РАСКРЫТИЕ ЦИФРОВЫХ АРХИВОВ: МЕЖДИСЦИПЛИНАРНЫЙ ПОДХОД К ПРОБЛЕМЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И ИЗНАЧАЛЬНО ЦИФРОВЫХ ДАННЫХ

UNLOCKING DIGITAL ARCHIVES: CROSS-DISCIPLINARY PERSPECTIVES ON AI AND BORN-DIGITAL DATA*

Лиз Джайан (Lise Jaillant), *Университет г. Лафборо, Лафборо, Великобритания (l.jaillant@lboro.ac.uk)*
Анналина Капуто (Annalina Caputo), *Центр ADAPT, Городской университет Дублина, Дублин, Ирландия (annalina.caputo@dcu.ie)*

Реферат. Авторы статьи, специалисты в области вычислительной техники (Анналина Капуто) и цифровых гуманитарных наук (Лиз Джайан), рассматривают актуальные в эпоху цифровых технологий проблемы учреждений, хранящих предметы культурного наследия, а именно закрытие подавляющего большинства архивных коллекций, содержащих изначально цифровые материалы. Особое внимание уделяется учреждениям культуры — библиотекам, музеям и архивам, к которым обращаются историки, литературоведы и другие ученые-гуманитарии. Доступ к большинству изначально цифровых документальных сведений, хранящихся в культурных организациях, закрыт в силу требований соблюдения конфиденциальности и авторских прав, а также по причине коммерческих и технических проблем. Даже в тех случаях, когда изначально цифровые данные являются общедоступными (например, веб-архивы), возможность ознакомиться с веб-страницами предоставляется пользователям лишь при условии личного присутствия в помещении учреждения, например Британской библиотеки или Национальной библиотеки Франции. Однако наличие достаточного объема выборочных данных для изучения и обучения моделей позволяет использовать технологии искусственного интеллекта и, в частности, алгоритмы машинного обучения для улучшения и упрощения доступа к цифровым архивам, научив машины выполнять комплексные человеческие задачи. Они варьируются от обеспечения интеллектуальной поддержки в поиске по архивам до автоматизации утомительных и трудоемких операций. В данной работе обсуждается возможность проверки информации на конфиденциальность как практическое решение, позволяющее учреждениям разблокировать цифровые архивы и предоставить доступ к информации, не являющейся конфиденциальной. Однако перспектива сделать архивы более доступными содержит потенциальные опасности, а именно: неизбежные ошибки, подходы по принципу «черного ящика», использующие непонятные алгоритмы, риски, связанные с предвзятой, неверной или неполной подачей информации. Основной вывод авторов статьи заключается в том, что реализация потенциала искусственного интеллекта может сделать цифровые коллекции архивных материалов более доступными, создавая при этом новые проблемы, особенно с точки зрения этики. В заключительной части работы авторы указывают на важность приверженности принципам справедливости, подотчетности и прозрачности в процессе расширения доступности цифровых архивов.

Ключевые слова: архивы изначально цифровых материалов, искусственный интеллект, конфиденциальность, авторское право, проверка на конфиденциальность, этика.

Введение

Специалисты в области вычислительной техники (Анналина Капуто) и цифровых гуманитарных наук (Лиз Джайан) рассматривают актуальные в эпоху цифровых технологий проблемы учреждений, хранящих предметы культурного наследия, а именно закрытие подавляющего большинства архивных коллекций, содержащих изначально цифровые материалы. Особое внимание уделяется учреждениям культуры — библиотекам, музеям и архивам, к которым обращаются историки, литературоведы и другие ученые-гуманитарии. Доступ к большинству изначально цифровых документальных сведений, хранящихся в культурных

* https://link.springer.com/article/10.1007/s00146-021-01367-x

организациях, закрыт в силу требований соблюдения конфиденциальности и авторских прав, а также по причине коммерческих и технических проблем. Даже в тех случаях, когда изначально цифровые данные являются общедоступными (например, веб-архивы), возможность ознакомиться с веб-страницами предоставляется пользователям лишь при условии личного присутствия в помещении учреждения, например Британской библиотеки (BL) или Национальной библиотеки Франции (BnF). Однако наличие достаточного объема выборочных данных для изучения и обучения моделей позволяет использовать технологии искусственного интеллекта (AI, ИИ) и, в частности, алгоритмы машинного обучения для улучшения и упрощения доступа к цифровым архивам, научив машины выполнять сложные человеческие задачи. Они варьируются от обеспечения интеллектуальной поддержки в поиске по архивам до автоматизации утомительных и трудоемких задач.

