УДК 004.8+930.25

Киселёв Игорь Николаевич
Всероссийский научно-исследовательский институт документоведения и архивного дела
г. Москва, Российская Федерация
kiselev_in@mail.ru

Научная статья

О ПРИМЕНЕНИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В РАСПОЗНАВАНИИ ТЕКСТОВ

Представлены описания основных этапов распознавания архивных документов с использованием искусственного интеллекта. Приведены примеры проектов по распознаванию и применяемых технологических инструментов.

Ключевые слова: архивные документы, искусственный интеллект, распознавание текстов.

Для цитирования: Киселев И.Н. О применении искусственного интеллекта в распознавании текстов // Вестник ВНИИДАД. 2024. № 1. С. 84–95.

Original article

ON THE AI APPLICATION IN TEXT RECOGNITION

The main stages of archival documents recognition using artificial intelligence are presented. Examples of recognition projects and used technological tools are provided.

Key words: archival documents, artificial intelligence, text recognition.

For citation: Kiselev I.N. On the AI application in text recognition. *Vestnik VNIIDAD = Herald of VNIIDAD*. 2024;1:84–95. (In Russian).

Среди приложений искусственного интеллекта (ИИ) в архивном деле распознавание рукописных и старопечатных документов – направление, приносящее в настоящее время наиболее зримые и практически значимые результаты.

Главное преимущество распознанных документов, загруженных в Интернет, – возможность их индексирования поисковыми машинами Интернета с последующим поиском по всему содержанию текста.

Поступила в редакцию: 09.01.2024
Поступила после рецензирования: 15.01.2024
Принята к публикации: 17.01.2024

Received: 09.01.2024
Revised: 15.01.2024
Accepted: 17.01.2024

Конечными целями распознавания являются предоставление доступа к распознанным копиям архивных документов, публикация документов, облегчение поиска документов, оптимизация исполнения запросов, формирование основы для научных исследований историков, палеографов, филологов.

В распознавании текстов можно выделить два основных направления: полное дословное распознавание исходного текста и целевое распознавание в тексте именованных сущностей – имен, дат, географических названий, типов действий, связей между ними и т.п. В последнем случае распознанная информация обычно служит основой для создания исследовательских баз данных, генеалогических и картографических построений, иных исследовательских проектов.

Несмотря на многолетние усилия сотен специалистов и осуществление множества научных и практических проектов, проблема распознавания архивных документов в целом не решена.

В 2019 году генеральный секретарь Международного совета архивов (МСА) Антея Селес в докладе «Дивный новый мир: Искусственный интеллект и архивы» на 14-й Общей конференции и семинаре в Токио (Япония) утверждала: «Автоматизация анализа почерка все еще находится на ранних стадиях исследований» [1]. С тех пор ситуация кардинально не изменилась.

Как правило, распознавание архивных документов осуществляется в рамках отдельных проектов. В зависимости от цели проекта определяется набор компьютеризированных инструментов его реализации. При этом только в части этих инструментов используется ИИ (к ним чаще всего относят искусственные нейронные сети). Непременным элементом распознавания текстов является участие специалистов-гуманитариев, которые выполняют ручную работу – прежде всего на этапе формирования обучающего и тестового массивов.

Исходным материалом для распознавания всегда служит отсканированный документ, результатом распознавания (полного или фрагментарного) является текст в форматах, позволяющих его редактировать и индексировать.

Процесс распознавания файла включает ряд основных этапов:

- выбор материала для распознавания;
- подготовка исходного материала;
- подготовка обучающего и тестового массивов;
- выделение текстовых фрагментов на изображении (анализ макета документа);
- сегментация текста;
- обучение нейронной сети (формирование языковой модели);
- распознавание текста.

Методам и программным инструментам реализации каждого этапа посвящено множество исследований, здесь мы дадим только их краткую характеристику.

Исходный материал для распознавания должен быть объемным и может содержать многие сотни страниц текста или большой массив однородных (однотипных) документов. Ясно, что для распознавания небольших текстов или нескольких документов не стоит затевать длительный и дорогостоящий проект.

При подготовке исходного материала производится очистка изображения от просвечивания обратной стороны листа, клякс и т.п., которая осуществляется графическими редакторами. На этом этапе производится также устранение перекосов в изображении, выпрямление наклона текста.

