Раздел 4
ИСТОРИКО-КУЛЬТУРНОЕ НАСЛЕДИЕ И ИНСТИТУТЫ ПАМЯТИ: МОДЕЛИ ВЗАИМОДЕЙСТВИЯ

УДК 930+02+069

Ю. Ю. Юмашева

ИСТОРИЧЕСКАЯ НАУКА, АРХИВЫ, БИБЛИОТЕКИ, МУЗЕИ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ГОД СПУСТЯ

Статья представляет собой краткий обзор научных мероприятий, результатов реализации исследовательских проектов и практики применения методов и методик Data Science (искусственного интеллекта, машинного обучения, глубокого машинного обучения, компьютерного зрения и т. п.) в современных исторических исследованиях, а также в основных направлениях архивной, библиотечной и музейной деятельности. Автор описывает новации, которые произошли в данной области в течение 2021 – первой половины 2022 гг.

Ключевые слова: науки о данных, наборы данных, искусственный интеллект, машинное обучение, компьютерное зрение, исторические исследования, архивы, библиотеки, музеи.

Весной 2021 г. в ДАИС была опубликована статья [Юмашева, 2021], посвященная обзору применения технологий Data Science (далее DS) в практической деятельности крупнейших фондообразователей – архивов, библиотек, музеев, а также в научных исследованиях специалистов-историков.

В рамках настоящего обзора рассмотрим изменения, которые произошли за год в вопросах применения и обсуждения методов DS на уровне международных профильных организаций, изучим некото-

© Юмашева Ю. Ю., 2022

рые опубликованные результаты научных исследований теоретической и прикладной гуманитарной направленности, а также обозначим перспективы применения DS в архивной и музейной деятельности.

Однако, прежде чем начать рассмотрение сформулированных вопросов, напомним, что DS является своеобразным «зонтичным брендом», который подразумевает не одну, а целый набор дисциплин из разных областей знания, отвечающих за сбор (преобразование, извлечение), обработку (*data wrangling* (*data preparation*)) и анализ данных (*analysis*), а также их использование (поиск на их основе оптимальных решений) и архивирование (*using and archiving*). С течением времени круг задач и методов DS значительно расширяется и в настоящее время, кроме математической статистики и машинного обучения, которые были своеобразным «началом отсчета» в развитии DS в рамках «наук о данных», развиваются «computer science», «deep learning» («глубокое обучение»), «искусственный интеллект» (включая нейросети), компьютерное зрение и т. п.

Взрывной рост количества исследований, выполненных в разных областях с применением методов DS, и прежде всего искусственного интеллекта (далее – Artificial Intelligence, AI) [Artificial Intelligence Index Report 2022], стал катализатором создания автоматизированных поисковых систем (платформ), отслеживающих публикации о подобных проектах. Одна из первых подобных систем была создана и введена в эксплуатацию осенью 2021 г. – Globalpolicy.AI (URL: https://globalpolicy.ai/en/). Инициаторами ее создания выступили восемь международных организаций, наделенных полномочиями регламентации работ в области развития и применения AI. Платформа позволяет отслеживать значимые новости о развитии AI, опубликованные в разных странах на наиболее распространенных языках мира.

Одновременно с развитием общей поисковой системы каждый из ее учредителей осуществляет собственные проекты и программы и разрабатывает нормативные акты, призванные регламентировать применение искусственного интеллекта в курируемых отраслях. В частности, 24 ноября 2021 г. Генеральная ассамблея ЮНЕСКО приняла Рекомендации по этике искусственного интеллекта [The Recommendation…], которые являются основополагающим документом для развития и применения методов DS в нескольких отраслях, находящихся в прямом или косвенном «подчинении» ЮНЕСКО. В свою очередь каждая из профильных международных организаций, действующих под эгидой ЮНЕСКО – IFLA, ICOM, ICA, – также предпринимают определенные усилия в вопросе изучения возможностей применения DS и, прежде всего, искусственного интеллекта в практике «подведомственных» им структур.

Кратко рассмотрим итоги их деятельности за последние пару лет, уделив основное внимание архивной проблематике.

Библиотеки

В декабре 2021 г. Международная федерация библиотечных ассоциаций и учреждений (IFLA) на базе Национальной библиотеки Франции провела тематический семинар и объявила о создании специализированной группы (AISIG), которой поручается координация и обобщение результатов библиотечных исследований в области AI [Conférence…, 2021]. Одновременно было обновлено Заявление IFLA 2020 г. [IFLA Statement…] о библиотеках и искусственном интеллекте, в котором отмечались основные тенденции, которые повлияют на деятельность библиотечного сообщества, прежде всего, цифровых библиотек в связи с «достижениями в области искусственного интеллекта», которые позволят:

«а) следующему поколению веб-браузеров выйти за рамки анализа ключевых слов и оценить конкретное содержание веб-сайтов/страниц (семантическая сеть);

б) сетевым устройствам сочетать распознавание речи, машинный перевод и синтез речи для поддержки многоязычного голосового перевода в реальном времени; и

в) осуществить облачную краудсорсинговую проверку перевода текста веб-страницы».

Проведенный под эгидой IFLA Парижский семинар был третьим в серии семинаров, которые с 2018 г. проводит сетевое сообщество и платформа ai4lam (Artificial intelligence for libraries, archives and museums, URL: https://sites.google.com/view/ai4lam/home), созданное в 2018 г. пятью библиотечными и научными организациями Норвегии, США, Великобритании и Франции. Целью данного сообщества является координация совместной деятельности библиотек, архивов и музеев по изучению, обмену опытом и повышению уровня знаний об искусственном интеллекте и его использовании в сопредельных библиотечно-музейно-архивных областях. Кроме тематических рабочих групп, которые функционируют в рамках сетевого сообщества, на соответствующей странице сайта ai4lam представлен список информационных ресурсов, реализованных с помощью методов DS и доступных онлайн.

Проанализировав данный список, а также опубликованные результаты уже осуществленных исследований, можно отметить, что наиболее распространенными и популярными технологиями DS, применяемыми в библиотеках, являются технологии автоматизированно-

го поиска (например, на платформе EBSCO [Expert.ai's…], машинного обучения, оптического распознавания рукописных и иероглифических/пиктографических текстов, автоматизированной предметной индексации и классификации машиночитаемых текстов (например, программное обеспечение Annif, URL: http://www.annif.org/), создание информационно-поисковых систем на основе семантических сетей и анализа изображений (AIDA – Image Analysis for Archival Discovery, URL: https://projectaida.org/), использование методов анализа конкретных тем «в контекстах», а также создание наборов открытых данных на основе библиотечных каталогов и машиночитаемой информации и т. п.

Музеи

Удивительно, но международные организации, курирующие музейное сообщество (ICOM, ICOMOS), одним из первых начавшие применять искусственный интеллект и робототехнику в работе с посетителями, до настоящего времени так и не выработали общих подходов и не опубликовали никаких руководящих документов, посвященных применению DS и AI в практической музейной деятельности. Однако в ноябре 2020 г. в Южной Корее был проведен Первый Всемирный музейный форум, посвященных проблемам применения AI в музеях. К сожалению, запланированное четырехдневное мероприятие было сокращено до двух дней и проводилось в режиме онлайн [ICOM Korea…].

Большинство проектов, реализованных в музейной сфере, как правило, осуществляется либо «временными творческими коллективами» – объединениями музейных специалистов, либо конкретными музеями в инициативном порядке. К примеру, в 2019 г. английскими и американскими музейщиками было создано сетевое сообщество «Museums+AI» (URL: https://themuseumsai.network/), которое провело серию семинаров по обсуждению возможностей применения AI в музеях и опубликовало в феврале 2020 г. документ, излагающий результаты дискуссий и исследований [AI: A Museum Planning Toolkit].