Авторы исследования уделяют особое внимание проверке информации на конфиденциальность как практическому решению проблемы раскрытия цифровых архивов. В заключительной части работы подчеркивается важность соблюдения принципов справедливости, подотчетности и прозрачности в процессе расширения доступности цифровых архивов.

Как обеспечить более свободный доступ к цифровым материалам? И какова роль ИИ в разблокировке закрытых для пользователей «темных» архивов? Золотой век громоздких архивов на физических носителях остался позади: электронные письма пришли на смену обычным, документы в форматах PDF или Word заменили бумажные отчеты, а вместо тяжелых многотомных изданий пользователи обращаются к онлайн-энциклопедиям. Поскольку изначально цифровые архивы по умолчанию размещаются на цифровых носителях, их обслуживание, на первый взгляд, кажется более дешевым, а обеспечение доступа к ним — более простым. Однако неуклонное расширение интернет-пространства, которое сначала развилось до сети контекста с семантической сетью и контекстуальным контентом, а далее до концепции Интернета вещей, включая весь контент, генерируемый мириадами небольших «умных» устройств, привело к массовому порождению изначально цифровых данных.

Создание архивов раньше являлось прерогативой организаций, которые располагали ресурсами и временем для ведения подобной деятельности. По сложившейся традиции государственные архивные учреждения собирали правительственные и административные документы. Национальный архив Франции, например, был создан в 1790-х гг. для размещения архивов центральных учреждений, закрытых в результате Французской революции, а также архивов церковных учреждений Парижской епархии. В XIX в. к этому списку добавились архивы министерств. Лишь после Второй мировой войны возникли новые области архивирования: архивы предприятий, личные и семейные архивы. Под влиянием феминистических и социально-культурных движений 1960—1970-х гг. во Франции и других странах личные архивы стали более разнообразными по своему характеру, а сфера их охвата расширилась от узкого интереса к деяниям «великих мужей» до документирования достижений женщин и различных меньшинств [1; 2; 3].

С популяризацией персональных компьютеров в 1980-х гг. и мобильных устройств в начале XXI в. человечество начало создавать огромные объемы данных. Современные архивы стали более «личными», поскольку, с одной стороны, они отражают эпизоды жизни отдельных людей, а с другой — в них хранится личная информация. Технологические компании, такие как Google и Facebook**, выступают экспертами в области использования накопленных персональных данных для анализа привычек пользователей и прогнозирования их поведения в будущем. Так, С. Зубофф [4] отмечает, что Google и Facebook** являются архетипами «капиталистов-соглядатаев», находящихся в центре экономической системы, которая в одностороннем порядке заявляет, что человеческий опыт — это бесплатное сырье для переработки в поведенческую информацию. По словам автора, наши данные подпитывают экономическую машину, которая мало заботится о частной жизни граждан [4].

Библиотеки и архивы, как некоммерческие организации, выполняют совершенно иные задачи, нежели технологические компании. Подавляющее большинство изначально цифровых сведений, хранящихся в культурных организациях, недоступны в силу требований соблюдения конфиденциальности и авторских прав, а также коммерческих и технических проблем. Даже в тех случаях, когда изначально цифровые данные являются общедоступными (например, веб-архивы), читателям приходится лично посещать учреждения, такие как BL или BnF, с целью ознакомления с веб-страницами. Пользователи, стремящиеся получить доступ к своим личным данным, собранным Google, Facebook** и другими компаниями, часто сталкиваются с многочисленными препятствиями, включая отсутствие ответа на запросы о доступе к данным [5]. Трудно разобраться в том, какую именно информацию коммерческие организации

** Принадлежит компании Meta Platforms Inc, которая решением Тверского суда Москвы от 21.03.2022 признана экстремистской организацией, ее деятельность запрещена на территории Российской Федерации.

собирают о своих пользователях и что они с ней делают. Спрятанные от пользователей цифровые данные стали для подавляющего большинства людей «темными» архивами.