После очистки изображения проводится его бинаризация – преобразование цветного (или в градациях серого) изображения в черно-белое. Традиционно алгоритмы бинаризации являются статистическими. Однако в ходе экспериментов была показана эффективность использования ИИ для бинаризации [2, 3].

При распознавании текстов документов со сложной структурой (таблиц, анкет и т.п.) необходимо производить анализ макета документа с целью определения областей, содержащих непосредственно текст.

Выделение текстовых фрагментов проделывалось обычно вручную в графическом редакторе, но может быть использован ИИ в виде так называемых «полностью сверточных нейронных сетей» [4]. Этот метод позволяет одновременно выполнять автоматизированную сегментацию выделенных текстов на строки и определять для строк базовые линии. При дословной транскрипции текста выделяются отдельные буквы и символы. Автоматическая сегментация может производиться с помощью ИИ [5, 6] или статистическими алгоритмами без обучения.

Собственно распознавание производится чаще всего нейронной сетью. Пока не разработана универсальная нейронная сеть, которая могла бы распознавать тексты на любых языках, написанных любым почерком в любой исторический период.

Для каждого конкретного массива документов необходимо адаптировать нейронную сеть, т.е. создать языковую модель. Этот процесс называется обучением модели. Первым шагом является подготовка обучающего материала, который состоит из отсканированного образца из выбранного документного массива и его точной расшифровки, представленной в современной машиночитаемой кодировке. Совокупность образца и его расшифровки обозначается в литературе и документации термином ground truth (в русскоязычных публикациях термин переводится как «основная истина», «наземная правда» и др.).

Подготовка «основной истины» производится вручную и является наиболее трудоемким и длительным этапом расшифровки текста. В этой работе, помимо архивистов, зачастую участвуют палеографы, филологи, историки, а также иногда используется краудсорсинг.

Обычно рекомендуется выбрать для рукописного текста в качестве обучающего материала массив объемом от 5 000 до 15 000 слов (около 25–75 страниц) расшифрованного текста [7].

Критически важной является точность «основной истины», поэтому в дополнение к ручной проверке точности разрабатываются методы автоматизированной проверки [8].

После бинаризации, анализа макета документа и сегментации производится обучение модели. Изображение текста загружается в «движок» (нейронную сеть), а на выходе получается расшифровка текста. В начале обучения полученная расшифровка обычно содержит значительное количество ошибок. Точность модели измеряется отношением (в процентах) количества ошибок в полученной расшифровке к общему количеству символов в точной расшифровке, т.е. в «основной истине». Этот показатель называется «коэффициентом символьных ошибок» (Character Error Rate, CER) и является главным критерием качества модели и ее пригодности для использования.

Минимально удовлетворительной считается модель со значением CER от 2 до 8% для рукописного текста и от 0,5 до 2% для печатного.

После корректировки полученной первоначально расшифровки процесс распознавания запускается снова.

Компания ReadCoop на своем сайте разместила в свободном доступе 134 бесплатные модели[1], уже обученные на платформе Transkribus. Цель размещения моделей – предоставить пользователям возможность не начинать обучение с нуля, а использовать эти модели в качестве исходных (базовых). Если модель дает на материале пользователя низкий CER, то можно проводить распознавание всего текста. Если же CER оказывается высоким, то следует производить дообучение модели. Среди моделей представлены пять для текстов на русском языке – «Русский почерк начала XX века»[2], «Российские записи актов гражданского состояния конца XIX века»[3] и др.

Критически необходимым модулем каждой системы является «движок» – программа, которая выполняет непосредственно распознавание текста, содержащегося в документах. Наибольшее применение получили движки PyLaia[4], Kraken[5] [9], Calamari [10], Tesseract[6]. Исходные коды перечисленных движков и документация к ним размещены на сайте GitHub в открытом доступе.

В настоящее время функционируют несколько платформ, предлагающих комплексные решения задач распознавания текстов исторических документов. Платформы содержат полные наборы инструментов, необходимых для распознавания, документацию по практической работе пользователей и другие вспомогательные материалы.

