УДК 004.8:930.25

Мащенко Наталья Евгеньевна, канд. экон. наук, доцент, доцент кафедры информационных систем управления ФГБОУ ВО «Донецкий государственный университет», г. Донецк, Россия

e-mail: maschenko_n@inbox.ru

Сафьянова София Сергеевна, студентка 2 курса направления подготовки «Системный анализ и управление» ФГБОУ ВО «Донецкий государственный университет», г. Донецк, Россия

e-mail: sofiasafianova@mail.ru

СИСТЕМНО-АНАЛИТИЧЕСКИЙ ПОДХОД К ПРОЦЕССУ ОЦИФРОВКИ РУКОПИСНЫХ АРХИВНЫХ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация. В статье исследуется применение системно-аналитического подхода к процессу оцифровки рукописных архивных документов с использованием технологий искусственного интеллекта. Оцифровка рассматривается как сложная социотехническая система, включающая технологические, организационные и экспертные компоненты. Раскрыта структура процесса цифровизации, включающая этапы сканирования, предобработки изображений, анализа особенностей рукописного текста, распознавания, постобработки и контроля качества. Особое внимание уделено применению нейросетевых методов распознавания рукописного текста (HTR), сверточных и трансформерных моделей, а также языковых моделей для контекстной верификации и коррекции результатов распознавания. Проанализированы современные инструменты (Transkribus, Kraken, языковые модели) и практика их применения в архивной деятельности на примере проекта «Архивный десант». Выявлены основные проблемы цифровизации и обоснована эффективность гибридной модели взаимодействия человека и

искусственного интеллекта.

Ключевые слова: оцифровка, архивные документы, рукописные документы, системный анализ, искусственный интеллект, нейросетевые модели, распознавание рукописного текста, цифровые архивы

Mashchenko Natalia Evgenievna, PhD in Economics, Associate Professor, Associate Professor of the Department of Information Management Systems of the Donetsk State University, Donetsk, Russia

e-mail: maschenko_n@inbox.ru

Safyanova Sofia Sergeevna, 2nd year student of the field of study " System analysis and management ", Donetsk State University, Donetsk, Russia

SYSTEM-ANALYTICAL APPROACH TO THE DIGITIZATION OF HANDWRITTEN ARCHIVAL DOCUMENTS USING AI

Annotation. The article explores the application of a system-analytical approach to the digitization of handwritten archival documents using artificial intelligence technologies. Digitization is considered as a complex socio-technical system that includes technological, organizational, and expert components. The structure of the digitization process is revealed, including scanning, preprocessing, recognition, post-processing, and quality control stages. Special attention is paid to neural network handwriting recognition (HTR), convolutional and transformer models, and language models for contextual verification. Modern tools (Transkribus, Kraken, language models) and their practical application are analyzed. The effectiveness of a hybrid human–AI interaction model is substantiated.

Keywords: digitization, archival documents, handwriting recognition, artificial intelligence, neural networks, digital archives.

Цифровизация архивного наследия является одной из приоритетных задач современной гуманитарной науки и информационных технологий.

В крупнейших архивохранилищах Российской Федерации, в частности в Российском государственном архиве древних актов [2], сосредоточены значительные массивы рукописных документов, доступ к которым ограничен физическим форматом хранения. Это создает риски утраты и затрудняет научное использование источников. И в России также была создана единая система удалённого доступа к архивам – ГИС УИАД [3], объединяющая фонды 13 федеральных архивов и предоставляющая доступ к десяткам миллионов описей и дел через интернет. Это яркий пример того, как цифровизация содействует научному и общественному использованию источников.

Оцифровка позволяет решать задачи сохранения культурного наследия, расширения доступа к документам и интеграции архивов в цифровую исследовательскую среду [4]. Вместе с тем обработка рукописных документов представляет собой сложный многоуровневый процесс, требующий применения системно-аналитического подхода и технологий искусственного интеллекта [7].

Целью исследования является обоснование эффективности системно-аналитического подхода в сочетании с технологиями искусственного интеллекта при оцифровке рукописных архивных документов.

Методологическую основу исследования составляет системный анализ, процессное моделирование и сравнительный анализ технологий распознавания текста.

Процесс оцифровки рассматривается как сложная социотехническая система, включающая взаимосвязанные технологические, организационные и экспертные компоненты. В рамках системного подхода процесс цифровизации структурируется на взаимосвязанные этапы: создание цифровых копий документов; анализ особенностей архивного фонда; проектирование архитектуры обработки данных; моделирование и обучение нейросетевых моделей [5]; внедрение, эксплуатация и контроль качества. Такой подход обеспечивает управляемость процесса и позволяет минимизировать технологические и интерпретационные ошибки.

