Историческая информатика
Правильная ссылка на статью:
Юмашева Ю.Ю. К вопросу о применении искусственного интеллекта в исторических исследованиях // Историческая информатика. 2025. № 1. DOI: 10.7256/2585-7797.2025.1.73578 EDN: PQTZJT URL: https://nbpublish.com/library_read_article.php?id=73578

К вопросу о применении искусственного интеллекта в исторических исследованиях

Юмашева Юлия Юрьевна

ORCID: 0000-0001-8353-5745

доктор исторических наук

Заместитель генерального директора ООО "ДИМИ-ЦЕНТР"

105264, Россия, г. Москва, бул. Измайловский, 43

Juliayu@yandex.ru

Статья из рубрики "Искусственный интеллект и наука о данных"

DOI:

10.7256/2585-7797.2025.1.73578

EDN:

PQTZJT

Дата направления статьи в редакцию:

04-03-2025

Аннотация: Статья посвящена дискуссионной проблеме применения искусственного интеллекта в исторических исследованиях. Во введении кратко рассматривается история возникновения «искусственного интеллекта» (ИИ) как направления в информатике, эволюция этого определения и взглядов на области применения ИИ; анализируется место методов искусственного интеллекта на разных этапах конкретно-исторических исследований. В основной части статьи на основе анализа историографических источников и собственного опыта участия в зарубежных проектах автор анализирует практику реализации проектов распознавания рукописного текста с помощью различных информационных технологий и методов ИИ, в частности, описываются и обосновываются требования к созданию электронных копий распознаваемых источников, необходимость учета фактуры носителей информации, писчих материалов, техники и технологии создания текста; разновидности и способы создания палеографических, кодикологических, дипломатических наборов данных, историко-лексикологических словарей, возможности использования больших языковых моделей и т.п. В качестве

методологической основы автор использовал системный подход, историко-сравнительный, историко-хронологический и описательный методы, а также анализ историографических источников. Учитывая то, что в российской исторической науке применение технологий и методов искусственного интеллекта является довольно редким явлением, анализ опыта осуществления подобных зарубежных проектов весьма актуален, так же как и характеристика профильных научных ассоциаций, научных и научно-вспомогательных ресурсов (порталов и сайтов с наборами данных и исследовательским инструментарием), размещенных в сети Интернет, и сборников научных трудов по изучаемой проблематике, неизвестных в России, о которых идет в речь в статье. В заключение делается вывод перспективности применения технологий искусственного интеллекта не только в качестве вспомогательного инструментария, но и как исследовательских методов, помогающих в установлении авторства исторических источников, уточнении их датировки, выявления подделок и т.п., а также в создании новых видов научно-справочных поисковых систем архивов и библиотек. Вместе с тем, использование технологий искусственного интеллекта отличается большой затратностью и капиталоемкостью, что является серьезным препятствием для широкого внедрения данных технологий в практику исторических исследований.

Ключевые слова:

искусственный интеллект, исторические источники, автоматизированное распознавание текстов, палеография, кодикология, дипломатика, историческая лексикология, наборы данных, большие языковые модели, информационные технологии

Введение

Заданная редакцией журнала «Историческая информатика» тема очередного выпуска «Искусственный интеллект в исторических исследованиях и образовании», на наш взгляд, предполагает необходимость уточнения двух понятий: «искусственный интеллект» и «этапы исторического исследования».

Оставив за границами рассмотрения историю идеи о «мыслящем искусственном существе» («мыслящей машине»), которые витали в воздухе еще со времен Аристотеля, обратимся к более близким временам, а именно к рассмотрению термина Artificial Intelligence (AI, искусственный интеллект, ИИ), который был сформулирован и впервые введен в научный оборот на основе практически реализованных проектов (в т.ч. разработанной в начале 1950-х гг. первой модели нейронных сетей – The Stochastic Neural Analog Reinforcement Calculator (SNARC)[1]) на Dartmouth workshop [2], который проходил в 1956 г. в Ганновере, штат Нью-Гемпшир. В то время организаторы этого семинара (Дж. Маккарти, Марвин Мински, Натаниэль Рочестер и Клод Шеннон) и докладчики (среди которых были Аллен Ньюэлл и Герберт Саймон – авторы программы «Logic Theorist», разработанной в 1955 г. [3]) не формулировали определение этого термина, справедливо полагая, что для корректной дефиниции необходимо сначала дать определение понятию «интеллект». Однако еще на этапе подготовки к семинару Дж. Маккарти так обозначал предмет обсуждения: «всякий аспект обучения или любой другой признак интеллекта может в принципе быть настолько точно описан, что машину можно заставить его симулировать». Это положение стало основанием для выделения в 1957 г. основных направлений применения AI: «машинный перевод, машинное обучение, автоматизированное распознавание (образов/письма/звучащей речи) и принятие

решений» [4].

Прошло без малого 70 лет, и после определенных успехов AI в конце 1950-х–1970-х гг., (разработки в 1958 г. архитектуры искусственной нейронной сети – персептона [5], в 1959 г. первой программы «машинного самообучения» – Samuel Checkers-playing (компьютерная программа игры в шашки) [6], языков программирования List Processing language (LISP, «язык обработки списков» [7]), одной из первых компьютерных шахматных программ [8]*, программы ELIZA [9, 10]**, экспертных систем, методов имитационного моделирования [11, 12, 13, 14, 15, 16, 17]*** и т.п.), и последовавшей за этим периодом «зимы AI», длившейся почти 20 лет и связанной с микрокомпьютерной революцией и нехваткой данных, с середины 1990-х гг. возникла новая, повторная волна интереса к AI, детерминированная технологическими сдвигами, взрывным увеличением мощностей компьютерной техники, появлением направления Data Science и развитием Data Engineering.

К концу 1990-х гг. было сформулировано максимально общее определение AI, которое давало и дает широкий простор для именования этим термином многих технологий, связанных с обработкой данных: «Искусственный интеллект – это область компьютерных наук, направленная на создание систем, способных выполнять задачи, требующие человеческого интеллекта». Претерпели изменения и области применения AI. К четырем ранее определенным направлениям, сохранившим свое значение, были добавлены роботехника со встроенным AI, распознавание и воспроизведение человеческих эмоций (робот Kismet [18]) и т.п.

Прошло еще 20 лет, в течение которых произошло огромное увеличение вычислительных мощностей, накопление больших объемов данных, появление генеративного AI и больших языковых моделей (LLM); разработкой новых технологических решений, внедрением виртуальных помощников и ботов, способных общаться на естественном языке…

Все эти события вновь поставили вопрос об уточнения определения понятия AI и областей его применения. Погружаясь в изучение историографии AI, трудно не согласиться с мнением немецких исследователей А. Каплана и М. Хейнлейна, которые писали, что «AI по-прежнему остается на удивление размытой концепцией, и многие вопросы, связанные с ней, остаются открытыми… Мы предполагаем, что AI – это не один монолитный термин…», а совокупность технологий, изучение которой возможно только «через призму эволюционных стадий (искусственный узкий интеллект, искусственный общий интеллект и искусственный суперинтеллект) или сосредоточившись на различных типах систем AI (аналитический ИИ, ИИ, вдохновленный человеком, и очеловеченный AI)» [19].

Это же понимание сущности AI зафиксировано и в наиболее близкой, по мнению автора, к содержанию этой статьи дефиниции, закрепленной в российском ГОСТ: «Искусственный интеллект – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека» [20].

Резюмируя краткий экскурс в историю представлений об AI, отметим, что изначально определенные направления его применения остаются практически неизменными, но при этом они проникают во все новые отрасли знания. Этот факт дает возможность

сосредоточиться на рассмотрении одной из основополагающих задач использования AI, а именно на **методах распознавания текстов/изображений и создании машиночитаемых источников**, как базового условия взаимодействия с компьютерными приложениями для проведения дальнейшего анализа информации.

Теперь рассмотрим содержание понятия «этапы исторического исследования». В отечественной историографии существует несколько вариантов списка этапов исторического исследования, детерминированных областью исследований, однако, большинство авторов рассматривают его на примере проблемно-исторических исследований событий, процессов, явлений, исторических персон и т.п. и включают такие обязательные пункты как:

1. Выбор темы на основе первичного анализа историографии;

2 . Подробный историографический анализ и определение объекта и предмета (исследовательских задач) исследования;

3. Разработка рабочей гипотезы;

4 . Выявление и отбор источников в соответствии с объектом, предметом, поставленными задачами и рабочей гипотезой (эвристика);

5. Определение и применение методов изучения источников (в том числе – подготовки их информации для последующего анализа), **адекватных источниковой базе**;

6 . Определение и применение методов анализа источников, адекватных источникам и поставленным задачам, получение результатов;

7. Интерпретация (теоретическое объяснение) полученных результатов;

8. Корректировка первоначальной рабочей гипотезы, формулирование выводов;

9. Интеграция выводов в исторический контекст;

1 0 . Определение результатов, не укладывающихся в имеющиеся концепции исторического развития;

11. Постановка новой исследовательской задачи и т.п.

Очевидно, что в пп. 1–3 вряд ли возможно применение AI – для этого было бы необходимо оцифровать всю научную литературу в мире, постоянно поддерживая этот процесс новыми публикациями, разработать совершенные методы ее *содержательного анализа* (с учетом традиций национальных и (у́же) научных школ) и выявления неисследованных тем, проблем, вопросов, т.е. историографических «лакун», которые могли бы стать объектом изучения, и при этом, были бы обеспечены источниковой базой, или к которым могли бы быть применены новые методы исследования.

На сегодняшний день такие задачи под силу только человеку, работающему в избранной предметной области и обладающему экспертными знаниями.

Примерно такие же выводы можно сделать и в отношении пп. 7, 8 и 9, т.к. применение генеративного AI для решения этих задач не может в принципе удовлетворить исследователей, поскольку результат генерации ответа всегда основан на уже введенных в машину знаниях, в то время как искомый итог любого научного

исследования должен быть нацелен *на получение новых выводов* («приращение научного знания» – © Л.И.Бородкин****).

Пп. 10 и 11 являются в некотором смысле «факультативными» и/или могут быть раскрыты в нарративном описании итогов исследования.

Набор этапов исследования в не менее важных исторических исследованиях в области источниковедения и вспомогательных исторических дисциплин мало чем отличается от перечня этапов проблемно-ориентированных работ, однако, методы исследования в них отмечены большей междисциплинарностью, т.е. использованием не только информационных технологий, но и привлечением методов естественно-научных дисциплин, что выводит результаты этих исследований на уровень *эмпирического знания*.