В документе в частности подчеркивалось, что музеи используют различные технологии (веб-сайты, чат-боты) и методы DS (аналитическая статистика, машинное обучение, AI и компьютерное зрение) для работы с посетителями (в том числе для выяснения их настроений и восприятия произведений искусства) и, что в настоящее время особенно важно, – для автоматизации внутримузейных процедур (аутентификации предметов [Introducing…], определения подделок, аннотирования, описания, классификации музейных предметов, обогащения

и изучения метаданных [Explore…]), а также для поиска произведений, близких по содержанию, композиции и другим критериям. Цель большинства подобных исследований – «предоставить искусствоведам инструменты для навигации по огромным базам данных изображений».

Перечисленные направления являются основными и в продолжающемся в Великобритании крупномасштабном проекте «Национальная коллекция» (URL: https://www.nationalcollection.org.uk/), одна из целей которого – применение AI для изучения культуры и историко-культурного наследия.

Весьма неожиданное применение методам DS было найдено Интерполом, который разработал мобильное приложение для определения украденных ценностей ID-art (URL: https://www.interpol.int/en/Crimes/Cultural-heritage-crime/ID-Art-mobile-app). Аналогичное программное обеспечение на основе машинного обучения и AI создается в Институте исследований компьютерной графики Фраунгофера в Дармштадте (Германия).

Таким образом, направления применения методов и технологий DS в музейной сфере чрезвычайно разнообразны, что позволяет высказать предположение о том, что AI и машинное обучение станут в ближайшее время одними из самых востребованных информационных технологий, используемых в музеях.

Архивы

В начале 2020-х гг. ни для кого уже не является секретом, что цифровая трансформация превращает старые и новые архивы в хранилища данных. Как следствие, автоматизация в форме методов DS, и прежде всего искусственного интеллекта, все чаще применяется как для масштабирования традиционной деятельности по ведению документации, так и для экспериментов с новыми способами сбора, организации, хранения и доступа к документам.

Архивы, изначально предназначенные для хранения огромного числа документов, превращаются в организации, которые либо уже хранят данные (в форме электронных документов), либо документную информацию, которая имеет потенциал к преобразованию в данные. Таким образом, преобразование традиционных архивов из собраний документов в коллекции данных (больших данных) является одним из основных последствий цифровой трансформации [Moss, et al., 2018]. Работа с этими массивами документированной информацией должна хотя бы частично базироваться на применении DS, в том числе на искусственном интеллекте и системах машинного обучения, без использования которых практически невозможно справиться с преобразованием документов на традиционных носителях в цифровой вид.

Казалось бы, эта реальность должна подталкивать Международный совет архивов и его региональные и профильные отделения к интенсификации изучения применения методов DS в архивном деле, тем более что в 2018 г. секретарь ICA Антея Селеш выступила с докладом, пообещав, что изучение возможностей применения DS и AI будет развиваться, как никакое другое направление. Результатом этого заявления стало создание профильной группы ICA – Центра исследований искусственного интеллекта в архивах (Le Hub for Artificial Intelligence Research in Archives, HAIRA) и соответствующего информационного ресурса (URL: https://haira.clir.org/artificial-intelligence-and-archives/), о начале деятельности которых было заявлено на Международной неделе архивов в июне 2021 г. (URL: https://www.ica.org/en/artificial-intelligence-and-diversity-a-turning-point-in-archival-practice). К сожалению, деятельность группы и портала не отличается большой активностью.

Значительно бóльшую активность проявляют национальные и профильные общественные объединения архивистов и специалисты крупнейших архивов мира, в частности, архивных организаций Великобритании, США, Нидерландов, Израиля и др.

К примеру, в Великобритании функционирует сеть AURA (AURA Network, URL: https://www.aura-network.net/), которая зимой 2020 – весной 2021 гг. провела три мероприятия на тему «AI и архивы» и публикует сборники статей по данной проблематике.

Англо-американский информационный ресурс – портал AEOLIAN (Artificial Intelligence for Cultural Organizations, URL: https://www.aeolian-network.net/) представляет результаты деятельности одноименной сети по применению AI в культурных учреждениях, преимущественно в архивах. В рамках этого проекта организуются шесть семинаров для специалистов по цифровым гуманитарным наукам, компьютерщиков, архивариусов и т. п.; публикуются обзоры, посвященные стратегиям применения DS и AI в архивах разных стран [Case Study 1].

Кроме упомянутой «сетевой» активности, следует подчеркнуть, что с середины 2010-х гг., когда были предприняты первые попытки использования методов DS и AI в архивной науке и практике, количество публикаций в профессиональных изданиях по данной проблематике исчисляется уже сотнями и даже тысячами. При этом практически все исследования могут быть сгруппированы в четыре тематические группы:

1. Вопросы теории архивного дела в контексте DS, в том числе теоретическое осмысление перспектив профессии архивиста, ее мо-

дификации в связи с активным проникновением информационных технологий в архивную практику, а также в связи с возникновением междисциплинарного направления Computational Archival Science, требующем значительных изменений в традиционных архивном мышлении и подходах [Lemieux].

2. Практические результаты автоматизации процессов комплектования и учета архивной документации, прежде всего электронных документов и электронной почты. Среди обсуждаемых экспериментов – создание и применение онтологий, баз знаний, машинного обучения, семантических сетей, нейролингвистического программирования (НЛП) и методов формирования и состава метаданных как инструментов преодоления «дезинтеграции» электронных документов[1] [Büttner, 2017; Spencer, 2017]; выявления полных и частичных дубликатов, хранящихся в разных информационных системах и на разных носителях, проведения экспертизы ценности и отбора документов на постоянное хранение, определения сроков хранения и критериев оценки устойчивости форматов файлов для долгосрочного хранения. Резюмируя итоги экспериментов по изучению возможностей применения DS и AI в практической архивной деятельности, специалист по цифровому архивированию в The National Archives (TNA) и руководитель проекта «AI for Digital Selection» С. Венката отмечает, что инструменты AI могут помочь архивистам в работе с документами, но заменить человеческий опыт машина пока не может [Venkata, 2021].

3. Организация доступа к архивной документации, прежде всего выявление документов с конфиденциальной (персональной, коммерческой и т. п.) информацией, не подлежащих представлению в открытый доступ (при этом, как отмечают сами архивисты – участники проектов, «проверка с помощью технологий никогда не будет на 100 % точной») [The National Archive]; редактирование конфиденциальной/личной/коммерческой информации с использованием инструментов цифровой криминалистики с целью сокрытия конфиденциальной информации (например, адресов электронной почты, номеров телефонов и т. п.); методов совершенствования поиска (в том числе при помощи замены ключевых слов оптимизированными AI наборами метаданных и их тематическим моделированием[2] – созданием «капсульных»

[1] Например, разработанный метод автоматизированного создания метаданных и выявления с их помощью латентно присутствующих взаимосвязей между документами для обнаружения дополнительных содержательных аспектов. Метод основан на семантических правилах автоматической классификации официальной документации и апробирован на документах Совета Европы.

[2] Тематическое моделирование группирует слова в кластеры на основе их сходства. Опираясь на неконтролируемые и контролируемые методы машинного обучения, этот метод интеллектуального анализа текста можно использовать для выделения основных

тематических моделей [Coeckelbergs, van Hooland, 2020]), «дистанционного чтения» для индексации не только текстов, но и изображений в архивных документах; ускорения получения (извлечения) информации и удобства ее использования; повышения доверия к оцифрованному/цифровому контенту и увеличения внимания к содержанию документов (а не только их метаданных. – *Ю. Ю.*) в архивной практике, и даже «датафикация[3], интеграция, интероперабельность и доступности для анализа [Bikakis, et al., 2021]». Так, в целях автоматического извлечения (data wrangling) и индексации контента активно разрабатываются и применяются известные программные продукты типа Transcribus[4], используется интеграция в них архивных «исторических индексов» (персональных, географических, предметных и т. п. указателей) [Ranade, 2016]. Последний подход позволяет создавать более гибкие механизмы поиска в архивных информационных системах, а также значительно ускорить процессы индексации [Colavizza, et al., 2019] и создания связанных данных, в том числе с ресурсами музеев и библиотек [Wilde, Hengchen, 2017].