Технологические этапы цифровизации включают следующие стадии.

Начальный этап предполагает создание качественных цифровых копий документов с соблюдением требований к разрешению, освещению и сохранности метаданных.

После сканирования осуществляется предобработка изображений, включающая выравнивание, очистку, повышение контрастности и устранение искажений. Качество данной стадии оказывает непосредственное влияние на точность последующего распознавания. Далее осуществляется анализ особенностей рукописного текста: особенности почерка, историческая орфография, языковые нормы соответствующей эпохи, наличие сокращений и специфических символов [7].

Классические OCR-системы (Google Vision, ABBYY, Яндекс) демонстрируют высокую эффективность при обработке печатных текстов, однако этот метод часто не справляется с плохими сканами или искривлёнными линиями текста, тем более при распознавании рукописей их возможности существенно ограничены. В связи с этим применяются специализированные HTR-решения (Handwritten Text Recognition), предполагающие обучение моделей на размеченных выборках конкретного фонда [5]. На практике используется гибридный подход: автоматическое распознавание с корректурой модели.

Проектирование системы предполагает формирование архитектуры обработки данных: сегментацию изображения, распознавание текста, постобработку, проверку и публикацию.

На этапе моделирования и обучения ИИ создаются и тестируются нейросетевые модели [5].

Современные решения в области распознавания рукописного текста базируются на использовании сверточных нейронных сетей и трансформерных моделей, анализирующих текст как целостную структуру [8]. Они анализируют слово целиком, а не отдельные символы, что существенно повышает точность. Такой подход обеспечивает повышение точности по сравнению с

посимвольным распознаванием.

Среди используемых специализированных инструментов выделяют:

Transkribus – платформа для обучения моделей под конкретные типы почерка [10];

Kraken – гибкая система с открытым исходным кодом [11];

крупные языковые модели, включая Qwen [1], применяемые для контекстной проверки и коррекции предварительно распознанного текста.

Использование языковых моделей позволяет повысить согласованность текста и снизить количество ошибок при автоматическом распознавании.

После этапа тестирования система интегрируется в рабочую среду, где осуществляется автоматическая обработка документов, проводится выборочная экспертная проверка и осуществляется дообучение моделей на основе исправленных текстов [5].

Процесс приобретает циклический характер: результаты экспертной корректировки используются для повышения точности последующих распознаваний и дообучения моделей.

Примером прикладной реализации системно-аналитического подхода является проект «Архивный десант» (Научный Центр исторической памяти при Президенте РФ) [9], направленный на научно-просветительскую деятельность по сохранению памяти о Великой Отечественной войне. В рамках проекта осуществляется поиск, расшифровка и публикация архивных документов, связанных с судьбами жертв нацистской политики. В процессе работы используются цифровые инструменты распознавания текста и языковые модели для предварительной дешифровки рукописных материалов [6].

В рамках проекта «Архивный десант» [9] осуществляется практическое применение современных инструментов искусственного интеллекта при работе с архивными документами, позволяющие осуществлять предварительное распознавание текста, ускорять анализ документов и повышать точность обработки данных. Впоследствии эксперты проверяют правильность текста и проводят археографическую обработку документов.

В частности, для поддержки процесса дешифровки рукописных источников используются нейросетевые модели, включая языковую модель Qwen (чат-интерфейс chat.qwen.ai) [1].

В целом, алгоритм обработки цифровых материалов включает получение цифровых копий архивных документов; предварительное распознавание текста с помощью нейросетей; контекстную проверку через языковые модели (Qwen); экспертную корректировку; публикацию проверенных материалов.

Алгоритм обработки предполагает передачу предварительно распознанного текста в языковую модель для проведения контекстной верификации и коррекции ошибок с опорой на исходный документ. Применение данной технологии позволяет автоматизировать первичный анализ сложных буквенных сочетаний, реконструкцию вероятных словоформ и устранение типичных ошибок распознавания [7].

Полученные результаты демонстрируют эффективность гибридной модели взаимодействия человека и искусственного интеллекта в архивной деятельности [8]. ИИ выполняет функции предварительной интеллектуальной обработки текста, тогда как эксперт осуществляет контроль достоверности, уточнение интерпретаций и финальную редакцию. Подобная модель взаимодействия обеспечивает повышение скорости обработки материалов при сохранении научной точности.