Таким образом, применение AI в исторических научных исследованиях на сегодняшний день ограничено этапами:

- выявления и отбора источниковой базы (п. 4) (создание новых разновидностей архивного НСА в виде наборов данных (баз знаний, dataset, ds), семантических поисковых систем и, возможно, систем, использующих генеративный AI***** [21]). Внедрение AI на этом этапе знаменует *новую стадию развития системы научно-справочных средств архивов*, а также *библиотечной и архивной эвристики (эвристика информационной эпохи)*;

- применения специализированных методов подготовки комплексов источников для их последующего изучения (п. 5) (преобразование исторических источников в цифровой вид, формирование наборов данных и терминологических словарей, разработку больших языковых моделей (LLM) на основе исторической лексики и т.п.). На этом этапе использование AI имеет следствием *трансформацию источниковедения и вспомогательных исторических дисциплин* и формирование исследовательского инструментария, адекватного цифровой среде (цифровые палеография, дипломатика, сфрагистика, кодикология, филигранология, историческая лексикология и т.п.) [22, 23].

- непосредственного использования методов AI для анализа подготовленных источников, получения определенных результатов (п. 6) в рамках проблемно-исторических исследованиях для их дальнейшего синтеза, интерпретации и интеграции в исторический контекст на следующих этапах.

Очевидно, что на каждом из перечисленных этапов использование методов AI будет иметь свои ограничения, детерминированные спецификой источниковой базы и исследовательскими задачами исследований.

Завершая затянувшееся Введение, отметим, что дальнейшее изложение автор посвятит возможностям применения AI для развития источниковедения и вспомогательных исторических дисциплин, распознавания текстов и изображений в исторических источниках и их подготовке для проведения аналитических процедур. Т.о., **предлагаемый материал представляет собой обзор и обобщение подходов и проблем, имеющихся в мировой практике, и собственного опыта участия автора в некоторых уже осуществленных и осуществляемых в настоящее время проектах, целью которых было создание полностью распознанных (с точность не менее 95–97%) машиночитаемых текстов комплексов рукописных и машинописных источников для их представления в электронной среде и формирования специализированных наборов данных, которые могут являться как целью**

исследования, так и вспомогательным инструментарием в масштабных проектах распознавания. (Особо подчеркнем, что автор *не будет анализировать* проекты *целевого распознавания* отдельных элементов текста или распознаванием текста, которое осуществляется *в качестве вспомогательного этапа в проблемно-исторических исследованиях).*

Наборы данных, машинное обучение, AI и другие информационные технологии в автоматизированном распознавании текстов рукописных и машинописных источников (на примере распознавания текстов)

Проблема автоматизированного распознавания изображений (текстов вообще и рукописных текстов в частности) не нова. Впервые она возникла более 50 лет назад, в середине 1970-х гг., когда за рубежом стали проводиться научные конференции [24, 25, 26, 27, 28, 29], возникли профессиональные ассоциации [30, 31], объединившие специалистов-историков и информатиков, начали публиковаться тематические сериальные издания, сборники статей [32, 33], в которых излагают свои взгляды те, кто постоянно участвует в этих проектах, и те, кто анализирует их «издалека», на основе чужих публикаций, презентаций, ни разу не попробовав подготовить до́лжным образом электронную копию архивного документа, создать различные наборы данных (палеографические, кодикологические, дипломатические, историко-лексикографические и т.п.) для обучения нейросетей, привлечь знания экспертов, имеющиеся коллекции эталонов (бумаги и писчих материалов), разметить текст в соответствии с традициями дипломатики или с учетом разнообразных полиграфических формуляров документов и т.п.

К сожалению, последний подход наиболее распространен в отечественной историографии и чреват грустными последствиями, которые не только подрывают веру в возможности информационных технологий и сводят решение данной задачи исключительно к мозговым штурмам в виде хакатонов для поиска «чудо-алгоритмов», но и не замечают и не хотят замечать огромной подготовительной работы, которая осуществляется историками, архивистами, филологами, лингвистами и лежит в основе любого успешного проекта распознавания письменных источников.

Сразу отметим, что в большинстве проектов распознавания текстов (в т.ч. рукописных документов) успех делится в пропорции 70% на 30%, где 70% относятся к подготовке массивов исторических источников для перевода и переводу их в цифровой вид с помощью разнообразных информационных технологий, формированию вспомогательных наборов данных для машинного обучения, или лишь 30% – собственно на методы AI-распознавания, среди которых первое место занимают *нейросети*.

В целом, применение информационных технологий (в том числе AI) для автоматизированного распознавания текстов/изображений можно разделить на три группы:

- использование различных ИТ-методов для создания и улучшения качества (технических параметров) электронных копий, которые будут распознаваться (подготовительный этап работы, связанный с внешними особенностями подлинников, воспроизведенных на электронных копиях);

- применение ИТ (в том числе AI) для формирования специализированных наборов данных разного назначения и машинное обучение алгоритмов AI на подготовленных ds;

- собственно использование AI-алгоритмов для непосредственного распознавания.

В рамках этой статьи *кратко охарактеризуем основные виды подготовительных работ, используемые технологии, проблемы создания электронных копий (ЭК) и наборов данных*, без которых получение удовлетворительного результата распознавания текстов исторических источников затруднительно. (Как справедливо отмечал в своем докладе на Круглом столе, проходившем во ВНИИДАД 10.04.2023 г., директор Государственного архива Тульской области Д.Н. Антонов, ключевым требованием к осуществлению проектов распознавания является источниковедческий подход и связанные с ним методы научной критики источников, предназначенных для обработки [34]). Оставив в стороне рассмотрение эволюции программных средств собственно распознавания текстов – от программных приложений середины 1970-х – начала 1990-х гг., до движков и платформ, функционирующих с помощью различных типов нейросетей, упомянем только отдельные программные решения, которые используются для автоматизации создания ЭК, datasets, машинного обучения и распознавания.

1. *Создание и обработка электронных копий письменных источников,* которые могут быть использованы в проектах автоматизированного распознавания.

1.1. Разрешение, режим сканирования, формат

В отечественной историографии в последние годы утвердилось мнение о том, что для автоматизированного распознавания можно использовать любое электронное изображение письменного источника. Между тем, на основе опыта реализации большого количества проектов научного исследования рукописей, инкунабул и старопечатных книг [35], а также распознавания текстов еще на рубеже 2010-х гг. были определены оптимальные характеристики создания сканов: разрешение при сканировании должно быть не ниже 400 dpi для документов формате А4 (в идеале – 600 dpi), режим сканирования – «оттенки серого», формат сжатия файла Tiff (документы меньшего формата сканируются с бóльшим разрешением).

Эти параметры имеют четкое обоснование, связанное с:

- соотношением средней величины строчного символа (знака: буквы или цифры), размером бумаги и предъявлением этого соотношения на экране монитора при увеличении не меньше 200%, которое позволяет различать тонкие линии в начертании букв/цифр;

- основным принципом оптического распознавания, который базируется на *оценке (сравнении) степени яркости и контрастности пиксел, из которых состоит компьютерное изображение*[36]. Знаки (символы), написанные чернилами, отпечатанные на машинке, принтере, типографским способом и т.п., будут наиболее темными участками (группами пиксел) на изображении по сравнению с незаполненным полем бумаги. Учитывая то, что знаки (буквы/цифры) имеют разную толщину линий и «насыщенность чернилами» (контрастность) (даже при анализе машинописного текста), определение их «границ» и начертаний будет более точным в режиме оцифровки «оттенки серого», который лучше улавливает и отражает нюансы яркости и контрастности, позволяя определять начертание знака, «захватывая» самые светлые из темных пикселей, которые составляют элементы буквы/цифры (например, «отлетающие» росчерки, «хвостики» букв, части литер, написанные без нажима).

Очевидно, что использование цветного режима сканирования добавит «лишней»

информации в анализируемое изображение, поскольку в «состав» каждого из оттенков цветов RGB или CMYK, представляемого на экране монитора, в той или иной пропорции включены яркость и контрастность; а черно-белый режим, в который рекомендуют преобразовывать цветные изображения многие авторы, «обрезают» необходимую информацию о яркости и контрастности (процесс «бинаризации» используется *только в случае невозможности* создания целевого комплекса изображений). Таким образом оба режима искажают и огрубляют изображение и являются источником ошибок при распознавании.

Давать комментарий по формату Tiff представляется излишним, отметим только, что этот формат сохранят изображение практически без искажений, что чрезвычайно важно при распознавании.

1.2. Необходимость учета текстуры писчих материалов, способов фиксации текстовой и изобразительной информации и степени сохранности распознаваемых документов

Перечисленные вопросы, как правило, относят к предметной области работы реставраторов и не принимают в расчет при создании электронных копий и разработке систем распознавания. Однако, как показывает опыт, непонимание физических особенностей и механизмов создания документов приводит к ошибкам при выборе инструментов сканирования и распознавания. Рассмотрим несколько практических примеров, проанализировав носители письменной информации (глиняные таблички, поверхность камня, папирус, пергамент, бомбицину, бумагу, кальку и т.п.).

Каждый из перечисленных носителей информации обладает собственной спецификой. Однако самым распространенным и «недооцененным» с точки зрения проблем носителем, является, безусловно, *бумага.* Многие ошибки при распознавании рукописных текстов, созданных до начала XIX вв. (в России до конца первой трети XIX вв.), связаны с тем, что тряпичная бумага, использовавшаяся в то время, имеет неровную «кочковатую» поверхность, обусловленную ручным способом измельчения сырья и «отлива» листов бумаги. Такая поверхность бумаги по-разному впитывает чернила на разных частях одного и того же листа, что влияет на начертание букв (они могут расплываться, их размеры и границы «плывут», чернила впитываются и проходят насквозь, создавая на обратной стороне листе «мусор» и т.п.), а яркость и контрастность на изображении становятся менее определенными.

Аналогичные проблемы могут возникать в случае использования бумаги, отлитой в маленьких мануфактурах, где в сетчатых формах для литья бумажных листов в качестве вержеров и понтюзо, в также для создания рисунка филиграни [37], использовалась довольно толстая проволока, оставляющая в готовом листе менее плотные бороздки бумаги.

Изобретение бумагоделательной машины Робера (1799 г.) и ее внедрение в мануфактурное производство улучшило качество бумаги [38, 39], и бумага документов XIX в. уже не создает таких проблем при распознавании.