Вместе с тем применение AI для поиска информации в архивных каталогах не так однозначно. В частности, изучая и анализируя в течение нескольких лет российские и зарубежные архивные информационные ресурсы, автор не может не упомянуть возражения скептиков в отношении применения методов DS для профессионального поиска и выявления документов для целей исторических исследований. В первую очередь здесь необходимо назвать отчет американских архивистов по изучению проблем обеспечения доступа к архивным документам [Ingram, Johnson, 2021], в котором отмечается «предвзятость» результатов применения поисковых механизмов, основанных на использовании искусственного интеллекта, с помощью которых осу-

тем в наборе данных – например, в метаданных каталога, описывающих большую коллекцию и создавая «модель» терминов, присущих описаниям той или иной предметной области.

[3] «Датафикация… представляет разнообразную информацию в виде машиночитаемых, поддающихся количественному измерению данных для целей агрегирования и анализа [Southerton, 2020]». Под датафикацией подразумевают также представление в машиночитаемом формате содержания архивных документов и связанных с ними метаданных.

[4] Transkribus (URL: https://readcoop.eu/transkribus/?sc=Transkribus) – инструмент, который объединяет модели распознавания изображений и текста для облегчения распознавания рукописных и оптических символов. Инструмент, доступный через графический интерфейс или программно, через API, разработан, чтобы «аккуратно вписаться в архивный рабочий процесс, напрямую используя растущие репозитории оцифрованных изображений исторических текстов». Transkribus все чаще используется архивами и другими учреждениями, играя решающую роль в расширении использования AI для извлечения содержимого контента.

ществляется поиск в различных каталогах − от архивных до больших поисковых машин − и делается попытка найти пути нивелирования этого отрицательного эффекта. К числу способов, которые предлагаются для решения этой задачи, относится активное использование методов анализа контекста, машинного обучения, или так называемого «критического поиска», смысл которого заключается в применении «инструментальной критики» на каждом этапе процесса поиска в цифровых архивах (своеобразная адаптация и цифровая трансформация методов традиционной эвристики к специфике цифровой среды).

4. Новые формы цифровых архивов. Речь идет о том, что растущие объемы оцифрованных и born-digital архивных документов требуют переосмысления и автоматизации организации хранения и доступа к ним. Среди методов, которые могут быть использованы для решения данных проблем, рассматриваются распределенное хранение/блокчейн (технология распределенного реестра − Distributed Ledger Technology, DLT), внедрение хеш-кодов и т. п.

Краткий итог анализу публикаций по проблемам применения DS в архивах можно подвести словами венгерского разработчика программного обеспечения для архивов и библиотек Петера Кирали, который, оценивая перспективы исследований в части применения DS в практике архивного дела, перечислил следующие направления:

– применение аналитических методов (например, анализ текста, анализ эмоций, сетевого анализа) к архивным материалам;

– анализ текстов документов для архивной обработки, определения возможности доступа к документам, идентификация личных данных, оценки, регистрации, описания документов;

– масштабируемость бизнес-процессов, реализуемых в архивных службах (например, идентификация, сохранение, создание метаданных, проверка целостности, нормализация, проверка, связанные данные, распознавание объектов, анонимизация, удаление/редактирование для целей использования);

– изучение возможностей применения DS для комплектования (архивирования) и научно-технической обработки новых информационных ресурсов, получаемых из Интернет, социальных сетей, электронной почты и т. п., а также организация и функционирование аудиовизуальных репозиториев, блокчейн и т. п.;

– ИТ-инфраструктура для архивных исследований, создания и хранения коллекций;

– цифровая консервация, реставрация и сохранение;

– краудсорсинг и архивы;

– большие данные и архивная теория и практика;

– большие данные и построение памяти/идентичности; конкретные технологии больших данных (такие как базы данных NoSQL) и их приложения; архивные корпуса больших данных и справочные коллекции; связанные данные и архивы; большие данные и происхождение; подготовка данных для исследований из архивных материалов; юридические и этические вопросы [Király, 2018].

За три года с момента публикации данного перечня для некоторых упомянутых в нем направлений уже выполнены научные исследования [Nix, 2021] и реализованы программные решения (например, разработано несколько пакетов программного обеспечения для архивирования и научно-технической обработки электронной почты, в том числе ePADD (URL: https://epadd.stanford.edu/epadd/about). Однако большинство обозначенных тем до настоящего времени являются все еще актуальными. К тому же их дополнение и развитие не заставит себя ждать. Как справедливо отмечалось в обзоре Стэнфордского университета, «по мере того, как люди привыкают к конкретной технологии, она перестает считаться AI, и появляются новые технологии AI [Stanford University]», которые потребуют проведения новых исследований.

* * *

Рассмотрим результаты нескольких реализованных и представленных в сети проектов по трем основным направлениям применения DS в гуманитарных областях.

Data wrangling (data preparation) (Преобразование текстов)

Данное направление основано на сочетании применения средств оптического распознавания OCR и методов DS (AI и машинное обучение) для автоматического распознавания рукописных и идеографических/пиктографических текстов.

Зимой 2022 г. в журнале Nature опубликована статья [Assael, et al., 2022], авторы которой представляют результаты разработки программного обеспечения под названием «Итака» (Ithaca) с использованием AI, которое может помочь историкам восстановить древнегреческие тексты с 72 % точностью. Программа также способна определить возраст древних надписей с погрешностью всего лишь в 30 лет и высказать предположение с 70 % точностью о том географическом месте, где был создан текст. Разработку осуществили специалисты британской компании Deep Mind в сотрудничестве с историками.

Аналогичные проекты реализованы:

– в Чикагском университете, где в рамках проекта DeepScribe с помощью машинного обучения, компьютерного зрения и AI были

прочтены и расшифрованы тысячи глиняных табличек Персепольского фортификационного архива (PFA), датируемого древнеиранской империей Ахеменидов (550–330 до н. э.). [Persepolis…];

– во всемирно известном Massachusetts Institute of Technology, который предложил собственную систему CSAIL, читающую многие «потерянные языки», но в первую очередь – угаритский [Translating lost languages…];

– в университете Оксфорда, где разработана программа для чтения древнегреческих текстов, высеченных на камнях (эпиграфика) с возможностью восстановления утраченных символов – «Пифия» (Pythia). Апробация программы показала, что «Пифия» совершает ошибки почти в два раза реже, чем специалисты по древнегреческой эпиграфике (30,1 % – «Пифия», и 57,3 % – человек) [Assael, et al., 2022].

Особый интерес вызывает проект распознавания древних рукописей MOJIZO (URL: http://mojizo.nabunken.go.jp/) который осуществляется в Японии, Корее и Китае. Цель проекта − оцифровка и распознавание очень распространенных на Дальнем Востоке в середине I тыс. н. э. иероглифических текстов моккан (木簡), которые написаны «ханджа» (китайскими иероглифами) на деревянных дощечках. Моккан является в некотором смысле промежуточным вариантом между алфавитным письмом и изображением.

Для расшифровки этих текстов еще в 2016 г. японскими специалистами был предложен аналог нейросети, реализованный на основе наборов данных MNIST, разработанных Яном ЛеКуном (Yann LeCun), автором многих продуктивных подходов в области глубокого машинного обучения.

В качестве наборов данных для обучения выступают две постоянно пополняемые базы данных: база данных Национального исследовательского института культурных ценностей Нара, которая содержит изображения моккан, найденные в Японии (примерно 1 800 знаков, 14 000 деревянных табличек и 89 600 символьных изображений), и база данных Историографического института Токийского университета, в которую включены различные изображения досовременной японской письменности (примерно 6 тыс. письменных знаков и изображения из 230 тыс. знаков).