Несмотря на технологический прогресс, процесс цифровизации сопровождается рядом трудностей: низкое качество сканов и повреждения документов снижают точность распознавания – требуется качественная съёмка и предобработка; историческая орфография и старые языковые формы требуют дообучения моделей [5] и использования словарей; неразборчивый почерк часто нуждается в ручной разметке и корректировке.

Подготовка данных ресурсоёмка, поэтому оптимально использование гибридного подхода, сочетающего автоматизированное распознавание и экспертную проверку с последующим дообучением моделей.

Проведённое исследование позволяет сделать вывод о том, что, что

оцифровка рукописных архивных документов представляет собой сложный многоуровневый процесс, требующий системной организации и междисциплинарного подхода. Применение системно-аналитической методологии обеспечивает структурирование этапов цифровизации, координацию технологических и экспертных компонентов, а также снижение вероятности ошибок на различных стадиях обработки.

Использование нейросетевых технологий, включая специализированные HTR-системы и крупные языковые модели, существенно повышает точность и скорость распознавания рукописных текстов. Особенно значимым является внедрение механизмов контекстной верификации, позволяющих минимизировать типичные ошибки автоматического распознавания и повысить согласованность итогового текста.

Таким образом, сочетание системного анализа и технологий искусственного интеллекта формирует методологическую и технологическую основу современной цифровой трансформации архивов, способствуя сохранению исторического наследия и расширению исследовательского доступа к источникам.

Список литературы:

1. Qwen Chat – китайский ИИ-ассистент [Электронный ресурс]. – URL: https://chat.qwen.ai (дата обращения: 25.03.2026).

2. РГАДА – российский государственный архив древних актов [Электронный ресурс]. – URL: http://rgada.info/ (дата обращения: 25.03.2026).

3. Архивные фонды России в одном окне: ЭЛАР вывел ГИС УИАД на новый уровень [Электронный ресурс] // Официальный сайт ЭЛАР. – 2025. – URL: https://elar.ru/press-center/news/arkhivnye-fondy-rossii-v-nbsp-odnom-okne-elar-vyvel-gis-uiad-na-nbsp-novyy-uroven/ (дата обращения: 25.03.2026).

4. Краснова Е. Л. Сохранение и трансляция культурного наследия в цифровую эпоху: к построению модели [Электронный ресурс] // Музей. Памятник. Наследие. – 2022. – № 1 (11). – С. 134-140. – URL:

https://museumstudy.ru/wp-content/uploads/2017/11/11-2022-ИТОГ-134-140.pdf (дата обращения: 27.03.2026).

5. Li Y., Chen D., Tang T., Shen X. HTR-VT: распознавание рукописного текста с использованием Vision Transformer [Электронный ресурс] // Pattern Recognition. – 2025. – Т. 158. – Статья 110967. – URL: https://www.sciencedirect.com/science/article/pii/S0031320324007180 (дата обращения: 05.04.2026).

6. ОРКРТ (ОСГ РМ) [Электронный ресурс] // Тенденции цифровизации в архивном секторе. 2020. — URL: https://www.osgrm.ru/info/news/tendentsii-tsifrovizatsii-v-arkhivnom-sektore/ (дата обращения: 25.03.2026

7. Garrido-Muñoz C., Ríos-Vila A., Calvo-Zaragoza J. Handwritten Text Recognition: A Survey (Распознавание рукописного текста: обзор) [Электронный ресурс] // arXiv. – 2025. – DOI: 10.48550/arXiv.2502.08417. – URL: https://arxiv.org/abs/2502.08417 (дата обращения: 05.04.2026).

8. Mahadevkar S., Patil S., Kotecha K. Повышение точности распознавания рукописного текста с использованием гибридного подхода на основе искусственного интеллекта [Электронный ресурс] // MethodsX. – 2024. – Т. 12. – Статья 102654. – URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10950881/ (дата обращения: 05.04.2026).

9. Архивный десант [Электронный ресурс] // Научный центр исторической памяти при Президенте Российской Федерации. — URL: https://russiancip.ru/nauchno-prosvetitelskij-proekt-dostupnyj-arhiv/arhivnyj-desant/ (дата обращения: 25.03.2026).

10. Transkribus – платформа для автоматического распознавания и транскрибирования исторических документов [Электронный ресурс] // Transkribus. – URL: https://www.transkribus.org/ (дата обращения: 27.03.2026).

11. Политика конфиденциальности Kraken [Электронный ресурс] // Kraken: официальный сайт компании. – URL: https://www.kraken.com/ru/legal/privacy (дата обращения: 27.03.2026).