправная точка для созданных в последние несколько лет различных информационных OCR/HTR систем и проектов, нацеленных на чтение скорописи и японского курсивного письма «кузусидзи» (kuzushiji), в частности: «Поиск по рукописному тексту с помощью AI» (AI Tegaki kuzushiji kensaku, 手書きくずし字検索, http://www.ai-kuzushiji.net/); «Скорописные формы ключевых знаков» (Bushu no kuzushiji, 部首のくずし字, http://komonjo.rokumeibunko.com/binran/bushu01.html) — сайт для расшифровки рукописных средневековых документов; «Поиск по базе данных Kuzushi-ji» (Kuzushiji dētabēsu kensaku (くずし字データベース検索, http://codh.rois.ac.jp/char-shape/search/) – поисковая система для поиска иероглифов в базе данных kuzushiji и др.

И, наконец, еще одно программное приложение — **База данных Кузусидзи** (KuLA, くずし字学習支援アプリ, https://apps.apple.com/us/app/くずし字学習支援アプリkula/id1076911000), позволяющее распознавать японские надписи в стиле кузусидзи, разработано специально для мобильных приложений и имеет функцию пополняемого новыми иероглифами словаря.

Особый интерес вызывает проект распознавания древних японских рукописей «Поиск соответствия изображений по мокканам или курсивным символам» — **MOJIZO** (解析：木簡・くずし字解読システム, https://aimojizo.nabunken.go.jp; описание: https://mojizo.nabunken.go.jp/doc/legend_en.pdf; сокращенный вариант системы – 検索画面：奈良文化財研究所 史的文字データベース連携検索システム – https://mojiportal.nabunken.go.jp/) который осуществляется в Японии, Корее и Китае уже более 10 лет. Цель проекта – оцифровка, разработка метода распознавания и распознавание древней иероглифической письменности, которая является в некотором смысле промежуточным вариантом между алфавитным письмом и изображением. Речь идет о переводе в цифровой формат текстов «моккан» (木簡), которые написаны китайскими иероглифами (яп. «кандзи», кор. «ханча») на деревянных дощечках. Это очень распространенные источники на Дальнем Востоке середины I тыс. н.э. На дощечки длиной 10–25 см, шириной 2–3 см и толщиной в несколько миллиметров наносили с помощью чернил (туши) тексты самого разного содержания и назначения. Дощечки использовали настолько широко, что их архивы сохранились в странах Дальнего Востока практически повсеместно. Однако изучение табличек осложнено степенью их сохранности и спецификой древнего языка, который имел собственные диалекты в разных странах.

Систематизация и описание дощечек моккан начались в Японии в середине 1960-х гг., в 1999 г. была презентована первая база данных (https://mokkanko.nabunken.go.jp/en/?c=about), но лишь в 2016 г. японскими специалистами был предложен аналог нейросети, реализованный на основе наборов данных MNIST, разработанных Яном Лекуном (Yann LeCun — http://yann.lecun.com), лауреатом премии Алана Тьюринга и автором многих продуктивных подходов в области глубокого машинного обучения с целью автоматизированного распознавания текстов.

В качестве наборов данных для обучения выступают две постоянно пополняемые базы данных: база данных Национального исследовательского института культурных ценностей Нара, которая содержит изображения моккан, найденные в Японии (примерно 1 800 знаков, 14 000 деревянных табличек и 89 600 символьных изображений),

и база данных Историографического института Токийского университета, в которую включены различные изображения досовременной японской письменности (примерно 6 000 письменных знаков и изображения из 230 000 знаков –木簡庫 奈良文化財研究所 – https://mokkanko.nabunken.go.jp/ja/). Кроме собственно изображений знаков система включает в себя несколько служебных баз данных, позволяющих максимально упростить процесс распознавания. Среди них базы данных личных имен, имен исторических деятелей, топонимов и т.п.

Механизм реализации функции распознавания прост – система предлагает возможности поиска текстовых изображений в БД: пользователь вводит текстовое изображение иероглифа в поисковую систему, и если в ней нет аналога, выбранного из 8 наиболее близких текстовых изображений, уже введенных в систему, то в системе создается новое изображение иероглифа-текста и делается попытка его перевода-интерпретации.