Однако, существенной проблемой для автоматизированного распознавания по-прежнему остается *плотность* самой бумаги. Особую сложность составляют документы, написанные на обеих сторонах неплотных листов (плотность менее 60 г/м2), на кальке (плотность менее 40 г/м2), на бязи, созданные с помощью копировальной бумаги (2 и последующие экземпляры) и т.п.

Методы борьбы с «кочковатостью» и просвечивающими с обратной стороны листа

строками, проступающими насквозь чернилами или рельефами (особенно знаками препинания), т.н. «артефактами», «мусором» и «шумом» на электронном изображении начинают применять не в процессе распознавания текста (поскольку никакая графическая обработка или разметка строк/текста на листе не способны компенсировать эти недостатки), а еще *на этапе сканирования документов с подбора соответствующих по цвету прокладочных листов*, использование которых позволяет избавиться от 85–90% «артефактов» и тем самым подготовить удовлетворительную по качеству для целей распознавания электронную копию.

В контексте анализа особенностей носителей письменных источников, оказывающих прямое воздействие на качество автоматизированного распознавания, следует также назвать *технику фиксации знаков (символов)*. Очевидно, что надписи на камне (эпиграфические источники), на глиняных табличках, бересте, пергаменте и т.п., – рельефные, и утрата красочного слоя (угасание текста) не оказывает существенного влияния на распознавание. Рельефными являются также и надписи на бумаге, сделанные пером, остро заточенным карандашом и пишущей машинкой. Эти рельефы («трассы») относятся к «низким», и для их выявления необходимо использование специально сконструированного сканирующего оборудования с последующей графической обработкой [40]. Разработанная технология, наравне с применением специализированных методов мультиспектральной [41, 42] и гиперспектральной фотосъемки и анализа, различных вариантов спектроскопий и др. с успехом используются для подготовки электронных копий угасших текстов рукописей и манускриптов для распознавания, а также выявления уничтоженных записей на пергаменте (палимпсестах).

Единственной техникой, которая на сегодняшний день составляет неразрешимую проблему при сканировании и распознавании угасающих текстов, является *факсимильная (факсовая) передача информации*. Это проблема обусловлена механизмом создания текста или изображения на факсовой термобумаге, при котором используется нагрев (химическая реакция – плавление) красителей, создающих текст/изображение, а физическое воздействие на носитель (бумагу) отсутствует, в результате чего рельеф не возникает.

Резюмируя рассмотрение этапа сканирования, еще раз подчеркнем, что игнорирование текстуры носителя и некачественная подготовка массивов электронных копий, предназначенных для автоматизированного распознавания, является причиной большого количества проблем, которые не могут быть решены программными средствами на следующих этапах исследования.

2. Наборы данных, их разновидности, специфика создания

Создание наборов данных для машинного обучения всех разновидностей программ автоматизированного распознавания и само машинное обучение – второй и третий по важности и наиболее продолжительные этапы подготовительных работ, которые могут длиться от нескольких месяцев до нескольких лет. Как правило, реализация этих этапов в последние годы является наиболее «закрытой» и редко афишируемой частью проектов, детерминированной особенностями комплекса распознаваемых источников и задачами, стоящими перед исследователями, и основанной на экспертном знании не только историков, источниковедов, палеографов и специалистов в области других вспомогательных исторических дисциплин, но и реставраторов и информатиков. Подчеркнем, что различные информационные методы (в том числе машинное обучение и AI) находят свое применение и на этих подготовительных этапах формирования наборов

данных[43].

Кратко охарактеризуем разновидности наборов данных, которые необходимо подготовить для машинного обучения систем распознавания письменных исторических источников.

Палеографические наборы данных

В 2000-е – середине 2010-х гг. создание палеографических наборов данных и их публикация в сети интернет были одним из самых популярных направлений в профессиональных исторических исследованиях. В этот период были созданы, например, один из первых палеографических наборов данных для машинного обучения – MNIST [44], в котором были аккумулированы варианты рукописного написания арабских цифр (за последние 20 лет ds неоднократно обновлялся), так называемая «Средневековая палеографическая шкала» (набор данных) для датировки исторических документов Нидерландов и Фландрии периода 1300–1550 гг. с интервалом в 25 лет[45], ставшая моделью для создания аналогичных наборов во многих европейских странах и разработки подходов к классификации почерков [46], создавались как персонифицированные наборы данных почерков отдельных людей, так и типологические палеографические модели для разных европейских и азиатских языков и хронологических периодов [47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59 60]. Созданные наборы данных стали основой для развития онлайн-платформ автоматизированного распознавания текстов типа Transcribus, eScriptorium, Tesseract и др. (в Европе), Hentaigana [61], KuLA (くずし字学習支援アプリ) [62], MOJIZO (解析：木簡・くずし字解読システム)[63] (в Японии) [64].

Цель формирования палеографических наборов данных очевидна – они служат обучающим материалом для алгоритмов автоматизированного распознавания средневековых текстов и текстов Нового времени (т.е. носят *«инструментальный характер»*), а также используются *в качестве метода исследования* для реализации ***проектов изучения конкретных архивных коллекций и рукописей исторических персон*** [65, 66]***, формирования коллекции автографов*** [67] (в том числе, например, установления/подтверждения авторства [68, 69, 70, 71, 72], выявления подделок рукописей [73, 74], уточнения датировок [75, 76] и т.п. Для этих целей создаются наборы данных, характеризующие эволюцию почерка конкретного лица *в течение всей его жизни*). Такие наборы данных широко применяются в зарубежной историографии, использовались они и в России (в частности – KHATT [77], доступный бесплатно) [78] и т.п.

В настоящее время существуют «типовые» палеографические ds для латиницы, греческого и арабского алфавитов, западноевропейской кириллицы, еврейского, армянского, грузинского и др. алфавитов. Эти наборы развиваются и дорабатываются с учетом специфики национальных языков и конкретных исторических периодов [79, 78]. Вместе с тем, к сожалению, следует констатировать, что в силу трудоемкости процесса создания палеографических наборов данных на сегодняшний день ни один язык не имеет полного комплекта палеографических моделей для всего периода существования национальной письменности, что затрудняет осуществление работ по тотальному распознаванию текстовых источников и их представлению в машиночитаемом виде. Сложившаяся ситуация привела к возникновению нового тренда – разработке искусственно созданных наборов данных на основе генеративного AI [81], возможности широкого внедрения которого еще требуют изучения.

В зарубежных университетах, архивах и библиотеках с начала 2000-х гг. создавались и продолжают развиваться ***проекты публикации коллекций электронных копий рукописей*** и созданных на их основе палеографических/изобразительных описанных (аннотированных, размеченных) ***наборов данных*** – шрифтов, почерков [82, 83], иллюстраций. Многие из этих работ представлены на обобщающих порталах или сайтах проектов [84, 85, 86], осуществляются в виде баз данных, онлайн-каталогов [87], связанных открытых данных [88, 89], интернет-платформ и онлайн-учебников/курсов для обучения студентов [90, 91], моделей палеографических наборов «изображение-текст»[92], публикаций ds на открытых площадках, создания наборов данных для разработки специальных инструментов работы с рукописями[93, 94, 95, 96], для расшифровки стенографических сокращений («Тиронские примечания» [97]), аббревиатур и т.п.

В странах Европы и Дальнего Востока (КНР, Япония и др.) для формирования баз знаний (наборов данных) специфических шрифтов и восточных идеографических языков активно разрабатываются собственные модели автоматизированного распознавания, например, созданная в Германии OCR-система для чтения готического шрифта Fractur[98] или система RURI (瑠璃)[99] в Японии. Последняя основана на применении Международной платформы обмена изображениями IIIF и глубоком обучении. На порталах и сайтах крупных проектов размещаются наборы данных, отражающие палеографические особенности иероглифики на разных носителях (шелке, бумаге, бамбуковых планках, бронзовых пластинах, керамике и т.п.) и разных каллиграфических стилей [100, 101, 102]. В качестве дополнения и расширения возможностей созданных ресурсов [103, 104], где опубликованы ds, на сайтах размещаются также лексикографические (тематические) наборы данных, облегчающие процесс распознавания.

К сожалению, в России работы по созданию палеографических систем и наборов данных практически не ведутся. В составе единственного близкого по тематике проекта «История письма европейской цивилизации» [105], разработанного в Санкт-Петербургском институте истории РАН (URL: https://gis.spbiiran.ru/), имеется коллекция оцифрованных исторических источников, однако, предложенное описание каждого документа и крайне низкое качество изображений не позволяют сформировать палеографические наборы.

Перечислив некоторые проекты формирования палеографических ds, остановимся на проблемах, возникающих при их создании. Так, серьезной проблемой является качество писчих материалов (чернила, краски, тушь, карандашные грифели, типографская краска, красящие наполнители ленты для п/машин, картриджи принтеров и т.п.), их яркость и контрастность на электронных копиях и т.п., которые напрямую зависят от содержания в них черного цвета [106].

Выцветание железо-галловых чернил, осыпь графитных карандашей, тусклость цветных красителей лент пишущих машинок, картриджей принтеров и т.п., т.е. так называемое «угасание текста», является одной из причин использования режима сканирования «оттенки серого», использования методов мультиспектральной фотосъемки [107] и анализа изображений и/или разработки методов графической обработки [108] электронных копий в таких графических редакторах как Adobe PhotoShop или Irfan. Автору известно, как минимум, несколько разработанных решений для этих программных приложений на основе сформированных наборов данных, созданных для разных типов

текстов. К сожалению, подобные разработки всегда носят исключительно источнико-ориентированный характер, поскольку тесно связаны с конкретными комплексами документов и степенью их сохранности, являются «техническими» аспектами проектов распознавания, и, как правило, даже не удостаиваются упоминания в статьях.

Особого рассмотрения заслуживает вопрос о писчих принадлежностях авторов письменных источников (калам, птичье и железное перо, перьевая, шариковая ручка, фломастер, пишущая машинка и т.п.), особенностях почерка писца и/или способах создания письменных источников.

К примеру, европейские писец-левша и писец-правша будут писать буквы с разным нажимом, т.е. у букв, написанных левшой, наиболее наполненная чернилами (т.е. наиболее темная, яркая и контрастная на электронном изображении) часть букв будет располагаться справа, а у правши – слева. Учитывая основной принцип, на котором базируется распознавание, упомянутый выше, эта «зеркальность» будет создавать проблемы «идентификации» букв, и требовать создания специализированных наборов данных и дообучения программ HTR. (Проблема «нажима» практически исчезла в связи с широким использованием шариковых, гелевых ручек и фломастеров).

