Имагология и компаративистика. 2025. № 23. С. 380–389
Imagology and Comparative Studies. 2025. 23. pp. 380–389

РЕЦЕНЗИИ

Рецензия
УДК 51-77
doi: 10.17223/24099554/23/18

Машинное обучение и представление информации: новые возможности цифровых архивов (рецензия на книгу: Artificial Intelligence, Archives and Manuscripts. New Relationships between the Virtual Archive and Its Referent. Edinburgh: University of Edinburgh, 2025. 584 p.)

Елена Наумовна Пенская

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия, e.penskaya@gmail.com

Аннотация. Представлена рецензия на книгу «Artificial Intelligence, Archives and Manuscripts. New Relationships between the Virtual Archive and its Referent» (2025). В данной коллективной монографии обсуждаются как технологические, так и правовые, интеллектуальные вопросы, с которыми сталкиваются исследователи и архивисты при автоматизированной работе с рукописным наследием, искусственным интеллектом и нейросетями.

Ключевые слова: машинное обучение, алгоритм, монография, исследование, искусственный интеллект, нейросеть

Источник финансирования: исследование проведено в Национальном исследовательском университете «Высшая школа экономики» в рамках проекта Российского научного фонда № 22-68-00066 «Культурное наследие России: интеллектуальный анализ и тематическое моделирование корпуса рукописных текстов».

Для цитирования: Пенская Е.Н. Машинное обучение и представление информации: новые возможности цифровых архивов (Рецензия на книгу: Artificial Intelligence, Archives and Manuscripts. New Relationships between

the Virtual Archive and Its Referent. Edinburgh: University of Edinburgh, 2025. 584 p.) // Имагология и компаративистика. 2025. № 23. С. 380–389. doi: 10.17223/24099554/23/18

Review
doi: 10.17223/24099554/23/18

Machine learning and information representation: New possibilities for digital archives (Book review: Nottorp, P. & Raymond, M. (eds) (2025) *Artificial Intelligence, Archives and Manuscripts. New Relationships between the Virtual Archive and Its Referent.* Edinburgh: University of Edinburgh)

Elena N. Penskaya

HSE University, Moscow, Russian Federation, e.penskaya@gmail.com

Abstract. The book *Artificial Intelligence, Archives and Manuscripts. New Relationships between the Virtual Archive and Its Referent* (2025) is presented. This collective monograph discusses both technological and legal, intellectual issues that researchers and archivists face in automated work with manuscript heritage, artificial intelligence and neural networks.

Keywords: machine learning, algorithm, monograph, research, artificial intelligence, neural network

***Financial Support*:** The research was conducted at the National Research University Higher School of Economics and supported by the Russian Science Foundation (RSF), Grant No. 22-68-00066: Cultural Heritage of Russia: Intellectual Analysis and Thematic Modeling of the Corpus of Handwritten Texts.

***For citation*:** Penskaya, E.N. (2025) Machine learning and information representation: New possibilities for digital archives (Book review: Nottorp, P. & Raymond, M. (eds) (2025) *Artificial Intelligence, Archives and Manuscripts. New Relationships between the Virtual Archive and Its Referent*. Edinburgh: University of Edinburgh). *Imagologiya i komparativistika – Imagology and Comparative Studies*. 23. pp. 380–389. (In Russian). doi: 10.17223/24099554/23/18

Цифровые рукописные коллекции стимулировали исследователей к разработке новых методов анализа больших данных, которые теперь эффективнее сохраняются и управляются благодаря программному обеспечению с открытым исходным кодом. Эти процессы сопровождаются множеством проблем, преимущественно открывающих новые перспективы, поскольку цифровые гуманитарные науки стали основополагающим направлением в академической среде. Тем не менее на этом пути возникают определенные сложности. Институции, чья деятельность ориентирована на сохранение культурного наследия, сталкиваются как минимум с тремя главными проблемами. Эти вопросы освещаются в совместной монографии, которая является результатом работы исследовательской группы из ведущих университетов, поддержанных финансированием университета в Эдинбурге. В книге рассматриваются современные вызовы, связанные с оцифровкой архивов, а также принципы и методы работы с ними. Также представлен дискуссионный форум, на котором обсуждаются разнообразные профильные темы.

В данной работе рассматриваются актуальные вопросы современности, связанные с оцифровкой архивных документов, а также с последующими принципами и методами их обработки. Кроме того, в ней выделяется и обсуждается широкий спектр дискуссионных тем, связанных с данной проблематикой.