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

На основе этих разработанных и апробированных методов в конце 2019 г. в Японии был открыт проект KuroNet (KuroNet Kuzushiji Ninshiki Sābisu (KuroNetくずし字認識サービス, URL: http://codh.rois.ac.jp/kuronet/; https://mp.ex.nii.ac.jp/kuronet/) – бесплатная онлайн платформа OCR-распознавания, которая позволяет пользователям конвертировать изображения документов, написанных на иероглифическом японском языке, в печатный текст.

Близкий по идеологии проект начали реализовывать в Центре восточных рукописей и ксилографов Института монголоведения, буддологии и тибетологии СО РАН (Улан-Удэ), где российские специалисты смогли «прочитать» (дешифровать) с помощью AI 500 страниц тибетских рукописей с точностью до 94 % распознаваемых символов, однако, с учетом всех особенностей тибетской письменности, правильность текстов в данный момент оценивается примерно в 80 % [Сегеда, 2022].

Однако распознавание, извлечение и преобразование информации, содержащейся в архивных документах, с помощью OCR и технологий DS в настоящее время активно дополняется другими методами. Например, методом обогащения данных. В качестве примера такой работы можно упомянуть проект EHRI (European Holocaust Research Infrastructure − Европейская инфраструктура исследования Холокоста, URL: https://www.ehri-project.eu/), в котором автоматически синхронизированные архивные метаданные и описания из нескольких источников дополняются пользовательскими метаданными и тематическими «виртуальными коллекциями», что придает новое информационное качество архивным материалам, значительно обогащая их и дополняя их контекст.

Не менее любопытным является проект американских исследователей [Grieggs, et al., 2021], которые объединили традиционные методы машинного обучения с визуальной психофизикой − методом измерения связей между физическими стимулами и психическими явлениями, такими как время, которое требуется для эксперта, чтобы распознать конкретный символ в рукописи, оценить качество почерка или определить использование определенных аббревиатур. Исследователи изучали оцифрованные латинские рукописи, которые были написаны в скриптории швейцарского монастыря Святого Галла в IX в. Цель проекта заключалась в разработке нового метода автоматизированного распознавания (транскрипции) и обеспечения быстрого «чтения» с возможностью поиска, имитирующего восприятие страницы с текстом глазами эксперта. Эксперты видели электронную копию документа, вводили свои транскрипции в специально разработанный программный интерфейс, а затем исследователи измеряли

время реакции во время ввода транскрипции, чтобы понять, какие слова, символы и отрывки были для экспертов максимально легкими или наоборот – трудными. Включение такого рода данных в программы машинного обучения позволило создать нейросеть, более соответствующую человеческому поведению, уменьшило количество ошибок и обеспечило более точное, более реалистичное чтение и расшифровку текста.

Преобразование аудио/видеодокументов и изображений

Проблемы преобразования информации статичных и динамических изображений, а также аудиодокументов в наборы данных носят более сложный характер. В рамках данного направления выполнено относительно немного исследований, однако они производят огромное впечатление.

Одним из наиболее интересных проектов, осуществленных библиотеками на основе работы с изображениями (иллюминатами) является проект Национальной библиотеки Франции «Mandragore» (URL: https://api.bnf.fr/recherche?f[0]=sources:198), который осуществляется с конца 1980-х гг. и к объектам которого были применены методы DS. Основу проекта составляет коллекция 205 тыс. иллюстраций из почти 7 тыс. рукописей, хранящихся в отделе рукописей и библиотеке Арсенала. Каждый идентифицируемый изобразительный элемент (иллюминат) индексируется с использованием многоязычного словаря, включающего в себя более чем 20 тыс. терминов. Кроме того, каждый иллюминат сегментирован на электронной копии с помощью программного обеспечения VGG Image Annotator (пример сегментирования элементов изображений − URL: https://api.bnf.fr/fr/mandragore-echantillon-segmente-2019), к нему созданы метаданные, и он имеет собственный API, что позволяет найти данное изображение в информационной системе (URL: http://mandragore.bnf.fr/html/accueil.html), а также предоставить пользователям все метаданные иллюминатов в виде наборов открытых данных, опубликованных на ресурсах цифровой библиотеки Франции Gallica в разных форматах (в частности, в форматах JSON, CVS и IIIF (International Image Interoperability Framework – Международная платформа совместимости изображений, URL: https://iiif.io/).

В рамках реализации проекта анализа иконографического содержания иллюминированных рукописей французскими исследователями были применены методы машинного обучения и искусственного интеллекта для автоматической индексации содержимого изображений (CBIR, поиск изображений на основе контента). В результате про-

веденного эксперимента был получен аннотированный тематический набор данных изображений на тему зоологии (URL: https://api.bnf.fr/fr/mandragore-jeu-dimages-annotees-sur-le-theme-de-la-zoologie), состоящий из 23 633 изображений, аннотированных 52 336 дескрипторами. Создание этого набора продемонстрировало возможности использования методов DS для индексации изображений на основе контента.

Проект по интеллектуальному поиску изображений реализуется и в государственной библиотеке в Берлине. Реализованный подход позволяет просматривать все иллюстрации по загруженному эталонному изображению, отображает связанные с темой изображения, а также позволяет осуществлять выбор области изображения и отдельные фрагменты, содержащие этот объект (URL: https://qurator.ai/innovationlab/staatsbibliothek-zu-berlin/).

Огромную работу по изучению возможности применения методов DS проводят зарубежные архивы аудиовизуальной информации. Среди этих проектов – VIVA (Визуальный поиск информации в видеоархивах, URL: https://projects.tib.eu/en/viva/projekt/), цель которого заключается в разработке инновационного программного обеспечения для видеомайнинга: используя искусственный интеллект, ПО автоматически индексирует оцифрованные архивные кадры, тем самым создавая новые возможности для поиска информации о сценах, людях и подобных изображениях в пока недокументированном материале.

Возможности AI не ограничиваются только датафикацией и методами анализа существующих разновидовых исторических источников. В современном мире исследователи активно используют методы DS для реконструкции и моделирования различных явлений, проблем, фактов, объектов, событий и их перспективного развития, и т. п.

Пример формирования базы знаний (набора данных) на основе архивной документации для последующего применения методов машинного обучения и AI в реконструкции и моделировании кинодокументов продемонстрировали китайские специалисты.

В связи со 100-летием Компартии Китая Федеральное архивное агентство РФ передало китайским коллегам копию цветного фильма о дне торжественного основания КНР [俄国家电视台], снятого советскими кинооператорами в 1949 г. [Федеральное архивное агентство, 2021]. Катушки фильма считались утраченными и были обнаружены в РГАКФД [开国大典的彩色胶片]. Все кадры переданной копии фильма (12 мин. 36 сек – огромный объем информации!), были оцифрованы китайскими специалистами в максимально высоком разрешении, а колористическая информация кадров составила базу знаний цветов и оттенков, которыми характеризовался этот исторический период в Китае. На основе созданной базы знаний и AI была осуществлена ве-

рификация колористического решения художественного фильма «Момент финальной битвы», снятого в 2019–2020 гг. (проще говоря, AI «перекрашивал» те цвета в кадре, которые не соответствовали базе), а использованные в фильме архивные черно-белые фотодокументы были «с нуля» раскрашены AI. Таким образом была осуществлена реконструкция и колоризация исторического прошлого.

Другой пример – из Европы, где по заказу Венского университета на основе единственной черно-белой фотографии с помощью AI и машинного обучения были воссозданы (реконструированы) и «раскрашены» утраченные во время Второй мировой войны картины Г. Климта – три полотна, представляющие Медицину, Философию и Юриспруденцию. Результат представлен на цифровой выставке, где можно познакомиться не только с картинами, но и историей их восстановления [Klimt vs. Klimt].