Все платформы работают по принципу «клиент – сервер». Клиентское программное обеспечение скачивается на персональный компьютер пользователя, где он задает все нужные параметры и управляет рабочими процессами на сервере. Теоретически платформу со всеми модулями можно установить на настольный современный компьютер и выполнять на нем всю работу, однако распознавание и другие процедуры нуждаются в очень больших ресурсах, доступных

[1] Public AI models in Transkribus. URL: https://readcoop.eu/transkribus/public-models/ (дата обращения: 24.11.2023).

[2] Russian Handwriting early 20th century. URL: https://readcoop.eu/model/russian-handwriting-early-20th-century/ (дата обращения: 24.11.2023).

[3] Russian Civil Records late XIX cent. URL: https://readcoop.eu/model/russian-civil-records-late-xix-cent/ (дата обращения: 24.11.2023).

[4] PyLaia. Pattern Recognition and Human Language Technology (PRHLT) Research Center. URL: https://github.com/jpuigcerver/pylaia (дата обращения: 24.11.2023).

[5] Kraken. Mittagessen. URL: https://kraken.re/main/index.html (дата обращения: 30.09.2023); Kraken. RESILIENCE project. URL: https://github.com/mittagessen/kraken (дата обращения: 30.09.2023); Training kraken. Mittagessen. URL: https://kraken.re/main/training.html#evaluation-and-validation (дата обращения: 24.11.2023).

[6] Tesseract-ocr. URL: https://github.com/tesseract-ocr/ (дата обращения: 24.11.2023); Tesseract User Manual. URL: https://tesseract-ocr.github.io/tessdoc/#external-projects (дата обращения: 24.11.2023).

только на мощном сервере. Выполнение работ, которые занимают на сервере несколько часов, потребовало бы на настольном персональном компьютере нескольких недель.

Платформы выполняют следующие основные функции: импорт изображений в систему; анализ макета; транскрипция строк текста, (преобразование изображения в машиночитаемый текст); формирование из распознанных строк целостного текста; экспорт итогового текста в выбранный пользователем формат.

Комплексное решение задач по распознаванию текстов предоставляет компания READ-COOP[7], образованная в 2019 году для поддержки и развития платформы Transkribus[8]. Ежемесячно к платформе обращаются около 1700 пользователей [11].

Функциональность платформы и анализ ряда проектов распознавания содержится в тематическом исследовании платформы «Преобразование научных знаний в архивах с помощью распознавания рукописного текста» [12], авторами которого указаны 54 специалиста.

С 2020 года использование платформы стало платным[9]. В 2024 году планируется ввод в эксплуатацию нового веб-приложения платформы.

Другими популярными платформами для распознавания рукописных текстов являются TEKLIA[10], eScriptorium[11] [13], OCR4all[12].

Одновременное существование нескольких платформ и движков распознавания ставит пользователей перед задачей выбора инструментов для обучения модели. Большая популярность платформы Transkribus не означает, по мнению специалистов, что она наиболее эффективна.

П.Б. Стробель (P.B. Ströbel) из Университета Цюриха предлагает для выбора платформы провести распознавание на нескольких платформах и остановиться на той, в которой модель покажет наилучшие результаты [14].

Перечислим несколько крупных проектов по распознаванию архивных документов, реализованных с применением ИИ.

Цель проекта Balsac[13] [15, 16] – создание демографической базы данных населения Квебека за период со второй половины XIX до начала XX века. Исходный материал – приходские метрические книги (реестры), хранящиеся в Национальной библиотеке и архивах Квебека (всего 44 742 реестра из 1 985 приходов). После расшифровки рукописных документов из текста извлекались необходимые именованные сущности – даты, лица (субъекты, родители, родст-

[7] We revolutionise Access to Historical Document. URL: https://readcoop.eu (дата обращения: 24.11.2023).

[8] Transkribus. URL: https://readcoop.eu/transkribus (дата обращения: 24.11.2023).

[9] Packages & Plans. URL: https://readcoop.eu/transkribus/credits/ (дата обращения: 24.11.2023).

[10] Automatic Document Processing with AI. Teklia. URL: https://teklia.com (дата обращения: 24.11.2023).

[11] Stokes P. RESILIENCE Tool: eScriptorium. RESILIENCE. 2020. URL: https://www.resilience-ri.eu/blog/resilience-tool-escriptorium/ (дата обращения: 24.11.2023).

[12] OCR4all: Optical Character Recognition (and more) for everyone. Centre for Philology and Digitality. URL: https://www.ocr4all.org (дата обращения: 24.11.2023).