Прежде всего, большие корпусы цифровых рукописных архивов существенно усложняют архивистам задачу оценки материалов. Как известно, внедрение искусственного интеллекта и машинного обучения в архивную практику все еще находится на стадии экспериментов. Тем не менее мы наблюдаем, как искусственный интеллект становится все более важным и неизменным помощником в архивных процессах. Для эффективного управления объемом данных и освоения инструментов тематического моделирования архивистам необходимы специальные методики, которые помогут им в принятии решений об оценке, отборе материалов, способах оцифровки и последующего анализа.

Во-вторых, множество цифровых коллекций рукописей сегодня остаются недоступными по самым разнообразным причинам, таким как технические ограничения, авторские права и вопросы защиты

данных. Авторы монографии справедливо подчеркивают, что невозможно найти оправдания для утверждения, что все цифровые данные должны быть открытыми и доступными. Тем не менее следует признать, что «темные зоны» внутри архивов, хранят в себе огромное количество важной информации для исследователей, включая массивы эгодокументов, черновики текстов. Увеличение доступности цифровых архивов является жизненно важной задачей для глубокого понимания нашего культурного наследия.

В-третьих, наука о данных и искусственный интеллект становятся важными инструментами, но очень немногие исследователи (особенно в области гуманитарных наук) прошли подготовку для овладения этими методами машинного обучения и компьютерного зрения. Включают ли современные образовательные программы те необходимые навыки, без которых сегодня студентам не обойтись? Кстати, присутствие современных исследований не только в качестве внушительного библиографического списка, но и материала для живого диалога, – полезное свойство «Artificial Intelligence, Archives and Manuscripts. New Relationships between the Virtual Archive and its Referent», расширяющее информационный ресурс монографии. Так, к примеру, достаточно подробно представлен комплекс идей, сформулированный в издании Института Алана Тьюринга «Проблемы и перспективы на стыке гуманитарных наук и науки о данных», поднимающем важный вопрос о том, как эффективно обучать и повышать квалификацию исследователей в области гуманитарных наук в применении количественных и вычислительных методов. Барбара Макгилливрей и ее соавторы подчеркивают необходимость интеграции основных принципов этих методов в учебные программы бакалавриата и магистратуры. Это позволит выпускникам гуманитарных специальностей освоить навыки для участия или руководства цифровыми проектами и даст им возможность развивать карьеру в области разработки исследовательского программного обеспечения и науки о данных, сосредоточенной на изучении работы нейросетей в гуманитарных дисциплинах.

Автоматизация, доступ и искусственный интеллект становятся неотъемлемыми инструментами для расшифровки нашей литературы и истории. Проблема заключается не в дефиците информации, а в ее избыточности. Доступ к рукописным архивам представляет собой

ключевую задачу, которую необходимо рассматривать в комплексном ключе, опираясь на методологии, основанные на данных. Вопросы, поднятые в книге «Artificial Intelligence, Archives and Manuscripts», являются особенно актуальными. Как можно обеспечить более широкий доступ к архивам, которые в настоящее время недоступны для общества? Какова роль автоматизации и искусственного интеллекта в этом процессе в целом и в работе с рукописными массивами данных? Каковы принципы «искусственного интеллекта во благо» в контексте анализа архивных коллекций? В сочетании с другими передовыми технологиями искусственный интеллект способен сделать цифровые архивы более полными и обеспечить аутентичность записей. Проблема утраты наследия и особенно трудновосстановимой рукописной его части является одной из самых болезненных тем, обсуждаемых в настоящее время. Как известно, документы могут быть уничтожены намеренно либо случайно. Процесс оценки и отбора источников играет ключевую роль в архивной практике. В свое время американский архивист Т.Р. Шелленберг отмечал, что запись имеет «первичную ценность» для ее создателя, но также может обладать «вторичной ценностью» (в качестве доказательства или сведений) для историков и будущих пользователей данной информации. Он подчеркивал, что архивисты должны не только иметь право, но и обязанность проверять все факты, которые государственные учреждения намерены уничтожить вместе с их материальным носителем.