В сложившихся условиях виртуального мира переосмысление и изменение способов обеспечения доступа к информации становится все более важным. Технологии ИИ обещают сделать изначально цифровые архивы более открытыми, например, путем выявления конфиденциальной информации, что позволит соответствующим учреждениям предоставлять доступ к информации, не являющейся конфиденциальной; или помечать документы как релевантные определенному поисковому запросу. Искусственный интеллект — это обширное понятие, обозначающее создание умных устройств, которые могут имитировать мыслительные способности и поведение человека. ИИ включает в себя множество подходов, но именно машинное обучение, одна из областей ИИ, связанная с обучением на основе данных без непосредственного программирования, в настоящее время становится ключевой технологией, потенциально способной произвести революцию во многих секторах. Более того, важность машинного обучения в рамках ИИ настолько велика, что эти два термина зачастую используются взаимозаменяемо. В контексте цифровых архивов эта технология может привести к обучению программ на базе существующих корпусов и созданию аннотированных наборов данных для автоматизации и упрощения сложных задач, таких как ручная проверка конфиденциальных или защищенных авторским правом материалов, а также предоставление поддержки пользователям в поиске и изучении этих архивов. Однако непрозрачные механизмы, с помощью которых используемые алгоритмы обучают модели, должны быть предметом тщательного рассмотрения, иначе недостатки и ошибки в данных, на которых они учатся, могут легко привести к неправильным решениям и искаженному представлению реальности.

Перспектива сделать архивы более доступными не лишена, однако, потенциальных ловушек и рисков. Во-первых, многим алгоритмам присущи неизбежные ошибки. Например, ePADD (программное обеспечение с открытым исходным кодом для управления архивами электронной почты) не всегда способно верно идентифицировать маркированные термины. Так, работая с архивными электронными письмами поэтессы Венди Коуп (Wendy Cope), ведущий куратор отдела современных архивов и рукописей BL Каллум Маккин (Callum McKean) заметил, что ePADD маркирует слово «гриб» в списке покупок В. Коуп[1]. Оказалось, это слово было внесено в словарь программы как «наркотик», что привело к ложному срабатыванию механизма. Во-вторых, предоставление ИИ возможности принимать решения в сложных ситуациях может создать этические и социальные проблемы[2]. Поскольку ИИ действует по принципу «черного ящика», трудно понять, почему машина принимает те или иные решения. Более того, используемые для обучения систем ИИ данные могут быть ангажированными, поскольку, например, представители европеоидной расы мужского пола занимают руководящие посты и лидируют во многих секторах, и документы, относящиеся к женщинам или этническим меньшинствам, могут быть расценены как менее важные. Риски получения предвзятой, фальшивой или неполной информации тесно связаны с внедрением ИИ.

Главная идея настоящей статьи заключается в том, что ИИ может оправдать возложенные на него надежды и сделать цифровые архивные коллекции более доступными, но он также создает новые проблемы, особенно с точки зрения этики. Разработка методов в области «объяснимого ИИ» (который позволяет людям осмыслить результаты, созданные алгоритмами машинного обучения) становится необходимым для понимания того, как машина пришла к конкретным решениям. С целью реализации этих новых задач следует обеспечить сотрудничество между специалистами в области архивной деятельности, цифровых гуманитарных наук и вычислительной техники. Несомненно, ключевые проблемы современности — от глобального потепления до социального неравенства — не могут быть решены в рамках одной дисциплины. То же самое относится и к трудностям, связанным с недоступными данными. Британо-ирландская сеть AURA и британо-американская сеть AEOLIAN (ИИ для культурных организаций) объединяют ученых-гуманитариев, специалистов по информатике, хранителей архивов, библиотекарей и музейных работников[3; 4]. Другие инициативы, такие как AI4LAM, используют более специализированный подход, ориентированный на работу с сотрудниками библиотек, архивов и музеев[5]. Подобные мероприятия показывают, что существует огромный спрос на обмен мнениями по проблеме применения ИИ как в области цифровых архивов, так и конкретных тематических исследований. Несмотря на то что возможности ИИ активно используются в широком спектре областей, в библиотечных и архивных учреждениях соответствующие технологии все еще находятся на экспериментальной стадии, а также отмечается недостаток убедительных тематических исследований [6].