[13] Balsac. Teklia. 2023. URL: https://teklia.com/research/projects/balsac/details/ (дата обращения: 24.11.2023).

венники, свидетели) и места; определялись связанные с ними действия – сведения о рождении, браке и смерти.

Проект библиотеки Сент-Женевьев[14] предназначен для перевода карточного каталога библиотеки в форму базы данных. Полностью автоматически удалось обработать 85% карточек.

Цель проекта «Секретный архив Ватикана» – произвести полную расшифровку реестров Ватикана, корпуса из более чем 18 тыс. страниц официальной переписки Римской курии, составленной в XIII веке [17].

Приведенные примеры зарубежных проектов можно продолжить, добавив проект «Картографирование средневековой Вены: социальная топография Вены в XV веке» [18], масштабный проект Национального архива Швеции по транскрибированию протоколов сыскной полиции Гетеборга за 1868–1902 годы [19], «Мастерские паспорта военного персонала» Тирольского государственного архива [20], расшифровку протоколов городских советов Белфорта [21] и многие другие.

В России реализуются пока немногочисленные проекты по распознаванию архивных документов с применением технологий ИИ.

Научно-исследовательский проект «Автографы Петра Великого: Чтение технологиями искусственного интеллекта» был инициирован Российским историческим обществом и ПАО «Сбербанк» [22–24]. Краткое описание этапов проекта содержится в разделе «Как это работает» на сайте «Digital Петр»[15]. На сайте «Автографы Петра I»[16] размещены 192 документа императора в виде изображений и их расшифровок с возможностью поиска по одному из двух хранилищ и по диапазону дат.

Проект «Расшифровка метрических книг, ревизских сказок и исповедных ведомостей с середины XVIII века до 1919 года» [25] реализует компания «Яндекс». Исходные документы хранятся в фондах Главархива Москвы и ряде госархивов других регионов.

Сложившееся к настоящему времени состояние технологий ИИ и опыта реализации проектов по распознаванию архивных документов позволяет констатировать, что происходит интенсивный процесс развития технологий и складывания методической и организационной базы этого направления. Осуществление проектов требует пока значительных финансовых и трудовых ресурсов и не включается в регулярную работу архивных учреждений. Важно отметить, что финансирование производится не из бюджетов архивов, а в рамках правительственных программ, а также за счет грантов исследовательских сообществ.

Важнейшим фактором развития распознавания архивных документов является широкое научное и практическое сотрудничество специалистов и профильных организаций в международном масштабе. Фактически сложилась целая индустрия этого направления. Она включает в себя размещение в открытом и бесплатном доступе алгоритмов ИИ по распознаванию,

[14] AI for cataloguing at the Sainte Geneviève library. Teklia. 2023. URL: https://teklia.com/blog/cataloguing-with-AI-at-BSG/ (дата обращения: 24.11.2023).

[15] Digital Пётр // Sber AI. URL: https://projects.tib.eu/en/viva/projekt/; https://fusionbrain.ai/digital-petr (дата обращения: 24.11.2023).

[16] Автографы Петра I: электронный архив // РИО. URL: https://peterscript.historyrussia.org/documents (дата обращения: 24.11.2023).

базовых моделей для разнообразных текстов, репозитарии наборов данных для машинного обучения[17], организацию конкурсов на лучшую технологию распознавания, репозитории научных и практических разработок[18].

Значительную роль в распознавании исторических документов играет деятельность Международной ассоциации по распознаванию образов (International Association for Pattern Recognition, IAPR)[19].

Ассоциация издает свой информационный бюллетень[20], спонсирует издание международного журнала по анализу и распознаванию документов[21].

Распознавание текстов является основной проблематикой ряда продолжающихся конференций, таких как Международная конференция по границам в распознавании рукописного текста (International conference on frontiers in handwriting recognition, ICFHR), Международная конференция по анализу и распознаванию документов (International Conference on Document Analysis and Recognition, ICDAR). Развитие инструментария распознавания текстов происходит также благодаря конкурсам, сопутствующим каждой конференции.

Значительный вклад в исследования применения ИИ в архивном деле вносит деятельность сообщества ученых и архивистов-практиков в рамках международного проекта InterPARES Trust AI (2021–2026). Работа участников проекта включает в себя проведение исследований, их публикацию, чтение лекций, проведение мастер-классов и семинаров (уже проведено 101 мероприятие), организацию конференций и симпозиумов (124 мероприятия)[22].