Изобретенные в конце XIX в. и остававшиеся самым популярным средством создания официальных документов в течение бóльшей части XX в. *механические пишущие машинки*, безусловно, улучшили «человеко-читаемость» письменных источников, но при этом создали новые специфические препятствия для автоматизированного распознавания текстов. Так, четкость (яркость и контрастность) воспроизведения символов в документах стали напрямую зависеть от «свежести» печатной ленты, чистоты рельефных букв (литер) на шрифтовых колодках, индивидуальной силы удара каждого из пальцев машинистки (при десятипальцевом методе печатанья; сила удара перестала быть определяющей во второй половине XX в. после изобретения электрических пишмашинок), плотности бумаги и использования копирки для создания нескольких экземпляров документа. К этому следует добавить, что каждый экземпляр пишущей машинки оснащался набором литер, имевшим практически незаметное, но при этом уникальное отличие в оттиске, что позволяло однозначно идентифицировать пишущее средство, но в настоящее время представляет дополнительную проблему при создании палеографических наборов данных для распознавания [109]. Совокупность этих особенностей создания машинописных документов является основной причиной того, что многие программные средства распознавания хуже распознают документы XX в., нежели рукописные источники.

Кодикологические и дипломатические наборы данных

Обязательные наборы данных, формируемые в целях автоматизированного распознавания текстов рукописных документов и старопечатных книжных памятников, связаны с кодикологией и дипломатикой.

Шрифты рукописных и печатных книг (начертание и размер), заставки и буквицы, орнаментика и иллюстрации, кустоды, клейма печатников (инициалы, имена и цифры), пропорции самого издания и наборной полосы, конфигурация текстового набора, размеры и соотношение полей на странице, количество строк на странице, формат строки (выключка), концовки; межбуквенные, межсловные и междустрочные интервалы (количество строк на странице, количество букв в строке); инструменты зрительного деления текста (абзацные отступы, втяжки, отбивки, шрифтовые и цветовые выделения, элементы рубрикации, маргиналии); колонцифры, колонтитулы, колонлинейки, фолиация,

пагинация и т.п. – все эти элементы являются объектами для формирования баз знаний (наборов данных [110, 111]), помогающих не только атрибутировать рукописи и книжные памятники и подтверждать их подлинность, но и проводить автоматизированную разметку, выделение фрагментов текста, строк [112, 113], установление участия писцов в создании рукописи [114], уточнение датировок письменных источников [115]. К настоящему времени в России и за рубежом реализовано несколько проектов [116], в которых активно использовались созданные кодикологические наборы данных и на их основе разрабатывались специальные программные приложения [117, 118].

Не менее важными являются и дипломатические datasets, позволяющие анализировать «шаблоны» (формуляры и формулы (структурные части текста) – объекты изучения формуляроведения – направления в цифровой дипломатике [119]) документов (например, с помощью технологии «изображение в изображении»[120, 121]), и тем самым облегчающие автоматизированную сегментацию текста [122, 123]. Подобные разработки особенно актуальны для таких видов документов как грамоты, акты, хартии, реестры и т.п. [124], текстов, созданных в средневековье [125], написанных на бланках, имеющих созданные полиграфическим способом графические элементы (таблицы, угловые штампы, линии, схемы и т.п.), которые мешают распознаванию собственно текстов. До недавнего времени для обработки таких источников создавались специализированные программные средства на основе т.н. «компьютерного зрения», которые «снимали» лишние элементы с изображения и оставляли только текст. В настоящее время для решения этой проблемы используются рекуррентные (RNN)[126], сверточные (CNN)[127] и иные виды нейронных сетей [128], обучение которых ведется на основе созданных дипломатических наборов данных[129], аккумулирующих и описывающих варианты оформления документов.

Необходимо подчеркнуть, что создание дипломатических ds актуально для работы с документацией, созданной до середины – второй половины XX в., когда во многих странах (и СССР в том числе) были разработаны и введены в действие стандарты оформления управленческой документации, что в совокупности с однозначно установленными и используемыми форматами бумаги и техническими средствами создания документов унифицировали документацию.

Особого упоминания заслуживают наборы данных и технологии, применяемые для сегментации и распознавания текстов и изображений на газетных полосах [130], картографических источниках [131, 132, 133, 134, 135, 136, 137, 138], чертежах и схемах в научно-технической документации [139, 140, 141], в научных статьях [142, 143] и т.п., которые всегда составляли проблемы для распознавания. Активное сканирование и публикация подобных источников позволили специалистам создать необходимые наборы данных, а развитие нейросетей в последнее десятилетие – приступить к решению задач распознавания подобных источников.

Вместе с тем, одной из самых сложных разновидностей документов для автоматизированного распознавания являются заполненные бланки разнообразных анкет (в т.ч. первичные листы переписей населения), налоговых деклараций и иной массовой документации середины XIX – третьей четверти XX вв., сочетающие в себе графические элементы, рукописные тексты и неоднозначно трактуемые символы, например, отметки в логических полях, в которых может стоять любой знак. Разрабатываемые с целью получения статистических данных, агрегирования информации

для последующего анализа, эти формуляры очень удобны для математической обработки и крайне сложны и капиталоемки для разработки систем автоматизированного распознавания. К примеру, в 2022 г. во Франции стартовал проект Socface (URL: https://socface.site.ined.fr/), целью которого является создание базы данных всех людей, живших во Франции в период с 1836 по 1936 гг. В качестве источниковой базы используются электронные копии именных списков 20 переписей населения всех департаментов Франции, которые должны быть распознаны в автоматизированном режиме с помощью AI. К сожалению, на сегодняшний день в части распознавания текстов в проекте особых успехов не наблюдается. Учитывая сложность работы с массовым анкетным материалом, с целью ввода информации большинство крупных архивов используют волонтеров и краудсорсинговые платформы [144].

Историко-лексикологические наборы данных – словари

Лексикологические наборы данных (словари) аккумулируют информацию об именах собственных, названиях географических объектов (ономастика и топонимика), учреждений, аббревиатурах, терминологии предметных областей и т.п. и их эволюции, отраженной в текстах документов. К сожалению, для формирования подобных ресурсов невозможно использовать разработанные и введенные в действие стандарты ISO, поскольку словари детерминированы содержанием исторических источников и той исторической реальностью, в которой документы созданы, поэтому для каждого комплекса источников подобные datasets формируются специально.

Существенные проблемы при создании историко-лексикологических ds составляют, например:

- различия в именованиях людей, явлений, процессов в разные исторические периоды;

- разный состав элементов имен у разных народов;

- омонимия – совпадение имен, названий, терминов;

- терминологические неоднозначности, обусловленные полисемией (многозначностью слов, имен);

- грамматические ошибки или специфика написания имен и терминов, обусловленная временем, местом, орфографией и т.п.,

- языковый меланж (смешение, соединение слов из разных языков в одном фрагменте текста);

- а также большое разнообразие написания одного и того же имени, названия, термина, затрудняющее его точную идентификацию и т.п.

Очевидно, что преодоление этих сложностей возможно только при проведении развернутого исторического, источниковедческого, палеографического и текстологического анализа, обеспечивающего однозначную идентификацию конкретной позиции в наборе данных и, одновременно, ее максимально полное описание [145]. Определенную помощь в создании подобных ds могут оказать ведущиеся традиционно в архивах именные и предметные указатели, однако, как показывает опыт, механический перенос таких указателей в электронную среду не дает искомого результата, их нужно перерабатывать и готовить специальным образом [146].

Фактически, «словарные» наборы данных представляют собой элементы *исторической*

лексикологии, целью которой является описание исторического метаязыка, отражающего эволюцию словарного состава языка (языков), на котором написаны документы, и являются своего рода переходным этапом от распознавания текстов источников к их семантическому анализу (в т.ч. с помощью больших языковых моделей). К сожалению, формирование исторических метаязыков даже в европейских странах далеко от завершения, поэтому использование LLM, созданных на основе современной лексики, для изучения исторических проблем и вопросов малоэффективно [147]. (Вариантом выхода из тупика может быть использование небольших языковых моделей, хотя понятие «небольшой» постоянно меняется: Phi-3 и 4 от Microsoft, Llama-3.2 1B и 3B, Qwen2-VL-2B, DeepSeek и др.).

3. Этап машинного обучения

На этом этапе проверяется точность созданных наборов данных, их репрезентативность и, как следствие, применимость для использования на всем массиве источников, которые подготовлены для распознавания.

Обычно созданные на предыдущих этапах ds делятся на три части:

- часть, предназначенную для использования в процессе обучения модели;

- часть, которая используется для верификации различных параметров и настроек модели с целью определения необходимости доработки набора и дообучения алгоритмов распознавания (т.н. настроек модели);

- часть, для тестирования окончательной версии обученной модели на тестовом массиве источников.

В историографии утвердилось мнение, что чем больше наборы данных, тем они успешнее справляются с поставленными задачами. Между тем, крупнейший авторитет в области машинного обучения профессор Стенфордского университета Эндрю НГ (URL: https://www.andrewng.org/) справедливо отмечал, что значительно важнее *качество данных и ориентация AI не на модели, а на данные* («во многих отраслях, где гигантских наборов данных просто не существует, думаю, акцент нужно сместить с больших данных на качественные данные. Наличия 50 хорошо продуманных и проработанных примеров может быть достаточно, чтобы объяснить нейронной сети то, чему вы хотите её научить»[148]). Эта точка зрения в настоящее время находит практическое подтверждение в работах по распознаванию текстов древних и средневековых рукописей, в эпиграфике и т.п., а созданные размеченные ds становятся базовыми для осуществления разнообразных исследований и многократного использования.

Вместо заключения

Задача автоматизированного распознавания текстов, изображений, аудио и т.п., как одно из ключевых направлений применения искусственного интеллекта, в последние 70 лет находится в фокусе внимания специалистов многих профессий, однако до окончательного ее решения еще очень далеко.