А 9 октября 2021 г. состоялась премьера исполнения дописанной искусственным интеллектом 10-й симфонии Бетховена, как известно, незавершенной композитором. Авторы этой работы, которая осуществлялась в австрийском институте Караяна (URL: https://karajan-institut.org), использовали генеральную совокупность произведений Бетховена, представленную в виде коллекции нотаций и законченных композиций, а также его эскизы к 10-й симфонии. С помощью AI было написано около 20 минут музыкального звучания, которые, по оценке специалистов, близки стилю Бетховена [Elgammal, 2021].

Самостоятельное направление в контексте Data wrangling составляют работы, связанные с преобразованием звуковой информации и, прежде всего, речи. За последний год зарубежными компаниями было сделано много заявлений о подобных сервисах. В частности, как утверждают специалисты из Facebook, они разработали самую крупную в своем роде модель автоматического распознавания речи (ASR), которая научилась понимать слова на 51 языке после обучения на более чем 16 тыс. часах голосовых записей. В своей статье авторы указывают, что система, которая содержит около миллиарда параметров, улучшает производительность распознавания речи до 28,8 % на одном бенчмарке по сравнению с базовыми показателями [Pratap, et al.]. При этом апробация модели в части работы с редкими языками проводилась на татарском языке и суахили. Видеопрезентация данного исследования размещена в сети [Facebook…].

Однако разработанные коммерческие решения не всегда подходят для архивов, поэтому многие зарубежные архивные учреждения делают попытки самостоятельно разрабатывать программное обеспечение [Cecchine, 2021a; 2021b].

Analysis (Анализ данных)

Это, пожалуй, один из самых насыщенных публикациями разделов DS, содержание которого условно можно разделить на две части. Первая является промежуточной, поскольку сочетает в себе методы подготовки данных и их анализа. Вторая представляет аналитические исследования, выполненные на уже готовых наборах данных.

Data wrangling+analysis

В рамках данной части активно развиваются несколько направлений.

Первое – создание различных онлайн-платформ, предназначенных для извлечения данных из оцифрованных коллекций исторических источников и анализа полученных наборов. В настоящее время в сети представлен большой спектр подобных разработок, однако срок их «жизни» невелик и часто обусловлен рамками участия авторов в конкретных финансируемых проектах.

Другим важным и интересным направлением исследований и разработок являются проекты, выполненные в рамках Архивной вычислительной науки (Computational Archival Science, CAS, URL: https://ai-collaboratory.net/cas/) – дисциплины, которая сочетает в себе подходы Big Data и методы DS в части AI и машинного обучения для создания максимально полных и релевантных баз данных на основе архивных коллекций. Прекрасным примером этого направления являются проекты Morgenthau Holocaust Collections (URL: https://www.fdrlibrary.org/morgenthauproject) и «Наследие рабства» (URL: http://slavery.msa.maryland.gov).

В контексте CAS большое место занимают проекты, выполненные в методологии «связанных данных» (linked data). Среди них стоит упомянуть проект «Голландские корабли и моряки» (URL: http://www.ghhpw.com/ships-and-sailors.php), который был разработан с целью совершенствования метода выявления связей между объектами в тематически различных наборах данных (набор описаний кораблей, набор данных о голландских морских событиях и т. п.), подготовленных на основе разных источников (архивных документов и исторических газет Национальной библиотеки Нидерландов [Balado, et al., 2014]). В проекте сочетались методы подготовки наборов данных, машинного обучения с оценкой, которая проводилась в ручном режиме, для выявления связей между двумя наборами данных, что доказывает не только возможность извлечения значительного количества релевантных связей между двумя наборами данных, но и то, что эти связи предоставляют новые возможности для анализа исходных наборов данных.

Успешный проект развивается в Финляндии. Речь о платформе «WarSampo» – «Вторая мировая война в Финляндии» (URL: https://www.sotasampo.fi/en/). Инициативный проект объединил 19 распределенных наборов данных, относящихся к Финляндии периода Второй мировой войны, которые представлены в виде связанных открытых данных и сервисов для проведения исследований и создания приложений по истории войны [Hyvönen, et al., 2016; Koho, et al., 2021]. Платформа WarSampo предоставляет доступ тем, у кого нет навыков работы со связанными данными, что позволяет выполнять поиск и просмотр на основе девяти различных «ракурсов», включая событие, место, человека, заключенного и воинскую часть. Частями платформы являются самостоятельные порталы BiographySampo, BookSampo, CultureSampo, NameSampo, TravelSampo, WarSampo, на которых были апробированы и реализованы сервисы онтологии, платформы публикации связанных данных и семантической сети, посвященной культурному наследию [Hyvönen, 2020]. К этим порталам обращаются миллионы пользователей, что повышает их ценность [Xia, et al., 2020].

Еще одним наглядным примером могут служить работы Шанхайской библиотеки, которая осуществила ряд проектов с использованием широкомасштабной оцифровки средневековых книг по генеалогии, OCR, датафикации, метода связанных данных и их обогащения, реализовав крупномасштабный ресурс по генеалогии – Китайскую платформу службы знаний по генеалогии (Jiapu, URL: https://jiapu.library.sh.cn/#/).

Примерно такой же по масштабам проект осуществляется на основе комплексов оцифрованной архивной документации коренных и испано-колониальных народов и испанской колониальной администрации, отложившихся в архивах Мезоамерики (Unlocking The Colonial Archive, URL: https://unlockingarchives.com/).

В завершение рассмотрения этой подтемы применительно к архивам и библиотекам необходимо сказать о совместном проекте Российского государственного военного архива и Венгерского национального архива, начатом в 2019 г. и уже имеющего представление в сети – «Пленные и интернированные венгры в лагерях ГУЛАГа» (URL: https://adatbazisokonline.hu/gyujtemeny/szovjetunioba-elhurcoltak). В БД представлены 681 955 записей и 1 363 910 изображений оцифрованных учетных карточек, что позволило использовать технологии OCR, машинное обучение и нейролингвистическое программирование для автоматизированного перевода записей из переданной БД на русском языке (URL: https://mnl.gov.hu/mnl/ol/hirek/elerheto_a_szovjetunioba_elhurcoltak_adatbazisa) и методы связанных данных. Все вместе позволило создать и опубликовать на сайте открытые наборы данных, востребованные исследователями.

Еще раз подчеркнем, что популярности и устойчивости этим проектам придает выполнение нескольких фундаментальных условий: активное использование исторических источников (архивных документов и библиотечных коллекций); участие в реализации подобных проектов архивистов, библиотекарей, исследователей гуманитарного профиля, хорошо знающих предметную область, и специалистов-информатиков; использование широкого круга технологий и методов, что превращает проект в полидисциплинарное исследование, обладающее потенциалом вторичного использования (т. е. одним из итогов проекта является создание информационного ресурса).

Изучая тему применения DS для создания наборов данных и их анализа, нельзя обойти вниманием работы, связанные с музейными артефактами.

Так, специалисты Института AI Стэнфордского университета, используя набор данных из более чем 81 тыс. картин и 440 тыс. человеческих реакций, разработали алгоритм искусственного интеллекта ArtEmis (URL: https://www.artemisdataset.org/), который научился предсказывать, как человек может отреагировать на произведение искусства, а затем обосновывать это предсказание словами [Achlioptas, et al.].

Исследователи из междисциплинарной группы под руководством физиков из Case Western Reserve University в Кливленде, штат Огайо осуществили разработку инновационной технологии атрибутирования живописи, позволяющей на основе анализа «мазка» (трехмерного скана фрагмента картины размером не более 1 мм, что является know-how данного проекта) с последующим обучением машины и применением сверхточных сетей установить авторство живописного полотна с вероятностью до 96 % [Sutton, 2022].