В эпоху цифровизации ручной просмотр больших корпусов документов стал практически невозможным. Архивистам необходимо справляться с колоссальным числом единиц хранения, распределенных по множеству мест. В монографии «Artificial Intelligence, Archives and Manuscripts» исследователи рассматривают серьезную современную угрозу: массовое внедрение доступных облачных технологий, а также смартфонов и других мобильных устройств способствует распространению явления, известного как теневой генеративный искусственный интеллект. Это означает, что государственные служащие могут использовать свои личные электронные почты или приватные мессенджеры, такие как WhatsApp и Skype, а также социальные сети, такие как Facebook и Twitter, для обмена информацией с коллегами. Это теневой искусственный интеллект существенно уве-

личивает и риск утечек данных и незаконного копирования рукописных источников, что вызывает искажение при последующем их воспроизведении, распространении, публикации и комментировании. В частности, в Великобритании законодательство в области архивов устанавливает правило, согласно которому правительственная документация, предназначенная для постоянного хранения, должна быть направлена в Национальный архив в течение двадцати лет с момента ее создания, вместо трех десятилетий, как это было принято ранее.

Для государственных экспертов в области управления знаниями и информацией теневой искусственный интеллект создает серьезное препятствие. Искусственный интеллект имеет потенциал для интеграции разрозненных архивов, что способствует их доступности для поиска и использования. Однако в данной ситуации мы сталкиваемся с парадоксом, который рассматривается с различных углов зрения в рецензируемой книге. В первую очередь это касается двойственной роли самого искусственного интеллекта, который может представлять риск для целостности архивной информации.

В монографии ставится провокационный вопрос: не лучше ли архивистам просто пассивно принимать поступающую цифровую информацию и хранить ее без изменений? Однако для многих экспертов в области управления знаниями и информацией этот подход неприемлем. Цифровая информация, создаваемая организациями, хаотична и неструктурирована. Чтобы облегчить поиск информации и оперативно реагировать на запросы, необходимо перерабатывать данные, аннотировать их метаданными и активно интерпретировать, объединяя в тематические коллекции, которые затем облекаются в форму артефактов. В этом процессе искусственный интеллект будет играть ключевую роль, автоматизируя добавление метаданных, извлечение имен и названий, а также содействуя тематическому моделированию – статистическому методу, применяемому в машинном обучении и обработке естественного языка для определения кластеров слов или тем в больших массивах данных.

Используя искусственный интеллект для автоматической идентификации конфиденциальных записей, архивисты могли бы более эффективно задействовать фактор тональности и «нечувствительные» данные. Анализ тональности, проводимый с помощью специализированного программного обеспечения, все еще находится на начальной

стадии, но обладает значительными перспективами, которые могут помочь выявить скрытое в архивах. Применение искусственного интеллекта к большим массивам данных может привести к созданию обширных реорганизованных архивов, свободных от конфиденциальной информации. В рамках книги «Artificial Intelligence, Archives and Manuscripts» поднимается важный вопрос: должна ли роль архивиста заключаться в формировании алгоритмов и написании кода, необходимых для принятия решений? Авторы работы выделили четыре ключевые развилки, которые требуют обсуждения.

Во-первых, архивисты отвечают за сохранность документов, которые необходимы историкам и другим исследователям. Во-вторых, они, обладая необходимыми профессиональными навыками, часто оказываются вне процесса разработки и внедрения технологий искусственного интеллекта. Обычно их приглашают лишь в последнюю очередь, что подчеркивает их незаслуженное игнорирование. В-третьих, архивисты не обладают достаточными возможностями и знаниями, необходимыми для того, чтобы стать настоящими экспертами в области современных цифровых операций, что гарантирует долгосрочное хранение новых архивных данных. В-четвертых, важно, чтобы архиввисты могли не только консультировать по вопросам использования алгоритмов искусственного интеллекта, но и разбираться в ключевых правовых аспектах. Применение искусственного интеллекта к архивам грозит искажением литературных и исторических документов, значит, и самой истории, а также нашей коллективной памяти. Один из авторов книги утверждает, что автоматизация становится не опцией, а необходимостью, но это вовсе не означает, что роль архивиста становится незначительной или второстепенной в сложной системе.

Одним из ключевых моментов в данной монографии становится описание проекта SIMULATION (2022–2023), разработанного для решения вопросов сохранения и представления целостности и аутентичности цифровых архивных материалов, рукописных в том числе. В рамках проекта были изучены возможности, которые предоставляют технологии машинного обучения, позволяющие добавлять данные в цепочку, при этом предотвращая их изменение, перезапись или удаление. В результате создается непрерывный список записей-кла-

стеров; каждый из них содержит разметку и другую важную информацию. Применение машинного обучения обеспечивает создание уникального цифрового отпечатка архивных материалов, что позволяет удостовериться в их подлинности.