Очевидным фактом является то, что применение любых информационных технологий в исторических исследованиях, в т.ч. OCR, HTR, AI вообще и нейросетей в частности, базируется на экспертном знании историков, источниковедов, специалистов по ВИД, реставраторов и информатиков. Созданные ими различные типы и варианты наборов данных, разработанные инструменты использовались и используются не только в

проектах автоматизированного распознавания текстов письменных источников, но обладают и собственной информационной ценностью, поскольку их применение не ограничивается только «инструментальной» (вспомогательной) ролью при переводе рукописного текста в машиночитаемый вид, но и «работают» на развитие архивной эвристики и на реализацию аналитического этапа конкретно-исторических исследований [149]. Результаты этой подготовительной работы являются «приращением исторического знания» и основанием для активного развития исторической науки, ее выводу на новый виток развития, соответствующий нынешнему уровню информационной эпохи [150].

К сожалению, автор вынужден констатировать, что отечественная историческая наука (за редким исключением), сосредоточившись на проблемно-ориентированных исследованиях, пропустила исторический временной период (середина 1990-х – начало 2010-х гг.), удобный для создания наборов данных, и теперь находится в догоняющем положении. Исправление этой ситуации возможно при увеличении внимания к источниковедению и архивоведению, внедрению в эти исторические дисциплины новых подходов и методов, о чем говорилось в выступлении академика-секретаря, руководителя секции истории Отделения историко-филологических наук Российской академии наук Е.И. Пивовара на III Петербургском историческом форуме в октябре 2024 г.

Примечания

1. Эта программа проиграла матч аналогичной советской программе для компьютера М-2, разработанной в лаборатории Московского института теоретической и экспериментальной физики (МИТЭФ) (рук. лаборатории А. Кронрод).

2. Первый чат-бот ELIZA был разработан в середине 1960-х гг. Д. Вайценбаумом. Бот мог общаться с человеком на естественном языке, имитируя работу психотерапевта. Долгое время считалось, что ELIZA утрачена, однако на основе сохранившихся распечаток кода Вайценбума чат-бот был восстановлен и представлен онлайн: ELIZA Archaeology – Try ELIZA // URL: https://sites.google.com/view/elizaarchaeology/try-eliza

3. В 1970-х – начале 1990-х гг. в отечественной исторической науке были осуществлены исследования, в которых в той или иной степени использовались методы и технологии AI. Среди этих работ необходимо упомянуть исследования группы математиков под руководством академика Н.Н. Моисеева (в частности – имитационное моделирование: модель Синопского сражения, моделирование процессов экономической динамики греческих полисов периода Пелопонесской войны V в. до н.э. [11, 12]), Лукова В.Б. и Сергеева В.М. (построение модели восприятия ситуации и принятия решения историческим деятелем на основе контент-анализа мемуаров Отто фон Бисмарка [14]), методы контрфактического моделирования развития экономики (монография Ю.П. Бокарева [15]), систему «Ретропрогноз» и экспертные системы АМСОР[16] и «ГИДРОНИМИКОН» [17] и др.

К сожалению, следует сказать, что в силу различных причин история развития и применения AI в СССР вообще и в исторической науке в частности менее известна и изучена, чем аналогичная зарубежная тематика.

4. Учитывая то, что историческая наука многогранна и разностороння, «приращение научного знания» может означать не только выявление и новое осмысление исторических фактов, событий, явлений, процессов и участия в них людей, но и решение теоретических и прикладных задач источниковедения, историографии, эвристики и иных

вспомогательных исторических дисциплин.

5. Одной из самых значимых отечественных работ в области применения AI в архивной эвристике является комплекс систем Искусственного интеллекта, который разрабатывается в ГА РФ и был представлен в докладе А.А. Колганова «Эволюция применения искусственного интеллекта в ГА РФ: 2021–2024 гг.» на XIX Конференции Ассоциации «История и компьютер» 15 ноября 2024 г. [21].

Библиография

1. Minsky M. A Neural-Analogue Calculator Based upon a Probability Model of Reinforcement. Harvard University Psychological Laboratories. Cambridge, Massachusetts. January 8, 1952 // Selected Publications of Marvin Minsky. URL: https://www.mit.edu/~dxh/marvin/web.media.mit.edu/~minsky/Bibliography.html
2. The Dartmouth AI archives // Ray Solomonoff's Home Page. URL: https://raysolomonoff.com/dartmouth/dart.html
3. Newell A., Simon H. A. The Logic Theory Machine. A complex information processing system. 12 July 1956. // RAND Corporation. 1956. Архивная копия от 17 октября 2014 на Wayback Machine. URL: https://archive.org/details/bitsavers_randiplP86ineJul56_3534001/mode/2up
4. John McCarthy's Home Page // URL: https://www-formal.stanford.edu/jmc//
5. Rosenblatt F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain // Psychological Review. November, 1958. Vol. 65. Pp. 386-408. Lancaster, PA and Washington, DC: American Psychological Association, 1958. Архивная копия от 17 октября 2014 на Wayback Machine. URL: https://web.archive.org/web/20080218153928/http://www.manhattanrarebooks-science.com/rosenblatt.htm
6. Samuel A. L. Some studies in machine learning using the game of checkers // IBM Journal of Research and Development. Jan. 2000. Vol. 44. No. 1.2. Pp. 206-226. DOI: 10.1147/rd.441.0206.
7. McCarthy J. Recursive functions of symbolic expressions and their computation by machine // Communications of the ACM. April 1960. Архивная копия от 17 октября 2014 на Wayback Machine. URL: https://web.archive.org/web/20131006003734/http://www-formal.stanford.edu/jmc/recursive.html
8. A chess playing program for the IBM 7090 computer // URL: https://dspace.mit.edu/handle/1721.1/17406
9. Killgrove K. "ELIZA", the world's 1st chatbot, was just resurrected from 60-year-old computer code // Live Science. 18 Jan. 2025. URL: https://www.livescience.com/technology/eliza-the-worlds-1st-chatbot-was-just-resurrected-from-60-year-old-computer-code
10. Lane R., Hay A., Schwarz A., Berry D. M., Shrager J. ELIZA Reanimated: The world's first chatbot restored on the world's first time sharing system // 12 Jan. 2025. URL: https://arxiv.org/abs/2501.06707
11. Моисеев Н. Н. Математика ставит эксперимент. М.: Наука, 1979. 223 с.
12. Гусейнова А. С., Павловский Ю. Н., Устинов В. А. Опыт имитационного моделирования исторического процесса // Под ред. и со вступ. ст. Н. Н. Моисеева. М.: Наука, 1984. 157 с.
13. Когнитивные методы за рубежом. Методы Искусственного Интеллекта в моделировании политического мышления. [Сб. ст.] / АН СССР, Ин-т США и Канады; [Отв. ред. В. М. Сергеев]. М.: Ин-т США и Канады, 1990. 148 с.
14. Луков В. Б., Сергеев В. М. Опыт моделирования мышления исторических деятелей:

Отто Фон Бисмарк, 1866–1876 гг. // Вопросы кибернетики. Логика рассуждений и её моделирование. [Сб. статей] / Под ред. Поспелова Д. А. М.: Науч. совет по комплекс. пробл. "Кибернетика" АН СССР, 1983. С. 149-172.
15. Бокарёв Ю. П. Социалистическая промышленность и мелкое крестьянское хозяйство в СССР в 20-е годы: источники, методы исследования, этапы взаимоотношений / Отв. ред. И. Д. Ковальченко; АН СССР, Ин-т истории СССР. М.: Наука, 1989. С. 148-166.
16. Бородкин Л. И. Что сделали ЭВМ для исторической науки // Арзамас. URL: https://arzamas.academy/materials/2284
17. Храмов Ю. Е. ГИДРОНИМИКОН – экспертная система по гидронимии Восточно-Европейской равнины // Информационный Бюллетень Комиссии по применению математических методов и ЭВМ в исторических исследованиях. 1992. № 5.
18. Kismet // 17 Oct. 2000. Архивная копия от 17 октября 2014 на Wayback Machine. URL: http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html
19. Kaplan A., Haenlein M. "Siri, Siri in my Hand, who's the Fairest in the Land?" On the Interpretations, Illustrations and Implications of Artificial Intelligence // Business Horizons. 2018. Vol. 62. Pp. 15-25. DOI: 10.1016/j.bushor.2018.08.004
20. ГОСТ Р 59895-2021 Технологии искусственного интеллекта в образовании. Общие положения и терминология // М.: ФГБУ "РСТ", 2021.
21. Колганов А. А. Эволюция применения искусственного интеллекта в Государственном Архиве РФ (2021–2024 годы) // Информационный бюллетень Ассоциации "История и компьютер". № 51, специальный выпуск, ноябрь 2024 г. Материалы международной научной конференции "Современная историческая информатика: Аналитика данных в исторических исследованиях" и XIX конференции Ассоциации "История и компьютер". Москва, 15-17 ноября 2024 г. М., 2024. С. 7. [Электронное издание].
22. Юмашева Ю. Ю. Цифровая трансформация вспомогательных исторических дисциплин. Современные неинвазивные методы изучения исторических артефактов [Видеолекция] // Международная летняя школа молодых ученых "Историческая информатика – 2022". 15.07.2022. URL: https://www.youtube.com/watch?v=jWUw8fWMcqw
23. Юмашева Ю. Ю. Цифровая трансформация вспомогательных исторических дисциплин [Видеолекция] // Международная летняя школа молодых ученых "Историческая информатика – 2023". 30.06.2023. URL: https://www.youtube.com/watch?v=4HQezjps7ig
24. International Conference on Document Analysis and Recognition (ICDAR) // URL: https://www.icdar.org/; http://www.iapr-tc11.org/mediawiki/index.php/Conferences
25. International Conference on Frontiers in Handwriting Recognition (ICFHR) // URL: http://www.iapr-tc11.org/mediawiki/index.php/Conferences
26. International Conference on Pattern Recognition Systems (ICPRS) // URL: https://www.icprs.org/
27. International Conference on Pattern Recognition and Artificial Intelligence (IEEE PRAI) // URL: https://www.prai.net/
28. Artificial Intelligence and Pattern Recognition (AIPR) // URL: https://www.aipr.net/
29. Japan-International Conference on Machine Learning and Pattern Recognition // URL: https://www.mlpr.org/
30. International Association for Pattern Recognition // URL: https://iapr.org/
31. History of IAPR // International Association for Pattern Recognition. URL: https://iapr.org/about-us/history-of-iapr/
32. IAPR Newsletter // International Association for Pattern Recognition. URL: https://iapr.org/articles/newsletter/
33. International Journal on Document Analysis and Recognition (IJDAR) // Springer-Verlag GmbH Germany. URL: https://www.springer.com/journal/10032/
34. Антонов Д. Н. Источниковедческие подходы к формированию базы данных