И, наконец, нельзя не упомянуть поражающий своим размахом проект Рейксмузеума в Амстердаме, который с помощью AI восстановил в первоначальном размере «Ночной дозор» Рембрандта (URL: https://www.rijksmuseum.nl/en/stories/operation-night-watch/story/artificial-intelligence).

Вместе с тем применение AI для изучения музейных предметов, и в частности, объектов живописи, не так однозначно. К примеру, осенью 2021 г. в профессиональном интернет-издании The Conversation была опубликована дискуссионная статья специалиста по средневековому изобразительному искусству Сони Дриммер о том, действительно ли AI помог в изучении Гентского алтаря [Sabetsarvestani, et al., 2021], восстановлении «Ночного дозора» Рембрандта, картин Пикассо и т. п., как об этом пишут разные издания и специалисты, полагаясь на заявления сотрудников компании Oxia

Palus, принимавшей участие в работе над этими проектами, или это лишь отвлекающий маневр, прикрывающий амбициозными проектами менее благостные направления применения AI [Drimmer]. Представляется, что опасения, высказанные автором, небеспочвенны, а вопрос о ценности AI как метода исследований в области истории искусства и искусствоведения сформулирован достаточно жестко и требует осмысления.

Analysis (Анализ подготовленных данных)

Эта подраздел Analysis не обеспечен большим количеством публикаций, и обнаружить среди множества опубликованных работ исследования, выполненные непосредственно в архивах, музеях и библиотеках на основе уже готовых, сформированных наборов данных, не так просто. Между тем, электронные каталоги библиотек, архивов и музеев и иные комплексы развернутых описаний объектов хранения представляют собой ценные массивы машиночитаемых данных, пригодных для анализа. Приведем только один пример подобной работы. В Национальном архиве Великобритании было проведено исследование каталожных описаний архивных документов Генерального совета здравоохранения (ок. 89 тыс. документов, описанных по отдельности), созданного правительством Великобритании после эпидемии холеры 1848–1849 гг. и функционировавшего как орган надзора за санитарными мерами по всей стране. Опираясь на тестовый набор из 1 967 подробных описаний архивных документов, исследователи применили тематическое моделирование, основанное на алгоритме, называемом «латентное размещение Дирихле» (LDA) и предназначенном для обнаружения абстрактных тем в коллекциях документов и сортировки их по нескольким группам, заданным исследователем. В результате исследования была, например, выявлена такая тема, как санитарные инспекции, которые занимали центральное место в деятельности Генерального совета здравоохранения в течение первого года его существования [Day].

Archiving and Using
(архивирование и повторное использование)

Данное направление заслуживает отдельной большой статьи, поскольку именно оно определяет будущее цифрового мира в вопросах хранения и использования информации. Но, к сожалению, на сегодняшний день именно в нем накопилось самое большое количество проблем, значительная часть которых не связана с аналитическими методами и технологиями DS, поскольку, как справедливо отмечают

многие специалисты, «AI может выполнить свое обещание сделать цифровые архивные коллекции более доступными, но это также создает новые проблемы, особенно с точки зрения этики [Jaillant, Caputo, 2022]».

Упомянутые проблемы не останавливают архивы, библиотеки и музеи от осуществления пока немногочисленных проектов по изучению возможностей современных технологий и поиска решения проблем хранения и обеспечения доступа к цифровой документной информации.

Одним из наиболее известных проектов по изучению возможностей блокчейна[1] и AI применительно к хранению архивных born-digital документов является проект ARCHANGEL (URL: https://www.archangel.ac.uk/), цель которого – гарантировать целостность и подлинность цифровых документов и укрепить к ним доверие, исключив такие наиболее важные недостатки цифровых архивов, как зависимость от эфемерных форматов файлов (т. е. фактически сделана попытка купировать необходимость конвертации файлов из формата в формат). В качестве элемента AI применяется байесовская сеть – графическая модель, представляющая набор переменных и условий, которые часто используются для вероятностного анализа. Сеть «снимает и хранит» «отпечаток пальца» (fingerprint, хеш-код) каждого блока документа. Одновременно в проекте была разработана модель оценки рисков цифрового архивирования (DiAGRAM), чтобы предоставить систему поддержки принятия решений, способную количественно оценивать риски и выгоды возможных вмешательств и помогать расставлять приоритеты для инвестиций [Barons, et al., 2021].

Подходы, заложенные в проекте ARCHANGEL, проходят апробацию в национальных архивах Великобритании, Австралии, Норвегии и Эстонии, а также в NARA (США), и не только для born-digital документов. К примеру, чтобы обеспечить целостность и достоверность электронных копий архивных документов и ограничить риск манипуляций с ними, NARA включил хэш-коды в метаданные электронных копий архива убийства Джона Ф. Кеннеди (URL: https://www.archives.gov/research/jfk).

[1] Блокчейн – серия блоков цифровых данных, хранящихся в базе данных, которую несколько организаций могут поддерживать, проверять, делиться и добавлять, но, что наиболее важно, данные в блокчейне нельзя изменить. Неизменность блокчейна подтверждается децентрализованным характером хранения блоков данных. Если делается попытка изменить какие-либо данные, это должно быть подтверждено другими блоками в цепочке, что делает редактирование практически невозможным. Иными словами, фокус доверия перемещается с отдельных содержательных блоков на сеть членов (организаций, узлов), которые принимают «участие» в создании документа и их действия в отношении «своего» блока, которые теперь необходимы для достижения консенсуса перед добавлением блока в цепочку [The ODI].

Среди проблем, решение которых с помощью DS активно обсуждается в профессиональном архивном сообществе, следует упомянуть:

– проблему архивов электронной почты и AI;

– возможности использования связанных данных и семантических сетей для целей архивирования и использования born-digital документации;

– тему применения AI для отбора born-digital документов для длительного хранения.

Вместе с тем следует подчеркнуть, что архивисты весьма сдержанны в оценке возможностей методов DS и AI и предпочитают философию «малых проектов», которые позволят получить максимально сфокусированный и точный ответ на вопрос о применении технологий DS в архивной практике [Sheridan, 2021; Johnston, 2021].

* * *

Завершая обзор наиболее ярких исследований в области применения DS в архивной, библиотечной, музейной сферах, а также в исторических исследованиях, результаты которых опубликованы в различных источниках, необходимо подчеркнуть, что данные технологии становятся все более востребованными, а их внедрение – все более скорым. Это наблюдение влечет за собой главный вывод – о необходимости подготовки профессиональных кадров, владеющих как знаниями предметных областей, так и информационных технологий и готовых к постоянно меняющемуся цифровому миру, в котором DS будет играть одну из ключевых ролей.

Сегеда Г. Тибетская мудрость в цифровом формате // Наука в Сибири. 2022. 7 февраля. URL: https://www.sbras.info/articles/nauka-dlya-obschestva/tibetskaya-mudrost-v-cifrovom-formate (дата обращения: 15.04.2022).

Федеральное архивное агентство информирует. 01.07.2021. // Федеральное архивное агентство : офиц. сайт. URL: https://archives.gov.ru/press/01-07-2021-china.shtml?fbclid=IwAR3OyYi6kgXZfm6wPSwtX5R7F3ub_1Q_QKNHWwerFaj1fJLlxzCMIuzaag (дата обращения: 15.04.2022).

Юмашева Ю. Ю. Историческая наука, архивы, библиотеки, музеи и искусственный интеллект: что день грядущий нам готовит? // Документ. Архив. История. Современность. Вып. 21. Екатеринбург: Изд-во Урал. ун-та, 2021. С. 247–279.

Achlioptas P., Ovsjanikov M., Haydarov K., Elhoseiny M., Guibas L. ArtEmis: Affective Language for Visual Art // Stanford University, LIX, Ecole Polytechnique, King Abdullah University of Science and Technology (KAUST). URL: https://arxiv.org/abs/2101.07396 (дата обращения: 15.04.2022).