Статья начинается с обзора типичных для архивных учреждений в эпоху цифровых технологий проблем, которые привели к закрытию подавляющего большинства коллекций, созданных в цифровой форме. Особое внимание авторы уделяют культурным организациям, таким как библиотеки, музеи и архивы, к которым обращаются истори-

ки, литературоведы и другие ученые-гуманитарии. Как обеспечить доступ к изначально цифровым материалам в архивах? Как использовать технологии ИИ в архивной деятельности? В статье изучается вопрос о проверке информации на конфиденциальность как практическое решение, направленное на раскрытие цифровых архивов. В заключительной части работы подчеркивается важность соблюдения принципов справедливости, подотчетности и прозрачности в процессе повышения доступности цифровых архивов.

Конфиденциальность и защита данных

Под «темными» архивами понимаются коллекции, скрытые от пользователей по целому ряду причин. Проблемы конфиденциальности и необходимость соблюдения законов о защите данных часто приводят к закрытию цифровых коллекций или к жестким ограничительным мерам, которые резко сокращают количество потенциальных пользователей [7; 8]. В Европе Общий регламент по защите данных 2018 г. дает субъектам данных право на удаление касающихся лица персональных данных без неоправданной задержки, которое, однако, ограничено нуждами архивирования в общественных интересах (Статья 17 Общего регламента). На практике многие архивные учреждения предпочитают закрывать целые коллекции вместо того, чтобы апеллировать к общественным интересам. Аналогичным образом Закон о защите персональных данных 2018 г., который отражает стандарты Общего регламента ЕС в Великобритании и отменяет утратившее силу законодательство о защите данных, часто толкуется очень узко. Так, получить доступ к изначально цифровым данным, хранящимся в британских архивах, часто невозможно, особенно когда речь идет о коллекциях, касающихся ныне живущих людей.

Например, литературоведы и другие пользователи, интересующиеся архивом британской поэтессы В. Коуп (род. в 1945 г.), не располагают доступом к информации из каталога BL «Изучение архивов и рукописей» (Explore Archives and Manuscripts), несмотря на то, что библиотека его приобрела еще в 2011 году. Гибридный архив, содержащий материалы как в бумажной, так и в электронной форме, состоит из 15 больших ящиков для хранения, а также файлов в формате Word. Кроме того, в коллекцию входит большое количество электронных писем, по первоначальным оценкам их было около 40 000 ед. [9; 10]. Рэйчел Фосс (Rachel Foss), заведующая отделом современных архивов и рукописей BL, отмечает, что сбор электронных писем был довольно сложным процессом, учитывая, что они создавались и хранились автором в нескольких папках, с некоторыми дубликатами, невидимыми файлами и т. д. Работа велась за несколько лет до того, как внедрили ePADD, и, не имея возможности использовать подобный инструмент обработки, сотрудники многое делали наугад. Как оказалось, были ошибки в подсчете, поскольку сообщения дублировались в папках. Общее количество писем в коллекции — 25 556 ед.[6] [39].

По просьбе библиотеки В. Коуп проредила свою коллекцию электронных писем, чтобы помочь экспертам с составлением заключения (отбором записей для сохранения). В ходе презентации в 2019 г. К. Маккин сообщил, что BL рекомендовала В. Коуп отобрать материалы, которые она хотела бы им передать, и избавиться от всего того, что она передавать не хотела. Он также добавил, что самой В. Коуп весь этот процесс показался довольно утомительным и трудным, а поскольку архив чрезвычайно объемный и содержит частную информацию, то с ним все равно возникнут серьезные проблемы в области защиты данных[7].

BL проделала большую работу в целях сохранения и обнародования материалов этого архива, но добилась лишь скромных успехов. В статье кураторов из отдела современных архивов Джонатана Пледжа (Jonathan Pledge) и Элеоноры Диккенс (Eleanor Dickens) разъясняются планы библиотеки по предоставлению доступа к избранным цифровым записям из читальных залов [11]. К. Маккин также отметил, что общий план библиотеки заключается в том, чтобы опубликовать материалы электронной почты, но в очень ограниченном режиме. Так, исследователям будет предоставлен доступ к текстовым файлам исключительно в формате PDF/A, по одному файлу за раз, по аналогии с доступом к бумажным материалам[8]. Однако по сей день архив В. Коуп остается закрытым для читателей.