Использование технологий распознавания архивных документов в работе российских архивов может и должно базироваться на запросе общества и научных потребностях. В настоящее время существует четкий запрос на расшифровку генеалогической архивной информации, сосредоточенной в метрических книгах, исповедных росписях, ревизских сказках.

В отношении распознавания текстов для различного рода исторических, источниковедческих исследований ситуация (и технологии) сложнее. Успехи могут быть достигнуты при создании среды для подобных работ, «питательного бульона» для проведения экспериментов и реализации практически значимых проектов. Необходимо провести подготовку и включение в образовательные программы для архивистов и документоведов курсов по базовым знаниям в области ИИ, изучить научные и общественные потребности в конкретных проектах путем опросов, создать общедоступные модели, проработать экономические аспекты проектов.

---

[17] См., например, DIVAHisDB Dataset of Medieval Manuscripts. University of Fribourg. URL: https://www.unifr.ch/inf/diva/en/research/software-data/diva-hisdb.html (дата обращения: 24.11.2023).

[18] См., например, About Zenodo. CERN data centre & Invenio. URL: https://about.zenodo.org (дата обращения: 24.11.2023).

[19] IARP. URL: https://iapr.org (дата обращения: 24.11.2023).

[20] IAPR Newsletter. IARP. URL: https://iapr.org/articles/newsletter (дата обращения: 24.11.2023).

[21] International Journal on Document Analysis and Recognition (IJDAR). Springer. URL: https://www.springer.com/journal/10032/ (дата обращения: 24.11.2023).

[22] Research Dissemination. InterPARES TRUST AI. URL: https://interparestrustai.org/trust/research_dissemination (дата обращения: 24.11.2023).

Список источников

1. Seles A. A brave new world: artificial intelligence and archives [О дивный новый мир: искусственный интеллект и архивы] // *14th EASTICA General Conference and Seminar.* 2019. URL: https://www.archives.go.jp/english/news/pdf/20191125_27e_01.pdf (дата обращения: 24.11.2023).
2. Calvo-Zaragozaa J., Gallego A.-J. A selectional auto-encoder approach for document image binarization [Выборочный подход автоматического кодирования для бинаризации изображений документов] // Published in Pattern Recognition. 2018. URL: https://arxiv.org/pdf/1706.10241.pdf (дата обращения: 24.11.2023).
3. Westphal F., Lavesson N., Grahn H. Document image binarization using recurrent neural networks [Бинаризация изображения документа с использованием рекуррентных нейронных сетей] // *Proceedings – 13th IAPR International Workshop on Document Analysis Systems,* DAS 2018. P. 263–268. URL: http://www.diva-portal.org/smash/get/diva2:1231250/FULLTEXT01.pdf (дата обращения: 24.11.2023).
4. Xu Y., Yin F., Zhang Z., Liu C.-L. et al. Multi-task layout analysis for historical handwritten documents using fully convolutional networks [Многозадачный анализ макета исторических рукописных документов с использованием полностью сверточных сетей] // *Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*. 2018. URL: https://www.researchgate.net/publication/326201381_Multi-task_Layout_Analysis_for_Historical_Handwritten_Documents_Using_Fully_Convolutional_Networks (дата обращения: 24.11.2023).
5. Fischer A., Wüthrich M., Liwicki M., Frinken V. Automatic transcription of handwritten medieval documents [Автоматическая расшифровка рукописных средневековых документов] // Conference: Proc. 15th Int. Conf. on Virtual Systems and Multimedia (VSMM'09). 2009. URL: https://www.researchgate.net/publication/228370463 (дата обращения: 24.11.2023).
6. Zhao L., Wu Z., Wu X., Wilsbacher G., Wang S. Background-insensitive scene text recognition with text semantic segmentation [Независимое от фона распознавание текста сцены с семантической сегментацией текста] // Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science. Vol. 13685. URL: https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136850161.pdf (дата обращения: 24.11.2023).
7. Model Training [Модельное обучение] // READ-COOP SCE. URL: https://readcoop.eu/glossary/model-training/ (дата обращения: 24.11.2023).
8. Clérice T. Ground-truth free evaluation of HTR on old french and latin medieval literary manuscripts [Настоящая бесплатная оценка HTR на старофранцузских и латинских средневековых литературных рукописях] // CHR 2022: Computational Humanities Research Conference, December 12–14, 2022, Antwerp, Belgium. URL: https://ceur-ws.org/Vol-3290/long_paper2081.pdf (дата обращения: 24.11.2023).
9. Kiessling B. Kraken – a Universal Text Recognizer for the Humanities [Kraken — универсальный распознаватель текста для гуманитарных наук] // DataverseNL. 2019. V2. URL: https://dh-abstracts.library.virginia.edu/works/9912 (дата обращения: 24.11.2023).
10. Wick C., Reul C., Puppe F. Calamari – a high-performance Tensorflow-based deep learning package for optical character recognition [Calamari — высокопроизводительный пакет глубокого обучения на основе Tensorflow для оптического распознавания символов] // Digital Humanities Quarterly. 2020. Vol. 14. № 2. URL: http://www.digitalhumanities.org/dhq/vol/14/2/000451/000451.html (дата обращения: 24.11.2023).