AI: A Museum Planning Toolkit // Goldsmiths, University of London. 2020. URL: https://themuseumsainetwork.files.wordpress.com/2020/02/20190317_museums-and-ai-toolkit_rl_web.pdf (дата обращения: 15.04.2022).

Artificial Intelligence Index Report 2022 // Stanford Institute for Human-Centered Artificial Intelligence. URL: https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf (дата обращения: 15.04.2022).

Assael Y., Sommerschield T., Prag J. Restoring ancient text using deep learning: a case study on Greek epigraphy. URL: https://www.deepmind.com/publications/restoring-ancient-text-using-deep-learning-a-case-study-on-greek-epigraphy (дата обращения: 15.04.2022).

Assael Y., Sommerschield T., Shillingford B., et al. Restoring and attributing ancient texts using deep neural networks // Nature. 2022. No. 603. Pp. 280–283. URL: https://www.nature.com/articles/s41586-022-04448-z; https://doi.org/10.1038/s41586-022-04448-z (дата обращения: 15.04.2022).

Balado A. B., de Boer V., Schreiber G. Linking Historical Ship Records to a Newspaper Archive // Aiello L. M., McFarland D. (eds) Social Informatics: SocInfo. 2014 Workshops. Barcelona, Spain, 2014. 11 November, Revised Selected Papers. Cham: Springer. Pp. 254–263.

Barons M., Bhatia S., Double J., Fonseca T., Green A., Krol S., Merwood H., Mulinder A., Ranade S., Smith J. Q., Thornhill T. Safeguarding the nation's digital memory: towards a Bayesian model of digital preservation risk // Archives and Records. 2021. No. 42(1). Pp. 58–78.

Bikakis A., Hyvonen E., Jean S., Markhoff B., Mosca A. Editorial: special issue on semantic web for cultural heritage // Semant Web. 2021. No. 12(2). Pp.163–167.

Büttner G. Auto-Classification in an International Organization: Report from a Feasibility Study // Comma. 2017. No. 2. Pp. 15–26. URL: https://doi.org/10.3828/comma.2017.2.2. (дата обращения: 15.04.2022).

Case Study 1: The National Archives (UK) / Dr. Lise Jaillant, Dr. Katherine Aske, Dr. Annalina Caputo (Eds) // AEOLIAN. URL: https://www.aeolian-network.net/wp-content/uploads/2021/11/AEOLIAN-Case-Study-1-The-National-Archives-UK.pdf (дата обращения: 15.04.2022).

Cecchine R. AI in audiovisual archives: what kinds of analysis are possible? 2021. 21 May // Beeld & Geluid. URL: https://beeldengeluid.nl/en/knowledge/blog/ai-audiovisual-archives-what-kinds-analysis-are-possible (дата обращения: 15.04.2022) (a).

Cecchine R. Bringing ai into the audiovisual archive. 2021. 14 June // Beeld & Geluid. URL: https://beeldengeluid.nl/en/knowledge/blog/bringing-ai-audiovisual-archive (дата обращения: 15.04.2022) (b).

Coeckelbergs M., van Hooland S. Concepts in Topics. Using Word Embeddings to Leverage the Outcomes of Topic Modeling for the Exploration of Digitized Archival Collections // Mugnaini R. (ed). Data and Information in Online Environments. DIONE 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 319. Springer, Cham. https://doi.org/10.1007/978-3-030-50072-6_4 (дата обращения: 15.04.2022).

Colavizza G., Ehrmann M., Bortoluzzi F. Index-Driven Digitization and Indexation of Historical Archives // Frontiers in Digital Humanities. 2019. No. 6 (March). URL: https://doi.org/10.3389/fdigh.2019.00004. (дата обращения: 15.04.2022).

Conférence les futurs fantastiques 2021 // The National library of France (BnF). URL: https://www.bnf.fr/en/les-futurs-fantastiques (дата обращения: 15.04.2022).

Day C. Computing Cholera? «Distant Reading» General Board of Health catalogue data // The National Archives. URL: blog.nationalarchives.gov.uk/computing-cholera-topic-modelling-the-catalogue-entries-of-the-general-board-of-health/ (дата обращения: 15.04.2022).

Drimmer S. How AI is hijacking art history. 2021. 1 November // The Conversation. Academic rigour, journalistic flair. URL: https://theconversation.com/how-ai-is-hijacking-art-history-170691 (дата обращения: 15.04.2022).

Elgammal A. How a team of musicologists and computer scientists completed Beethoven's unfinished 10th Symphony. 2021. 24 September // The Conversation. Academic rigour, journalistic flair. URL: https://theconversation.com/how-a-team-of-musicologists-and-computer-scientists-completed-beethovens-unfinished-10th-symphony-168160 (дата обращения: 15.04.2022).

Expert.ai's Natural Language Search Augments EBSCO Information Services' Existing Search and Discovery Technology // Cision. PR Newswire. URL: https://www.prnewswire.com/news-releases/expertais-natural-language-search-augments-ebsco-information-services-existing-search-and-discovery-technology-301293601.html (дата обращения: 15.04.2022).

Explore how a computer sees art // The Harvard Art Museums. URL: https://ai.harvardartmuseums.org/about (дата обращения: 15.04.2022).

Facebook разработал систему распознавания речи для редких языков! // Facebook.com. URL: https://www.facebook.com/watch/?v=347074320108504 (дата обращения: 15.04.2022).

Grieggs S., Shen B., Rauch G., Li P., Ma J., Chiang D., Price B., Scheirer W. Measuring Human Perception to Improve Handwritten Document Transcription // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021. No. 1. URL: DOI: 10.1109/TPAMI.2021.3092688 (дата обращения: 03.08.2022).

Hyvonen E. Linked open data infrastructure for digital humanities in Finland // Reinsone S., Skadina I., Baklāne A., Daugavietis J. (eds). Proceedings of the digital humanities in the Nordic Countries 5th Conference, 21–23 October 2020, online. Aachen: CEUR. Pp. 254–259.

Hyvonen E., Heino E., Leskinen P., Ikkala E., Koho M., Tamper M., Tuominen J., Makela E. WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History // Sack H., Blomqvist E., d'Aquin M., Ghidini C., Ponzetto S., Lange C. (eds). The Semantic Web. Latest Advances and New Domains. ESWC 2016, May 29 – June 2, 2016. Heraklion, Crete. Greece. LNCS. Vol. 9678. Cham: Springer. Pp 758–773. URL: https:// doi. org/ 10. 1007/ 978-3- 319- 34129-3_ 46 (дата обращения: 15.04.2022).

ICOM Korea – National Museum of Korea to hold the 1st World Museum Forum. 19.11.2020 // ICOM. https://icom-aspac.mini.icom.museum/icom-korea-national-museum-of-korea-to-hold-the-1st-world-museum-forum/ (дата обращения: 15.04.2022).

IFLA Statement on Libraries and Artificial Intelligence // IFLA. URL: https://www.ifla.org/wp-content/uploads/2019/05/assets/faife/ifla_statement_on_libraries_and_artificial_intelligence.pdf (дата обращения: 15.04.2022).

Ingram W. A., Johnson S. A. Ensuring Scholarly Access to Government Archives and Records. Report of project activities from February 2020 through October 2021 // URL: https://vtechworks.lib.vt.edu/bitstream/handle/10919/108067/IngramMellonFinalReport.pdf?sequence=2&isAllowed=y (дата обращения: 15.04.2022).

Introducing The Latest Tool In Art Authentication: AI // Jing Culture & Commerce. URL: https://jingculturecommerce.com/art-recognition-ai-art-authentication/ (дата обращения: 15.04.2022).