На основе этих разработанных и апробированных методов в конце 2019 г. в Японии был открыт проект **KuroNet Kuzushiji Ninshiki Sābisu** (KuroNet くずし字認識サービス, http://codh.rois.ac.jp/kuronet/; https://mp.ex.nii.ac.jp/kuronet/) — бесплатный онлайн-сервис OCR-распознавания, который позволяет пользователям конвертировать изображения документов, написанных на иероглифическом японском языке, в печатный текст [17]. В основу этого многосимвольного сервиса была положена модель распознавания RURI (瑠璃), соответствующая Международной платформе обмена изображениями IIIF и использующая методы глубокого обучения [18, 19]. В качестве дополнения и расширения возможностей ресурса созданы и представлены на сайте специальные тематические наборы данных, облегчающие процесс распознавания.

Рассказ о палеографических базах данных и системах распознавания средневековых рукописных текстов, созданных в Японии, был бы не полон без упоминания двух онлайн-учебников. Первый — **«Древние документы»** (古文書なび-古文書解読支援サイト, http://komonjo.rokumeibunko.com/index.html), предназначен для обучения японской палеографии и умению читать средневековые документы. Второй — **«Сеть древних документов»** (古文書ネット くずし字史料から歴史を紐解こう, https://komonjyo.net/index.html) включает в себя обширные руководства по чтению исторических японских текстов, в том числе текстов на гравюрах Укиё-э, а также методы изучения и примеры различных типов исторических источников.

Используя особенности графики японской письменности и опыт создания наборов палеографических данных, японские специалисты формируют наборы данных для распознавания лиц, мимики и жестов на гравюрах стиля Укиё-э (浮世絵). Ксилографии этого стиля, возникшего в конце XVI в. и просуществовавшего до начала XX в., были чрезвычайно популярны и распространены, в результате в разных музеях, библиотеках и архивах сохранились большие и разнообразные коллекции гравюр. На многих из них запечатлены исторические личности, да и просто типажи, известные из произведений классической японской литературы.

Для автоматизированного определения персон сформированы два набора данных. Первый — ARC Ukiyo-e Faces Dataset (http://codh.rois.ac.jp/ukiyo-e/face-dataset) — создан путем первоначально ручной разметки, последующего машинного обучения и автоматизированного «извлечения» областей с изображением лица с электронных копий гравюр. По состоянию на июнь 2021 г. набор содержит 16653 данных о лицах, изображенных на 9203 гравюрах Укиё-э (https://github.com/rois-codh/arc-ukiyoe-faces/). При этом набор создан таким образом, что позволяет даже проследить эволюцию изображений персонажа.

Второй — посвящен чертам лица (顔貌コレクション（顔コレ), Facial features — http://codh.rois.ac.jp/face/), которые также «извлечены» из изображений и позволяют проследить изменение мимики персонажей, изображенных на гравюрах или книжных иллюстрациях [20].

Работы по созданию наборов данных и программ автоматизированного распознавания рукописного текста активно ведутся и в Китае. К примеру, начатый в инициативном порядке проект **«Историческая лингвистика»** (Guyin xiaojing, 古音小鏡·歷史語言學, http://www.kaom.net/), предназначенный для представления и обмена материалами в области древнекитайской письменности (в частности, иероглифики Южных царств – письменности Чу) и совершенствования инструментов исторической лингвистики, стал развиваться благодаря интересу и посильному вкладу широкого круга исследователей-синологов. Ресурс состоит из шести блоков, в которых представлены создаваемые в режиме краудсорсинга наборы данных по темам: фонология разных периодов китайского языка 上古音; сравнительные реконструкции 構擬; памятники древний письменности 古文字; раздел карт географического распределения произношения одного и того же слова 漢語地理; раздел топонимов 地名; и раздел библиографических ссылок на исследования по тематике ресурса [21].