11. Nockels J.,Gooding P., Ames S., Terras M. Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research [Понимание применения технологии распознавания рукописного текста в контексте наследия: систематический обзор Transkribus в опубликованных исследованиях] // Archival Science. 2022. 22. P. 367–392. URL: https://doi.org/10.1007/s10502-022-09397-0 (дата обращения: 24.11.2023).
12. Muehlberger G. Transforming scholarship in the archives through handwritten text recognition Transkribus as a case study [Преобразование научных знаний в архивах посредством распознавания рукописного текста Transkribus как практический пример] // Journal of Documentation. 2018. Vol. 75. Issue 5. P. 954-976.
13. Stokes P.A., Kiessling B. The eScriptorium VRE for manuscript cultures [eScriptorium VRE для рукописного культурного наследия] // Classics. 2021. Vol. 18. URL: https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/ (дата обращения: 24.11.2023).
14. Ströbel P.B. Flexible Techniques for Automatic Text Recognition of Historical Documents [Гибкие методы автоматического распознавания текста исторического документа] // University of Zurich. 2023. URL: https://www.researchgate.net/publication/372501015_Flexible_Techniques_for_Automatic_Text_Recognition_of_Historical_Documents (дата обращения: 24.11.2023).
15. Maarand M. BALSAC project registers have been processed! [Реестры проектов BALSAC обработаны!] // Teklia. 2022. URL: https://teklia.com/blog/202202-balsac/ (дата обращения: 24.11.2023).
16. Tarride S. Large Scale Genealogical information extraction from handwritten Quebec Parish Records [Крупномасштабное извлечение генеалогической информации из рукописных приходских записей Квебека] // Research square. 2022. URL: https://doi.org/10.21203/rs.3.rs-2260181/v1 (дата обращения: 24.11.2023).
17. Firmani D., Maiorino M., Merialdo P., Nieddu E. Towards knowledge discovery from the Vatican Secret Archives. In codice ratio – Episode 1: Machine transcription of the manuscripts [На пути к открытию знаний из секретных архивов Ватикана. В кодовом соотношении – Эпизод 1: Машинная расшифровка рукописей] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). ACM, New York, NY, USA. 2018. URL: https://arxiv.org/abs/1803.03200 (дата обращения: 24.11.2023).
18. Ertl T., Schmidle W., Helmchen J., Duval T. Mapping. Medieval Vienna: The Social Topography of Vienna in the 15th Century [Картирование. Средневековая Вена: социальная топография Вены XV века] // Freie Universitat, Berlin. URL: https://www.geschkult.fu-berlin.de/e/fmi/bereiche/mittelalter/ab_ertl/Mapping-Vienna.html (дата обращения: 24.11.2023).
19. Karsvall O. Maskintolkning av handskrivna källmaterial [Машинная интерпретация рукописных исходных материалов]. *RIKSARKIVET*. URL: https://riksarkivet.se/htr (дата обращения: 24.11.2023).
20. How to make a complete collection accessible with Transkribus. A best-practice example from the Tyrolean State Archives [Как сделать полную коллекцию доступной с помощью Transkribus. Пример передового опыта из Тирольского государственного архива] // READ-COOP SCE. URL: https://readcoop.eu/success-stories/grundbuchblaetter/ (дата обращения: 24.11.2023).