метрических книг с целью оптического распознавания рукописного текста: Круглый стол "Практические задачи внедрения технологий ИИ в деятельность архивов" от 10 апреля 2023 г. // YouTube канал ВНИИДАД. М., 2023. URL: https://www.youtube.com/watch?v=KHzhpS42vqk&t=12179s
35. Шабанов А. В. Факторы, влияющие на выбор технологии оцифровки русских старопечатных и рукописных книг // Библиосфера. 2008. № 4. С. 46-48.
36. Impedovo S. Fundamentals in Handwriting Recognition // North Atlantic Treaty Organization. Scientific Affairs Division. NATO Advanced Study Institute on Fundamentals in Handwriting Recognition (NATO ASI Series). Berlin: Springer-Verlag, 1994. URL: https://link.springer.com/book/10.1007/978-3-642-78646-4
37. The memory of paper // URL: https://memoryofpaper.eu/BernsteinPortal/appl_start.disp
38. Муратова А., Гудков А. Бумага и бумажное производство в средние века и ранее новое время // Рукописная книга: традиция и современность. URL: https://manuscriptcraft.com/article_11
39. Есипова В. А. Бумага как исторический источник (по материалам Западной Сибири XVII-XVIII вв.). / Под ред. А. Н. Жеравиной. Томск: Изд-во Том. ун-та, 2003. 290 с.
40. ARCHiOx: seeing the unseen. Digitising objects in 3D will give more than the ability to zoom in and examine historical objects in detail // URL: https://oxford.shorthandstories.com/digital-archiox/index.html?fbclid=IwAR2LM19j6iFh1NUgEBddBmU0oZotufAEEs8G0vn2FzF97_dFd2c-TUUwGBs
41. Brown N. Collection Care welcomes a new multispectral imaging system // UK National Archives Blog, 2019. URL: https://blog.nationalarchives.gov.uk/collection-care-welcomes-a-new-multispectral-imaging-system/
42. Миклас Х., Бреннер С., Саблатниг Р. Мультиспектральная съемка для цифровой реставрации древних рукописей: устройства, методы и практические аспекты // Историческая информатика. 2017. № 3. С.116-134. DOI: 10.7256/2585-7797.2017.3.23697 URL: https://nbpublish.com/library_read_article.php?id=23697
43. Sánchez-DelaCruz E., Loeza-Mejía C. I. Importance and challenges of handwriting recognition with the implementation of machine learning techniques: a survey // Applied Intelligence. The International Journal of Research on Intelligent Systems for Real Life Complex Problems. 2024. Vol. 54. Pp. 6444-6465. DOI: 10.1007/s10489-024-05487-x
44. MNIST // Modified National Institute of Standards and Technology. URL: http://yann.lecun.com/exdb/mnist/; https://docs.ultralytics.com/ru/datasets/classify/mnist/
45. MPS – Medieval Paleographic Scale – The University of Groningen research portal // URL: https://research.rug.nl/en/datasets/mps-medieval-paleographic-scale
46. Житинева А. М. Палеография и эпиграфика: две дисциплины или одна? (К вопросу о палеографической классификации письменных источников X–XVII вв.) // URL: https://spbiiran.ru/paleografiya-i-epigrafika-dve-discziplinу-ili-odna-k-voprosu-o-paleograficheskoj-klassifikaczii-pismennyh-istochnikov-x-xvii-vv-doklad-a-m-zhitenevoj-na-zasedanii-drevneruss/
47. Leuven Database of Ancient Books // Portal Trismegistos. URL: https://www.trismegistos.org/ldab/
48. Papyri.info // URL: https://papyri.info/
49. Kölner Papyri (Fayum papyri) // URL: https://papyri.uni-koeln.de/
50. Stutzmann D. Dated and Datable Manuscripts: dataset // 2022. DOI: 10.5281/zenodo.6507965.
51. Clélice T. et al. CATMuS Medieval: A Multilingual Large-Scale Cross-Century Dataset in Latin Script for Handwritten Text Recognition and Beyond // Lecture Notes in Computer Science. 2024. Pp. 174-194. DOI: 10.1007/978-3-031-70543-4_11
52. DigiPal // URL: http://www.digipal.eu

53. Italian Paleography // URL: https://italian.newberry.t-pen.org/
54. DIVAHisDB Dataset of Medieval Manuscripts // University of Fribourg. URL: https://www.unifr.ch/inf/diva/en/research/software-data/diva-hisdb.html
55. HisDoc III Digital Analysis of Syriac Handwriting (DASH) // URL: http://dash.stanford.edu/
56. Fischer A., Bunke H., Naji N., Savoy J., Baechler M., Ingold R. The HisDoc Project. Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries. // In: Internationalität und Interdisziplinarität der Editionswissenschaft. DOI: 10.1515/9783110367317.91
57. French Renaissance. Paleography // URL: https://french.newberry.t-pen.org/
58. France-England: medieval manuscripts between 700 and 1200 // URL: https://manuscrits-france-angleterre.org/polonsky/en/content/accueil-en?mode=desktop
59. Scottish Handwriting // Scotland's People URL: https://www.scotlandspeople.gov.uk/scottish-handwriting
60. Al-Furqan's E-Database // Al-Furqan Islamic Heritage Foundation. URL: Al-Furqan Islamic Heritage Foundation
61. Hentaigana // URL: https://alcvps.cdh.ucla.edu/support/
62. KuLA (九郎) // URL: https://apps.apple.com/us/app/kula/id1076911000
63. MOJIZO (もじぞう: 文字の記録) // URL: https://aimojizo.nabunken.go.jp
64. Юмашева Ю. Ю. Автоматизированное распознавание рукописных текстов с помощью алгоритмов искусственного интеллекта: российский и зарубежный опыт // Цифровое востоковедение. 2023. Vol. 3. No. 1-2. DOI: 10.31696/S278240120026084-5
65. Shakespeare Documented // URL: https://shakespearedocumented.folger.edu/resource/family-legal-property-records
66. Тарасова Н. А. Новые методы изучения рукописного наследия Ф. М. Достоевского. Отчет о НИР (итоговый) // Федеральное государственное бюджетное учреждение науки Институт русской литературы (Пушкинский Дом) Российской академии наук, г Санкт-Петербург. 2021-2023. РНФ. Грант: 21-18-00333
67. Mains d'éru-dits (XVIe–XXe siècles) // Bibale. URL: https://mainsderudits.irht.cnrs.fr/
68. Peer M., Kleber F., Sablatnig R. Towards Writer Retrieval for Historical Datasets // In: Fink G. A., Jain R., Kise K., Zanibbi R. (eds). Document Analysis and Recognition – ICDAR 2023. Lecture Notes in Computer Science. 2023. Vol. 14187. Springer, Cham. DOI: 10.1007/978-3-031-41676-7_24
69. Christlein V., Marthot-Santaniello I., Mayr M., Nicolaou A., Seuret M. Writer Retrieval and Writer Identification in Greek Papyri. // In: Carmona-Duarte C., Diaz M., Ferrer M. A., Morales A. (eds). Intertwining Graphonomics with Human Movements. IGS 2022. Lecture Notes in Computer Science. 2022. Vol. 13424. Springer, Cham. DOI: 10.1007/978-3-031-19745-1_6
70. Fiel S., Sablatnig R. Writer Identification and Retrieval Using a Convolutional Neural Network // In: Azzopardi G., Petkov N. (eds). Computer Analysis of Images and Patterns. CAIP 2015. Lecture Notes in Computer Science. 2015. Vol. 9257. Springer, Cham. DOI: 10.1007/978-3-319-23117-4_3
71. Dhali Maruf A., Sheng He, Popovic M., Tigchelaar E., Schomaker L. A Digital Palaeographic Approach towards Writer Identification in the Dead Sea Scrolls // International Conference on Pattern Recognition Applications and Methods. 2017. DOI: 10.5220/0006249706930702
72. Волчкова М. А. Опыт персонификации писцов "Соборного уложения 1649 г." с применением цифровых технологий. Отчет о НИР/НИОКР (итоговый). 2015. Частное учреждение культуры Музей классического и современного искусства "Бурганов-Центр". Российский гуманитарный научный фонд. Грант: 14-01-00304

73. Cha S. H., Tappert C. C. Automatic detection of handwriting forgery // Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition. IEEE, 2002. C. 264-267.
74. Carrière G., Nikolaidou K., Kordon F., Mayr M., Seuret M., Christlein V. Beyond Human Forgeries: An Investigation into Detecting Diffusion-Generated Handwriting // In: Coustaty M., Fornès A. (eds). Document Analysis and Recognition – ICDAR 2023 Workshops. Lecture Notes in Computer Science. 2023. Vol. 14193. Springer, Cham. DOI: 10.1007/978-3-031-41498-5_1
75. Anmol H., Bibi M., Moetesum M., Siddiqi I. Deep Learning Based Approach for Historical Manuscript Dating // 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019. Pp. 967-972. DOI: 10.1109/ICDAR.2019.00159
76. Madi B., Atamni N., Tsitrinovich V., Vasyutinsky-Shapira D., El-Sana J., Rabaev I. Automated Dating of Medieval Manuscripts with a New Dataset // In: Document Analysis and Recognition – ICDAR 2024 Workshops: Athens, Greece, August 30-31, 2024. Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 2024. Pp. 119-139. DOI: 10.1007/978-3-031-70642-4_8
77. KFUPM Handwritten Arabic TexT // URL: http://khatt.ideas2serve.net/
78. Смирнов И. Н. О возможностях восстановления цифровых архивных текстов и распознавания рукописных арабских букв // Доклад на Международном форуме Казань-Экспо-2023 и Казанской цифровой неделе. URL: https://docs.yandex.ru/docs/view?url=ya-browser%3A%2F%2F4DT1uXEPRrJRXlUFoewruLkFYs7ubIAbSAY-xbL0IBKEaUp3AMQOVTSNPc-2YyqdfQrXgF3z9zrSTC_aAKNXel2yXz60D0C9kCdp5RwRSf9cFvtDbvmJ-yubbW85hEWb4ftUudW-2OSXY3dbwUtNbw%3D%3D%3Fsign%3DjlXgcIS8jxvD_9odPNQjyr4BS4YF5gk8ukUILjVYqjs%3D&name=Kazan-2023.docx&nosw=1
79. Public AI models in Transkribus // READ COOP. URL: https://readcoop.eu/transkribus/public-models/
80. AI Models For Transcribing German Text In Fraktur, Kurrent and Sütterlin // URL: https://blog.transkribus.org/en/3-ai-models-for-transcribing-german-text-in-fraktur-kurrent-and-sutterlin
81. Aswathy A., Maheswari P. U. Generative innovations for paleography: enhancing character image synthesis through unconditional single image models // Heritage Science. 2024. Vol. 12. No. 258. DOI: 10.1186/s40494-024-01373-4
82. Marti U. V., Bunke H. The IAM-database: an English sentence database for offline handwriting recognition // IJDAR. 2002. Vol. 5. Pp. 39-46. DOI: 10.1007/s100320200071
83. Mohammed H., Marthot-Santaniello I., Märgner V. GRK-Papyri: A Dataset of Greek Handwriting on Papyri for the Task of Writer Identification // 2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney, NSW, Australia, 2019. Pp. 726-731. DOI: 10.1109/ICDAR.2019.00121
84. Papers with Code // URL: https://paperswithcode.com/about; https://paperswithcode.com/datasets?task=optical-character-recognition&page=1
85. Hugging Face – The AI community building the future // URL: https://huggingface.co/datasets
86. HebrewPal // Hebrew Palaeography Album. URL: https://www.hebrewpalaeography.com/
87. Droby A., Vasyutinsky Shapira D., Rabaev I., Kurar Barakat B., El-Sana J. Hard and Soft Labeling for Hebrew Paleography: A Case Study // International Workshop on Document Analysis Systems. 2022. URL: https://link.springer.com/chapter/10.1007/978-3-031-06555-2_33
88. Digital Scriptorium // URL: https://digital-scriptorium.org/
89. Ressources // L'Institut de recherche et d'histoire des textes // URL:

https://www.irht.cnrs.fr/index.php/fr/qui-sommes-nous/lirht-en-bref
90. English Handwriting 1500–1700: An Online Course // Faculty of English. URL: https://www.english.cam.ac.uk/scriptorium/
91. Palaeography tutorial (how to read old handwriting) // The National Archives [Archived content] URL: https://webarchive.nationalarchives.gov.uk/ukgwa/20230801144244/https:/www.nationalarchives.gov.uk/palaeography/
92. MultiPal // URL: https://www.multipal.fr/en/welcome/
93. LAION-5B: A new era of open large-scale multi-modal datasets // LAION. URL: https://laion.ai/blog/laion-5b/
94. GRAPHOSKOP // URL: https://www.palaeographia.org/graphoskop/index.html
95. Millesimo (lancement) // URL: https://palaeographia.org/millesimo/index.html
96. Исаев Б. Л., Ляховицкий Е. А., Цыпкин Д. О., Чиркова А. В. "Vestigium" – комплекс программного обеспечения для анализа нетекстовой информации рукописных памятников // Историческая информатика. Информационные технологии и математические методы в исторических исследованиях и образовании. 2016. № 1-2(15-16). С. 72-83.
97. Deciphering medieval shorthand – can a digital tool solve the "Tironian Notes"? // Medievalists.net. URL: https://www.medievalists.net/2024/02/medieval-shorthand-tironian-notes/
98. OCR-D // URL: https://ocr-d.de/en/
99. Kitamoto Asanobu, Tarin Karanuwat. Kuzushi Character Recognition by AI and the Road to Full-text Search for Historical Materials // Specialized Library. 2020. Vol. 5. No. 300. Pp. 26-32.
100. CASIA-HWDB // URL: https://paperswithcode.com/dataset/casia-hwdb
101. CASIA Online and Offline Chinese Handwriting Databases // URL: https://nlpr.ia.ac.cn/databases/handwriting/home.html
102. Chinese Calligraphy Styles by Calligraphers // URL: https://www.kaggle.com/datasets/yuanhaowang486/chinese-calligraphy-styles-by-calligraphers
103. KuroNet Kuzushiji Ninshiki サービス (KuroNet 九郎) // URL: http://codh.rois.ac.jp/kuronet/; https://mp.ex.nii.ac.jp/kuronet/
104. Cursive Japanese and OCR: Using KuroNet // The Digital Orientalist. URL: https://digitalorientalist.com/2020/02/18/cursive-japanese-and-ocr-using-kuronet/
105. Сиренов А. В. Проект "История письма европейской цивилизации": коллекции памятников письменности академических институтов Санкт-Петербурга – оцифровка и изучение // Труды Отделения историко-филологических наук 2021: Ежегодник / Отв. Ред. В. А. Тишков. Том 11. М.: РАН, 2022. С. 125-134. DOI: 10.26158/OIFN.2022.11.1.010.
106. Tsypkin D. O., Tereschenko E. Yu., Balachenkova A. P., Vasiliev A. L., Lyakhovitsky E. A., Yatsishina E. B., Kovalchuk M. V. Comprehensive Studies of the Historical Inks of Old Russian Manuscripts // Nanotechnologies in Russia. 2020. Vol. 15. № 9-10. Pp. 542-550.
107. Ляховицкий Е.А., Цыпкин Д.О. Инфракрасная визуализация текста в изучении памятников древнерусской письменности // Историческая информатика. 2019. № 4. С.148-156. DOI: 10.7256/2585-7797.2019.4.31588 URL: https://nbpublish.com/library_read_article.php?id=31588
108. Айсманн К., Палмер У. Ретуширование и обработка изображений в PhotoShop. М.: Вильямс, 2008. 600 с.
109. Keys to the Past – Typewriters in the Records of the Federal Government // NARA. URL: https://archives-20973928.hs-sites.com/keys-to-the-past?ecid=ACsprvumObuCwkwzawZGYsTfDoztaLW7YuCcPtmTh2XiZbavjZ7PL0CPbJS3LhzYw3NkhW

yAUjgt
110. Sfardata – צֹרָה // URL: https://sfardata.nli.org.il/#/startSearch_He
111. Beit-Arié M. The new website of SfarData: The codicological database of the Hebrew Palaeography Project // The Israel Academy of Sciences and Humanities. URL: https://www.academia.edu/38849781/The_new_website_of_SfarData_The_codicological_database_of_the_Hebrew_Palaeography_Project_The_Israel_Academy_of_Sciences_and_Humanities
112. Grüning T., Labahn R., Diem M., Kleber F., Fiel S. READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents // DOI: 10.48550/arXiv.1705.03311
113. Boillet M., Kermorvant C., Paquet T. Multiple document datasets pre-training improves text line detection with deep neural networks // In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021. Pp. 2134-2141.
114. Claudio De S., Fontanella F., Maniaci M., Marrocco C., Molinara M., Scotto di Freca A. Automatic Writer Identification in Medieval Books // 2018 Metrology for Archaeology and Cultural Heritage (MetroArchaeo), 2018. Pp. 27-32. DOI: 10.1109/MetroArchaeo43810.2018.13633
115. He Sh., Sammara P., Burgers J., Schomaker L. Towards Style-Based Dating of Historical Documents // 2014 14th International Conference on Frontiers in Handwriting Recognition. 2014. Pp. 265-270. DOI: 10.1109/ICFHR.2014.52
116. Фролов А.А. Опыт применения инструментов геоинформатики в кодикологическом исследовании писцовых книг // Историческая информатика. 2020. № 2. С.218-233. DOI: 10.7256/2585-7797.2020.2.33330 URL: https://nbpublish.com/library_read_article.php?id=33330
117. Чиркова А. В. создание программного обеспечения для комплексного кодикологического анализа рукописно-книжных памятников и документов. Отчет по НИР (итоговый) // Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт истории Российской академии наук, Санкт-Петербург. 2013-2015. РГНФ. Грант: 13-01-12010
118. Ринчинов О. С. Цифровые модели кодикологии тибетских книг // Oriental Studies. 2021. Т. 14. № 3. С. 541-549. DOI: 10.22162/2619-0990-2021-55-3-541-549
119. Володин А. Ю. Цифровая дипломатика: ресурсы, подходы, тенденции // Проблемы историографии, источниковедения и методов исторического исследования: Материалы V научных чтений памяти академика И. Д. Ковальченко, Москва, 13 декабря 2013 г. М.: Московский государственный университет им. М. В. Ломоносова (Издательский Дом (Типография), 2014. С. 179-185.
120. Isola P., Zhu J. Y., Zhou T., Efros A. A. Image-to-image translation with conditional adversarial networks // In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. Pp. 1125-1134.
121. Huang X., Liu M. Y., Belongie S., Kautz J. Multimodal Unsupervised Image-to-image Translation // In: The European conference on computer vision (ECCV). 2018. DOI: 10.48550/arXiv.1804.04732
122. Bayerisch-tschechisches Netzwerk digitaler Geschichtsquellen // Porta fontium. URL: https://www.portafontium.eu/?language=de
123. Baloun J., Král P., Lenc L. How to Segment Handwritten Historical Chronicles Using Fully Convolutional Networks? // In: Rocha A. P., Steels L., van den Herik J. (eds). Agents and Artificial Intelligence. ICAART 2021. Lecture Notes in Computer Science. Vol. 13251. Springer, Cham. DOI: 10.1007/978-3-031-10161-8_9
124. Diplomata Belgica // URL: https://www.diplomata-belgica.be/colophon_fr.html
125. Sources diplomatiques // TELMA. URL: https://telma.hypotheses.org/category/sources-