Развиваются и иные проекты в области древнекитайской палеографии и лингвистики, создание которых является основой для совершенствования систем OCR/HTR. Среди наиболее крупных проектов следует упомянуть: проект **«Рукописи на бамбуке и шелке»** Уханьского университета (中國古代簡帛字形辭例數據庫, http://www.bsm.org.cn/zxcl/); Многофункциональную Базу данных китайских иероглифов с архаичными формами (漢語多功能字庫, Multi-function Chinese Character Database, https://humanum.arts.cuhk.edu.hk//Lexis/lexi-mf/); проект «Интеграции древнего и современного текста» (簠 | 楚簡文字 | 開放古文字字形庫 | 古今文字集成, http://ccamc.org/cjkv_oaccgd.php?cjkv=簠&type=chujian); Базу данных древних китайских текстов (CHANT Database — Chinese Ancient Texts, https://www.cuhk.edu.hk/ics/rccat/en/database.html), а также совместный проект китайских и японских специалистов по созданию единой базы данных, позволяющей осуществлять сквозной поиск китайских исторических иероглифов по всем известным ресурсам (奈良文化財研究所 史的文字データベース連携検索システム, https://mojiportal.nabunken.go.jp/en/).

Не остались в стороне от реализации проектов формирования палеографических систем и словарей и разработчики программного обеспечения. Весной 2023 г. в Китае была анонсирована новая функция программного обеспечения WeChat (https://www.wechat.com/), предназначенная для сбора редко используемых китайских иероглифов. В программу можно загрузить предварительно сфотографированные символы, которые отсутствуют в компьютерных наборах китайского языка, внести их (после одобрения филологов) в библиотеку, чтобы затем редко используемые иероглифы стали доступными для цифрового отображения и распознавания. После запуска 20 апреля в первый же день своей работы мини-программа получила более 630 000 посещений при этом пользователи представили 1404 редких иероглифа. После профессиональной проверки одобренные редкие иероглифы будут закодированы и внесены в национальную библиотеку стандартных символов, что в конечном итоге обеспечит их безбарьерный ввод, отображение и распознавание на устройствах и информационных системах, таких как компьютеры и мобильные телефоны. К слову, сейчас в китайском языке, доступном для набора с компьютера, 90 тыс. иероглифов, а согласно китайской же статистике, в Китае более 60 млн имен людей и мест, древних книг и диалектов содержат редкие иероглифы, большинство из которых не были оцифрованы, и даже

имена при компьютерном наборе пишутся искаженно [22]. Таким образом формируются словари/наборы данных, которые со временем позволят создавать и применять системы OCR/HTR.

Подводя итог краткому обзору программного инструментария, который создается и используется для автоматизированного распознавания рукописных текстов, необходимо подчеркнуть, что успешность применения подобных программных средств и методов Data Science определяется наличием и качеством проведенных подготовительных работ, полнотой и уровнем детализации сформированных наборов данных и тематических словарей, что ярко подтверждается не только практическими примерами реализации проектов распознавания текстов, написанных с помощью алфавитных систем кодирования информации, но и с помощью иероглифической и идеографической письменности.