21. Kermorvant C. Belfort city archives: a pilot project for automatic recognition of city council registers [Городские архивы Белфорта: пилотный проект автоматического распознавания реестров городского совета] // Teklia. 2022. URL: https://teklia.com/blog/202211-belfort-en/ (дата обращения: 24.11.2023).
22. Базарова Т., Димитров Д., Потанин М., Проскурякова М. Распознать и транскрибировать: автографы Петра Великого и технологии искусственного интеллекта // Воронцово поле. 2020. № 4. С. 64–71. URL: https://portal.historyrussia.org/img/news/VP_4.2020.pdf#page=33 (дата обращения: 24.11.2023).
23. Владимир Аракчеев: Идея проекта Петру пришлась бы по душе // Российской историческое общество. 2021. URL: https://historyrussia.org/sobytiya/ideya-proekta-petru-prishlas-by-po-dushe.html (дата обращения: 24.11.2023).
24. Сиренов А.В. Здесь мы видим действительно научный метод // Воронцово поле. 2020. № 4. С. 74–75.
25. Андреев А. Нейросеть поможет прочитать старинные рукописи // Стимул. 2023. URL: https://stimul.online/articles/science-and-technology/neyroset-pomozhet-prochitat-starinnye-rukopisi/ (дата обращения: 24.11.2023).

References

1. Seles A. A brave new world: artificial intelligence and archives. *14th EASTICA General Conference and Seminar.* 2019. URL: https://www.archives.go.jp/english/news/pdf/20191125_27e_01.pdf (accessed: 24.11.2023).
2. Calvo-Zaragozaa J., Gallego A.-J. A selectional auto-encoder approach for document image binarization. Published in Pattern Recognition. 2018. URL: https://arxiv.org/pdf/1706.10241.pdf (accessed: 24.11.2023).
3. Westphal F., Lavesson N., Grahn H. Document image binarization using recurrent neural networks. *Proceedings – 13th IAPR International Workshop on Document Analysis Systems,* DAS 2018. P. 263–268. URL: http://www.diva-portal.org/smash/get/diva2:1231250/FULLTEXT01.pdf (accessed: 24.11.2023).
4. Xu Y., Yin F., Zhang Z., Liu C.-L. et al. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. *Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*. 2018. URL: https://www.researchgate.net/publication/326201381_Multi-task_Layout_Analysis_for_Historical_Handwritten_Documents_Using_Fully_Convolutional_Networks (accessed: 24.11.2023).
5. Fischer A., Wüthrich M., Liwicki M., Frinken V. Automatic transcription of handwritten medieval documents. *Conference: Proc. 15th Int. Conf. on Virtual Systems and Multimedia (VSMM'09*). 2009. URL: https://www.researchgate.net/publication/228370463 (accessed: 24.11.2023).
6. Zhao L., Wu Z., Wu X., Wilsbacher G., Wang S. Background-insensitive scene text recognition with text semantic segmentation. *Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science*. Vol. 13685. URL: https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136850161.pdf (accessed: 24.11.2023).
7. Model Training. *READ-COOP SCE*. URL: https://readcoop.eu/glossary/model-training/ (accessed: 24.11.2023).
8. Clérice T. Ground-truth free evaluation of HTR on old French and Latin medieval literary manuscripts. *CHR 2022: Computational Humanities Research Conference, December 12–14, 2022, Antwerp, Belgium*. URL: https://ceur-ws.org/Vol-3290/long_paper2081.pdf (accessed: 24.11.2023).