diplomatiques
126. Breuel T. M., Ul-Hasan A., Azawi M. I. A. A., Shafait F. High-performance OCR for printed English and Fraktur using LSTM networks // In: 2013 12th international conference on document analysis and recognition. 2013. Pp. 683-687.
127. Shi B., Bai X., Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition // IEEE Trans Pattern Anal Mach Intell. 2017. Vol. 39(11). Pp. 2298.
128. Rahal N., Vögtlin L., Ingold R. Layout Analysis of Historical Document Images Using a Light Fully Convolutional Network // In: Fink G. A., Jain R., Kise K., Zanibbi R. (eds). Document Analysis and Recognition – ICDAR 2023. Lecture Notes in Computer Science. 2023. Vol. 14191. Springer, Cham. DOI: 10.1007/978-3-031-41734-4_20
129. Martínek J., Lenc L., Král P. Building an efficient OCR system for historical documents with little training data // Neural Comput & Applic. 2020. Vol. 32. Pp. 17209-17227. DOI: 10.1007/s00521-020-04910-x
130. Fleischhacker D., Kern R., Göderle W. Enhancing OCR in historical documents with complex layouts through machine learning // Int J Digit Libr. 2025. Vol. 26, 3. DOI: 10.1007/s00799-025-00413-z
131. Digimap // URL: https://digimap.edina.ac.uk/
132. Chiang Y. Y., Knoblock C. A. Recognizing text in raster maps // Geoinformatica. 2015. Vol. 19. Pp. 1-27. DOI: 10.1007/s10707-014-0203-9
133. Weinman J. Historical Maps. Research. CompSci.Grinnell // URL: https://weinman.cs.grinnell.edu/research/maps.shtml#data
134. Weinman J., Chen Z., Gafford B., Gifford N., Lamsal A., Niehus-Staab L. Deep neural networks for text detection and recognition in historical maps // In: 2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney, NSW, Australia, 2019. Pp. 902-909.
135. Historical Atlas of the Low Countries (1350–1800) – GIS of the Low Countries // URL: https://datasets.iisg.amsterdam/dataset.xhtml?persistentId=hdl:10622/PGFYTM
136. Li Z., et al. ICDAR 2024 Competition on Historical Map Text Detection, Recognition, and Linking // In: Barney Smith E. H., Liwicki M., Peng L. (eds). Document Analysis and Recognition – ICDAR 2024. Lecture Notes in Computer Science. 2024. Vol. 14809. Springer, Cham. DOI: 10.1007/978-3-031-70552-6_22
137. Baloun J., Král P., Lenc L. ChronSeg: novel dataset for segmentation of handwritten historical chronicles // In: Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART). 2021. Pp. 314-322.
138. 歴史GIS. ROIS-DS歴史的地理情報システム – (CODH) // URL: https://codh.rois.ac.jp/historical-gis/
139. Riedl C., Zanibbi R., Hearst M. A., et al. Detecting figures and part labels in patents: competition-based development of graphics recognition algorithms // IJDAR. 2016. Vol. 19. Pp. 155-172. DOI: 10.1007/s10032-016-0260-8
140. Jamieson L., Francisco Moreno-García C., Elyan E. A review of deep learning methods for digitisation of complex documents and engineering diagrams // Artificial Intelligence Review. 2024. Vol. 57. P. 136. DOI: 10.1007/s10462-024-10779-2
141. Wang H., Shan H., Song Y., Meng Y., Wu M. Engineering Drawing Text Detection via Better Feature Fusion // In: Fujita H., Wang Y., Xiao Y., Moonis A. (eds). Advances and Trends in Artificial Intelligence. Theory and Applications. IEA/AIE 2023. Lecture Notes in Computer Science. 2023. Vol. 13925. Springer, Cham. DOI: 10.1007/978-3-031-36819-6_23
142. Gemelli A., Marinai S., Pisaneschi L., et al. Datasets and annotations for layout analysis of scientific articles // IJDAR. 2024. Vol. 27. Pp. 683-705. DOI: 10.1007/s10032-024-00461-2

143. Shen Z., Zhang R., Dell M., Lee B. C. G., Carlson J., Li W. LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis // In: Lladós J., Lopresti D., Uchida S. (eds). Document Analysis and Recognition – ICDAR 2021. Lecture Notes in Computer Science. 2021. Vol. 12821. Springer, Cham. DOI: 10.1007/978-3-030-86549-8_9
144. Citizen Archivist // National Archives. URL: https://www.archives.gov/citizen-archivist
145. Антонов Д. Н., Скопин Ю. А. Опыт разработки электронной системы отечественной генеалогии с применением искусственного интеллекта: использование документов Архивного фонда РФ в режиме удалённого доступа // Архивный вестник: Сборник статей и материалов Научно-методического совета архивных учреждений Центрального федерального округа РФ. Вып. 26 / Отв. ред. О. В. Акимова. М.: Главное архивное управление города Москвы, 2022. URL: https://www.mos.ru/upload/documents/files/7256/ArhivniivestnikVip26.pdf
146. Указатель церквей // Портал "Государственный архив Вологодской области". URL: https://gosarchive.gov35.ru/user/sign-in/login
147. Turchin P., Rio-Chanona R. M. del, Hauser J., Kondor D., Reddish J., Benam M., Cioni E., Villa F., et al. Large Language Models' Expert-level Global History Knowledge Benchmark (HiST-LLM) // Advances in Neural Information Processing Systems 37 (NeurIPS 2024). URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/38cc5cba8e513547b96bc326e25610dc-Abstract-Datasets_and_Benchmarks_Track.html
148. Ng A. Unbiggen AI // IEEE Spectrum. 09 Feb. 2022. URL: https://spectrum.ieee.org/andrew-ng-data-centric-ai#toggle-gdpr
149. Motor de búsqueda PARES con Inteligencia Artificial // PARES. URL: https://pares.cultura.gob.es/pares-htr/
150. Oberbichler S., Petz C. Working Paper: Implementing Generative AI in the Historical Studies (1.0) // Zenodo. 2025. DOI: 10.5281/zenodo.14924737

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Рецензируемая статья посвящена анализу возможностей и ограничений применения искусственного интеллекта (ИИ) в исторических исследованиях, с акцентом на задачи источниковедения и специальных исторических дисциплин, о чем читателю становится известно в середине текста. Основной содержательный фокус сделан на современных возможностях автоматизированного распознавания текстов на основе электронных копий рукописных и машинописных исторических источников, а также на подготовке машиночитаемых наборов данных. Автор подробно рассматривает этапы исторического исследования, выделяя те, где использование ИИ наиболее перспективно (эвристика, обработка источников, создание баз знаний).
Предмет исследования сформулирован общо, стоит отметить, что работа носит скорее обзорно-аналитический характер, нежели предлагает какие-то эмпирические результаты исследования. Стилистически текст является выступлением в дискуссии, нежели исследованием. Тем не менее, именно такой подход позволяет охватить широкий спектр проблем, связанных с интеграцией ИИ в историческую науку.
Статья включает историко-научный анализ эволюции понятия ИИ и его прикладных направлений, авторскую систематизацию этапов исторического исследования с точки зрения их совместимости с технологиями ИИ, описание технических аспектов обработки исторических источников (сканирование, создание наборов данных, машинное обучение).

Автор опирается на обширный международный и отечественный опыт, приводя примеры проектов (Transkribus, eScriptorium, Socface) и технологических решений (нейросети, OCR/HTR). Однако методология могла бы выиграть от включения хотя бы одного-двух подробных разборов указанных примеров, а не простого их перечисления, или, например, количественного анализа эффективности конкретных инструментов (оценки вроде «успех делится в пропорции 70% на 30%», «с точность не менее 95–97%» не в счёт, так как указываются без сносок).

Тема статьи исключительно актуальна в контексте цифровой трансформации гуманитарных исследований. Рост доступности вычислительных мощностей, развитие генеративного ИИ и больших языковых моделей (LLM) открывают новые возможности, в том числи и для работы с историческими источниками. Автор справедливо подчеркивает необходимость адаптации традиционных методов источниковедения к цифровым вызовам, что особенно важно в условиях нарастающего объема оцифрованных исторических материалов. Хотя статья охватывает многие ключевые тренды, в ней почти не затронуты такие перспективные направления, как анализ больших данных для выявления исторических закономерностей или применение ИИ в виртуальной реконструкции объектов историко-культурного наследия.

Научная новизна работы проявляется в следующих аспектах: систематизация этапов исторического исследования с точки зрения применимости ИИ; детальный анализ технических требований к оцифровке источников (разрешение, режимы сканирования, учет текстуры носителей), что редко обсуждается в исторической литературе; обзор специфики создания палеографических, кодикологических и лексикологических наборов данных, включая проблемы их формирования для языков с нелатинской графикой.

Структура статьи логична, стиль изложения научный. Примечания и библиография впечатляют объемом и разнообразием, демонстрируют глубокую проработку темы (хотя часть ссылок, например, на YouTube-лекции не вполне соответствуют академическим критериям).

Автор предвосхищает возможную критику, подчеркивая, что ИИ не заменяет экспертов на этапах интерпретации результатов или постановки гипотез — это важный контраргумент против скептиков, опасающихся девальвации роли историка в свете успехов машинного обучения последних лет.

Необходимо высказать несколько конкретных замечаний по статье. Во-первых, кажется, название статьи не полностью согласуется с содержанием, было бы уместно подчеркнуть роль специальных исторических дисциплин в связи с проблемами использования искусственного интеллекта в исторических исследованиях, а также соотнести название с главным содержанием статьи — описанием наборов данных для автоматизированного распознавания текстов рукописных и машинописных источников.

Во-вторых, этапы исторического исследования сформулированы в авторской версии, что необходимо прямо указать в тексте. Нельзя утверждать, что эти этапы соотносятся с устоявшимися эпистемологическими взглядами. Возможно именно по этой причине в работе с перечнем библиографии в 150 наименований, нет ни одной ссылки на методологические работы с дискуссиями об этапах исторического исследования, и нет ссылки хотя бы на пару примеров исследований, которые соответствовали бы предложенной теоретической схеме. К слову, сегодня многие исторические исследования строятся не вокруг гипотезы, а вокруг исследовательского вопроса, и можно заметить, что многие исторические исследования идут к этому вопросу не от историографии, а от источников.

И наконец, утверждения о способности или неспособности искусственного интеллекта решать задачи разных этапов исторического исследования в статье представлены как

авторские постулаты без каких-либо проведенных на этот счёт экспериментов или тестов, доказывающих такую (не)возможность. К примеру, тезис «Очевидно, что в пп. 1–3 вряд ли возможно применение AI – для этого было бы необходимо оцифровать всю научную литературу в мире…» наводит на размышления о том, а под силу ли такая задача человеку — и способен ли в такой максималистской постановке вопроса историк хоть как-то сформулировать исследовательскую гипотезу, или все его силы уйдут на освоение всей научной литературы в мире?

Статья будет безусловно полезна историкам, архивистам и цифровым гуманитариям, занимающимся оцифровкой, распознаванием и исследованием исторических источников; разработчикам программного обеспечения, работающим с культурным наследием; студентам, изучающим методы исторической информатики.

Статья «К вопросу о применении искусственного интеллекта в исторических исследованиях» соответствует тематике журнала «Историческая информатика», предлагает междисциплинарный взгляд на интеграцию ИИ в исторические исследования. Несмотря на высказанные замечания, работа вносит существенный вклад в дискуссию о цифровизации гуманитарного знания. Рекомендую к публикации в разделе «Дискуссии и обсуждения» на страницах журнала «Историческая информатика» после уточнения названия статьи.