Библиография / References

1. Программа Круглого стола «Искусственный интеллект в исторических исследованиях: автоматизированное распознавание текстов рукописных исторических источников», 11 февраля 2023 г. РАНХиГС. https://aik-hisc.ru/static/pdfs/aik_docs/семинар_ИИ_2023.pdf (дата обращения: 16.06.2023)
2. Видеозапись выступлений на Круглом столе 11 февраля 2023, РАНХиГС. https://www.youtube.com/watch?v=iP7kpaDBPP4 (дата обращения: 16.06.2023)
3. Базарова Т.А., Проскурякова М.Е. Автографы Петра I: чтение технологиями искусственного интеллекта и создание электронного архива // Историческая информатика. 2022. № 4 (42). С. 179–190.
4. Автографы Петра Великого и технологии искусственного интеллекта. Новости РИО. https://historyrussia.org/sobytiya/avtografy-petra-velikogo-i-tekhnologii-iskusstvennogo-intellekta.html (дата обращения: 16.06.2023)
5. Автографы Петра I. Электронный архив. https://peterscript.historyrussia.org/ (дата обращения: 16.06.2023)
6. Письма и бумаги императора Петра Великого. Том XIV. Выпуск I. Январь – июнь 1714 г. Издательство «Древлехранище», Москва, 2022. 928 с.
7. Литвак Б.Г. О достоверности сведений губернаторских отчетов XIX в. // Источниковедение отечественной истории. М., 1976. С. 125–144.
8. Минаков А.С. Годовые всеподданнейшие отчеты губернаторов: исследовательский опыт и источниковедческие перспективы /Археографический ежегодник за 2009-2010 годы. М., Наука. 2013. С. 37–55.
9. Штерман И. Сибирские ученые начали расшифровку старинных книг при помощи нейросети // Российская газета. Иркутск. 05.04.2022. https://rg.ru/2022/04/05/reg-dfo/sibirskie-uchenye-nachali-rasshifrovku-starinnyh-knig-pri-pomoshchi-nejroseti.html (дата обращения: 16.06.2023)
10. Базаров Б.В., Ринчинов О.С., Базаров А.А. Цифровая трансформация письменного наследия тибетского буддизма: состояние и перспективы // Oriental Studies. 2022;15(4):740-750. https://doi.org/10.22162/2619-0990-2022-62-4-740-750 (дата обращения: 16.06.2023)
11. International Conference on Document Analysis and Recognition (ICDAR) // https://icdar2021.org/; https://www.icdar.org/document-analysis/ (дата обращения: 16.06.2023)
12. Ranade S. Traces through Time: A Probabilistic Approach to Connected Archival Data // 2016 IEEE International Conference on Big Data (Big Data), 3260–65. Washington DC, USA: IEEE. https://doi.org/10.1109/BigData.2016.7840983 (дата обращения: 16.06.2023)
13. Colavizza, G., Ehrmann, M., Bortoluzzi, F. Index-Driven Digitization and Indexation of Historical Archives // Frontiers in Digital Humanities. 2019. №6 (March). https://doi.org/10.3389/fdigh.2019.00004 (дата обращения: 16.06.2023)

14. Wilde M. de, Hengchen S. Semantic Enrichment of a Multilingual Archive with Linked Open Data // Digital Humanities Quarterly. 2017. № 11(4).
15. Chauhan R. eScriptorium: Digital Text Production for Urdu, Hindi, and Bengali Print, part 1 // The Digital Orientalist. https://digitalorientalist.com/2022/11/15/escriptorium-digital-text-production-for-urdu-hindi-and-bengali-print-part-1/ (дата обращения: 16.06.2023)
16. Chauhan R. eScriptorium: Digital Text Production for Urdu, Hindi, and Bengali Print, part 2 // The Digital Orientalist. https://digitalorientalist.com/2023/01/31/escriptorium-digital-text-production-for-urdu-hindi-and-bengali-print-part-2/ (дата обращения: 16.06.2023)
17. Cursive Japanese and OCR: Using KuroNet // The Digital Orientalist. https://digitalorientalist.com/2020/02/18/cursive-japanese-and-ocr-using-kuronet/ (дата обращения: 16.06.2023)
18. Kitamoto Asanobu, Tarin Karanuwat. Kuzushi Character Recognition by AI and the Road to Full-text Search for Historical Materials // Specialized Library, No. 300, pp. 26–32, 2020/5 (北本 朝展, カラーヌワット タリン, 「AIによるくずし字認識と歴史的資料全文検索への道」, 専門図書館, No. 300, pp. 26-32, 2020年5月)
19. Tallinn Karanuwat, Kitamoto Asanobu. Evolution of Kuzushi Character Recognition and Development of Service // Humanities and Computer Symposium Jinmonkon 2020 Proceedings, pp. 3–10, December 2020 (カラーヌワット タリン, 北本 朝展, 「くずし字認識の進化とサービス化の展開」, 人文科学とコンピュータシンポジウム じんもんこん 2020 論文集, pp. 3–10, 2020年12月)
20. Yingtao Tian, Tarin Clanuwat, Chikahiko Suzuki, Asanobu Kitamoto. Ukiyo-e Analysis and Creativity with Attribute and Geometry Annotation // Arxiv.org. https://arxiv.org/pdf/2106.02267.pdf (дата обращения: 16.06.2023)
21. Poli M. The evolution of Kaom.net // The Digital Orientalist. https://digitalorientalist.com/2023/05/16/the-evolution-of-kaom-net/ (дата обращения: 16.06.2023)
22. Liu Yanling. Rarely used Chinese characters to be collected and made available online // Global Times. Apr 24, 2023. https://www.globaltimes.cn/page/202304/1289735.shtml (дата обращения: 16.06.2023)