Проект SIMULATION разработал прототип для создания хэшей, используя методы машинного обучения, особенно на примере рукописных изображений. Такой подход позволяет машине выявлять причины сбоев и шумов в этих записях, возникающих как в результате повторного кодирования и изменения формата, так и в случае нарушения целостности рукописного документа. Машинное обучение в SIMULATION позволяет загружать метаданные архивных коллекций с высокой точностью, что помогает идентифицировать конкретные рукописные документы. Одно из предложенных решений заключается в добавлении архивной ссылки и контрольной суммы записи – уникальной строки, созданной компьютером, которая изменяется при трансформации файла, фиксирующего рукописный источник. Эти данные затем размещаются в блоке, который не может быть изменен или удален бесследно. В завершение копия данных передается всем участникам сети. SIMULATION наглядно демонстрирует, что искусственный интеллект может продуктивно действовать в интересах архивных коллекций.

Монография разделена на две части. Первая часть, затрагивающая темы «Отбора, оценки, обнаружения и доступа», открывается интересным кейсом, где рассматривается использование искусственного интеллекта в рукописном архиве справочной библиотеки Frick Art Reference Library. Благодаря междисциплинарной команде, обладающей знаниями в области компьютерных наук, истории искусства, литературы и других дисциплин, эта коллекция стала более доступной и удобной для использования. Вторая глава, речь в которой идет о веб-архивах, также носит междисциплинарный характер и посвящена крупномасштабным цифровым рукописным коллекциям. Третья глава рассматривает дизайн-мышление, сосредоточенное на методах решения социальных проблем. В ней обоснованно подчеркивается необходимость сотрудничества исследователей, архивистов и математиков, использующих искусственный интеллект и другие современные подходы.

Второй раздел, озаглавленный «Использование архивов: искусственный интеллект и новые знания», начинается с главы, посвященной исследованию пользователей цифровых библиотек. Пауль Натторп, анализируя влияние электронного правового регулирования на британские академические депозитные библиотеки, подчеркивает важность прозрачности рабочих процессов и документирования данных в исследованиях пользователей, которые часто опираются на инструменты, такие как Google Analytics, но оказываются уязвимыми. Качество данных о посетителях библиотек также вызывает серьезные вопросы.

В пятой главе Мартин Реймонд и его коллеги изучают масштабный набор данных – анализ феноменологии эксперта при работе с нейросетью в сфере разметки и постановки поисковых задач. В данном разделе описаны сложности, с которыми сталкиваются эксперты в области применения нейронных сетей.

Шестая глава, написанная Тобиасом Зантельманом, сосредоточена на распознавании рукописных текстов (HTR). Контролируемые методы глубокого обучения продемонстрировали впечатляющие результаты в расшифровке рукописных материалов, однако появились и новые вызовы, касающиеся прозрачности. Зантельман также делится обзором машинного обучения, регулируемого ООН, которое включает тематическое моделирование и применяется для работы с большими объемами текстовых данных.

Продолжая обсуждение HTR, в седьмой главе Мелисса Маккарти акцентирует внимание на том, как радикально изменяет влияние этой технологии на доступ к нашему наследию, зафиксированному в рукописях. На основании опроса пользователей системы HTR на платформе Transkribus она обсуждает проблемы, возникающие при интеграции машинного обучения в литературные и исторические архивы. Маккарти, утверждая, что транскрипции, созданные с помощью HTR, потребуют нового подхода как к историографии, так и к вовлечению общественности, предлагает рекомендации по поддержке сообщества, использующего HTR для работы с материалами культурного наследия.

Наконец, послесловие Ричарда Донкинза очерчивает будущие перспективы взаимодействия технологий и архивов, открывая новые

пути для сохранения и обеспечения доступа к нашему коллективному прошлому.

Список источников

1. Artificial Intelligence, Archives and Manuscripts. New Relationships between the Virtual Archive and its Referent. Edinburgh : University of Edinburgh, 2025. 584 p.

References

1. Nottorp, P. & Raymond, M. (eds) (2025) *Artificial Intelligence, Archives and Manuscripts. New Relationships between the Virtual Archive and Its Referent*. Edinburgh: University of Edinburgh.

Информация об авторе:
Пенская Е.Н. – д-р филол. наук, профессор факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики» (Москва, Россия). E-mail: e.penskaya@gmail.com

Автор заявляет об отсутствии конфликта интересов.

Information about the author:
E.N. Penskaya, Dr. Sci. (Philology), professor, National Research University Higher School of Economics (HSE University) (Moscow, Russian Federation). E-mail: e.penskaya@gmail.com

The author declares no conflicts of interests.

Статья принята к публикации 20.11.2024.

The article was accepted for publication 20.11.2024.