Jaillant L., Caputo A. Unlocking digital archives: cross-disciplinary perspectives on AI and born-digital data // AI & Soc. 2022. URL: https://doi.org/10.1007/s00146-021-01367-x (дата обращения: 15.04.2022).

Johnston L. Achieving the Promise of AI with Modest Approaches to Machine // Unlocking our Digital Past. Workshop 2. 16 September 2021. URL: https://unlockingourdigitalpast.com/2021/09/21/leslie-johnston-achieving-the-promise-of-ai-with-modest-approaches-to-machine-learning-unlocking-our-digital-past/ (дата обращения: 15.04.2022).

Király P. Adat a könyvtárban (Data in the Library) // Gesellschaft für wissenschaftliche Datenverarbeitung mbH. Göttingen (GWDG). Göttingen, Németország. 2018. URL: https://www.researchgate.net/publication/328887550_Adat_a_konyvtarban (дата обращения: 15.04.2022).

Klimt vs. Klimt. Project // Google's Art and Culture. URL: https://artsandculture.google.com/project/klimt-vs-klimt; https://artsandculture.google.com/story/bgXxLsdwpiFriQ (дата обращения: 15.04.2022).

Koho M., Ikkala E., Leskinen P., Tamper M., Tuominen J., Hyvonen E. WarSampo knowledge graph: Finland in the second world war as linked open data // Semant Web. 2021. No. 12(2). Pp. 265–278. URL: https:// doi. org/10. 3233/ sw- 200392 (дата обращения: 15.04.2022).

Lemieux V. The Future of Archives as Networked, Decentralized, Autonomous and Global // Archival Futures / ed. by C. Brown. London: Facet Publishing, 2018. Pp. 33–44.

Moss M., David Th., Gollins T. The Reconfiguration of the Archive as Data to Be Mined // Archivaria. 2018. No. 86 (Fall). Pp. 118–151 (дата обращения: 15.04.2022).

Nix A. Finding light in dark archives: Using AI to connect context and content in email. URL: https://www.aura-network.net/wp-content/uploads/2021/04/Adam-Nix-slides.pdf (дата обращения: 15.04.2022).

Persepolis Fortification Archive // The Oriental Institute of the University of Chicago. URL: https://oi.uchicago.edu/research/projects/persepolis-fortification-archive?fbclid=IwAR3JdmbcHF_ybhyUijqO_EzECOdtGYDDo_KEFTW84klzPMAte2aJNYx77jI; https://news.uchicago.edu/story/how-ai-could-help-translate-written-language-ancient-civilizations?fbclid=IwAR08X5IZNrqCNrIFW7QFJMrZjCrWiWR-lt_kfvMjpBTdaWIl-zmsJ475Wxk (дата обращения: 15.04.2022).

Pratap V., Sriram A., Tomasello P., Hannun A., Liptchinsky V., Synnaeve G., Collobert R. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. URL: https://arxiv.org/pdf/2007.03001.pdf?fbclid=IwAR2d9m45bOJVvnHLYLCn7wbKcP-jYCsz1dga5dgBudpUtk20UX02J7ATolg (дата обращения: 15.04.2022).

Ranade S. Traces through Time: A Probabilistic Approach to Connected Archival Data // 2016 IEEE International Conference on Big Data (Big Data), 3260–65. Washington DC, USA: IEEE. URL: https://doi.org/10.1109/BigData.2016.7840983 (дата обращения: 15.04.2022).

Sabetsarvestani Z., Sober B., Higgitt C., Daubechies I., Rodrigues M. R. D. Artificial intelligence for art investigation: Meeting the challenge of separating x-ray images of the Ghent Altarpiece // Science Mag. Org. URL: https://advances.sciencemag.org/content/5/8/eaaw7416?fbclid=IwAR1nPJXyB5ln-ruEv5JCJW_mCXSJE_839o9wBfHO7HKI7fEK4dGCK4yWhs8 (дата обращения: 15.04.2022).

Sheridan J. Archives in the midst of the AI revolution // Unlocking our Digital Past. Workshop 2. 16 September 2021. URL: https://unlockingourdigitalpast.com/2021/09/21/john-sheridan-archives-in-the-midst-of-an-ai-revolution-unlocking-our-digital-past/ (дата обращения: 15.04.2022).

Southerton C. Datafication. // Schintler L., McNeely C. (eds). Encyclopaedia of big data. Springer,Cham. 2020. URL: https://doi.org/10.1007/978-3-319-32001-4_332-1 (дата обращения: 15.04.2022).

Spencer R. Binary Trees? Automatically Identifying the Links between Born-Digital Records // Archives and Manuscripts, 2017. No. 45(2). Pp. 77–99. URL: https://doi.org/10.1080/01576895.2017.1330158. (дата обращения: 15.04.2022).

Stanford University. Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence; Report of the 2015 // Study Panel. 2016. URL: https://ai100.stanford.edu/2016-report (дата обращения: 15.04.2022).

Sutton B. Researchers train AI to attribute paintings based on detailed brushstroke analysis // The Art Newspaper. 4 Jan. 2022. URL: https://www.theartnewspaper.com/2022/01/04/artificial-intelligence-attributes-paintings-brushstroke-analysis (дата обращения: 15.04.2022).

The National Archive. The Application of Technology-Assisted Review to Born-Digital Records Transfer, Inquiries and Beyond. 2016. URL: https://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf (дата обращения: 04.08.2022).

The ODI. How can smart contracts be useful for business? URL: https://theodi.org/wp-content/uploads/2018/05/378720579-How-can-smart-contracts-be-useful-for-businesses.pdf (дата обращения: 15.04.2022).

The Recommendation on the Ethics of Artificial Intelligence // UNESCO. URL: https://en.unesco.org/artificial-intelligence/ethics#recommendation (дата обращения: 15.04.2022).

Translating lost languages using machine learning // MIT News. URL: https://news.mit.edu/2020/translating-lost-languages-using-machine-learning-1021?fbclid=IwAR146qcC-ocbHP11StjhgYl315nZ-_GRtpDlXZSDbE7RXqhLsTyi0piSHXw (дата обращения: 15.04.2022).

Venkata S. Interview for the AURA project (Archives in the UK/ Republic of Ireland and AI), 21 May, 2021 // AURA Network. URL: https://www.aura-network.net/ (дата обращения: 15.04.2022).

Wilde Max De, Hengchen S. Semantic Enrichment of a Multilingual Archive with Linked Open Data // Digital Humanities Quarterly. 2017. No. 11(4). (дата обращения: 15.04.2022).

Xia C., Bao X. Dynamic digital humanities projects from Shanghai library in China // McAleer M., Deyrup M. M. (eds). Transformative digital humanities: challenges and opportunities. Ch. 6. Abingdon: Routledge? 2020. URL: https://www.taylorfrancis.com/chapters/edit/10.4324/9780429399923-9/dynamic-digital-humanities-projects-shanghai-library-china-cuijuan-xia-xue-ming-bao (дата обращения: 15.04.2022).

俄国家电视台，首次播放新中国成立彩色纪录片《中国的重生》 // Bilibili. URL: https://www.bilibili.com/video/BV1ME411Q7AJ?spm_id_from=333.905.b_72656c61746564.4&fbclid=IwAR3OyYi6kgXZfm6wPSwtX5R-7F3ub_1Q_QKNHWwerFa-j1fJLlxzCMIuzaag (дата обращения: 15.04.2022).

开国大典的彩色胶片 – 来自俄罗斯的珍贵礼物. 28.05.2021 // Consulate-General of the People's Republic of China in Rio de Janeiro. URL: https://www.fmprc.gov.cn/ce/cgrj/chn/ztzl/C1/t1879194.htm?fbclid=IwAR2LArf606kPJJTkFF5-LCE7Y-36Q65p3N3LnFJbCF417DE_A7O0sIgXzcZs (дата обращения: 15.04.2022).