9. Kiessling B. Kraken – a Universal Text Recognizer for the Humanities. *DataverseNL*. 2019. V2. URL: https://dh-abstracts.library.virginia.edu/works/9912 (accessed: 24.11.2023).
10. Wick C., Reul C., Puppe F. Calamari – A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *Digital Humanities Quarterly.* 2020;2(14). URL: http://www.digitalhumanities.org/dhq/vol/14/2/000451/000451.html (accessed: 24.11.2023).
11. Nockels J., Gooding P., Ames S., Terras M. Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Archival Science*. 2022;22:367–392. URL: https://doi.org/10.1007/s10502-022-09397-0 (accessed: 24.11.2023).
12. Muehlberger G. Transforming scholarship in the archives through handwritten text recognition Transkribus as a case study. *Journal of Documentation.* 2018;5(75):954-976.
13. Stokes P.A., Kiessling B. The eScriptorium VRE for manuscript cultures. *Classics.* 2021. Vol. 18. URL: https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/ (accessed: 24.11.2023).
14. Ströbel P.B. Flexible Techniques for Automatic Text Recognition of Historical Documents. *University of Zurich*. 2023. URL: https://www.researchgate.net/publication/372501015_Flexible_Techniques_for_Automatic_Text_Recognition_of_Historical_Documents (accessed: 24.11.2023).
15. Maarand M. BALSAC project registers have been processed! *Teklia*. 2022. URL: https://teklia.com/blog/202202-balsac/ (accessed: 24.11.2023).
16. Tarride S. Large Scale Genealogical information extraction from handwritten Quebec Parish Records. *Research square*. 2022. URL: https://doi.org/10.21203/rs.3.rs-2260181/v1 (accessed: 24.11.2023).
17. Firmani D., Maiorino M., Merialdo P., Nieddu E. Towards knowledge discovery from the Vatican Secret Archives. In codice ratio – Episode 1: Machine transcription of the manuscripts. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. ACM, New York, NY, USA. 2018. URL: https://arxiv.org/abs/1803.03200 (accessed: 24.11.2023).
18. Ertl T., Schmidle W., Helmchen J., Duval T. Mapping. Medieval Vienna: The Social Topography of Vienna in the 15th Century. *Freie Universitat,* Berlin. URL: https://www.geschkult.fu-berlin.de/e/fmi/bereiche/mittelalter/ab_ertl/Mapping-Vienna.html (accessed: 24.11.2023).
19. Karsvall O. Maskintolkning av handskrivna källmaterial. *RIKSARKIVET*. URL: https://riksarkivet.se/htr (accessed: 24.11.2023). (In Swedish).
20. How to make a complete collection accessible with Transkribus. A best-practice example from the Tyrolean State Archives. *READ-COOP SCE*. URL: https://readcoop.eu/success-stories/grundbuchblaetter/ (accessed: 24.11.2023).
21. Kermorvant C. Belfort city archives: a pilot project for automatic recognition of city council registers. T*eklia*. 2022. URL: https://teklia.com/blog/202211-belfort-en/ (accessed: 24.11.2023).
22. Bazarova T., Dimitrov D., Potanin M., Proskuryakova M. Raspoznat` i transkribirovat`: avtografy` Petra Velikogo i texnologii iskusstvennogo intellekta [Recognize and transcribe: autographs of Peter the Great and artificial intelligence technologies]. *Voronczovo pole = Vorontsovo Pole.* 2020;4:64–71. URL: https://portal.historyrussia.org/img/news/VP_4.2020.pdf#page=33 (accessed: 24.11.2023). (In Russian).

23. Vladimir Arakcheev: Ideya proekta Petru prishlas` by` po dushe [Vladimir Arakcheev: Peter would have liked the idea of the project]. *Russian Historical Society.* 2021. URL: https://historyrussia.org/sobytiya/ideya-proekta-petru-prishlas-by-po-dushe.html (accessed: 24.11.2023). (In Russian).
24. Sirenov A.V. Zdes` my` vidim dejstvitel`no nauchny`j metod [Here we see a truly scientific method]. *Voronczovo pole = Vorontsovo Pole.* 2020;4:74–75. (In Russian).
25. Andreev A. Nejroset` pomozhet prochitat` starinny`e rukopisi [A neural network will help read ancient manuscripts]. *Stimul.* 2023. URL: https://stimul.online/articles/science-and-technology/neyroset-pomozhet-prochitat-starinnye-rukopisi/ (accessed: 24.11.2023). (In Russian).

ИНФОРМАЦИЯ ОБ АВТОРАХ

Киселёв Игорь Николаевич, кандидат исторических наук, старший научный сотрудник отдела архивоведения Всероссийского научно-исследовательского института документоведения и архивного дела (ВНИИДАД), Москва, Российская Федерация.

INFORMATION ABOUT THE AUTHORS

Igor N. Kiselev, PhD (in history), senior researcher of Archival Science Department of the All-Russian Scientific and Research Institute for Records and Archives Management (VNIIDAD), Moscow, Russian Federation.