Отсутствие возможности обнаружить коллекцию В. Коуп (без указания в описи фонда при поиске по каталогу) означает, что лишь немногие пользователи осведомлены о том, что этот архив хранится в BL. Информация о предполагаемой дате, когда архив будет опубликован, не сообщается. Перестанет ли коллекция в ближайшее время быть «темной»? И если да, то каковы будут условия доступа? На сегодняшний день у пользователей нет возможности загружать и анализировать данные из архива. Политика предоставления библиотечных материалов (по одному цифровому файлу за раз) наводит на мысль о предупреждении Терри Кука (Terry Cook) о том, что «бумажные умы» не знают, как взаимодействовать с «электронной реальностью» [12]. Иными словами, процессы, используемые для предоставления доступа к бумажным архивам, также применяются к изначально цифровым файлам. Это влияет на методы исследования: например, вдумчивое чтение может использовать-

ся для анализа отдельных архивных электронных писем, однако применение количественных методологий более целесообразно, если исследователи имеют доступ к большим объемам данных.

Необходимость личного посещения читальных залов является еще одним препятствием для доступа, как показало закрытие культурных организаций во время пандемии COVID-19. Даже когда библиотеки и архивы открыты, предоставление документов для ознакомления лишь в стенах учреждения ставит преграды для людей, которые не могут путешествовать по состоянию здоровья, семейным обстоятельствам или в силу отсутствия финансирования. Как и в предыдущем случае, принятые процедуры аналогичны процессу работы с бумажными архивами: поскольку лишь небольшая часть бумажных коллекций была оцифрована по ряду причин (включая дефицит средств, хрупкость документов и размер коллекций), общепринятым подходом является требование очного присутствия пользователя. В редких случаях, когда электронные письма и изначально цифровые материалы становятся доступными, читателям приходится работать с ними на территории учреждения. Например, Библиотека им. Бейнеке Йельского университета предоставляет на месте доступ к избранным электронным файлам из архивной коллекции британского писателя Питера Акройда (Peter Ackroyd). Однако эти документы также можно было бы распространять в Интернете, например, в рамках защищенной онлайн-системы, доступной для зарегистрированных пользователей.

Архивные учреждения часто закрывают целые коллекции из соображений защиты личных данных. В качестве примера можно привести архив электронной почты британского романиста Яна Макьюэна (Ian McEwan), который хранится в Центре им. Гарри Рэнсома в Остине (США), и в настоящее время недоступен исследователям. Вопрос в том, чем вызваны подобные действия организаций. Они искренне обеспокоены правами субъектов данных на неприкосновенность частной жизни или же волнуются о собственной репутации и потенциальном риске судебного преследования? Поучительным представляется сравнение их подходов с практикой таких технологических гигантов, как Google. Компания Google неоднократно подвергалась штрафам за несоблюдение мер безопасности в области защиты личных данных, в том числе в 2020 г. на нее был наложен рекордный штраф в размере 100 млн евро за нарушение Общего регламента ЕС во Франции[9]. До сих пор эти постоянные штрафы мало повлияли на стремление компании Google собрать как можно больше данных о пользователях и извлечь из них выгоду. В то время как Общий регламент ЕС нацелен на минимизацию сбора данных и устанавливает принцип, согласно которому персональные данные должны быть надлежащими, актуальными и ограничиваться тем, что необходимо для достижения целей, для которых они обрабатываются (Статья 5), бизнес-модель Google основана на максимизации данных и претендует на служение общественным интересам, реализуя право быть информированными.

Архивные хранилища также полагаются на максимизацию данных, поскольку их основная функция заключается в сборе, организации, сохранении и обеспечении доступности материалов культурного наследия как для представителей исследовательского сообщества, так и для широкой общественности. Однако они стараются избегать тех рисков, которые готова взять на себя компания Google[10]. В поиске баланса между правами субъектов данных и общественными интересами они отдают предпочтение праву на неприкосновенность частной жизни (и их собственным интересам во избежание репутационных и юридических проблем). Компания Google, напротив, выдвигает на первый план общественные нужды и свободу информации (и свои собственные коммерческие интересы). Мы полагаем, что эти две радикально противоположные точки зрения одинаково ошибочны. Google в значительной степени переняла традиционную роль архивов и библиотек по поиску информации и обеспечению доступа к ней, но слишком часто игнорирует законное право людей на неприкосновенность частной жизни. При этом архивные учреждения пренебрегли законным правом граждан на доступ к информации. Архивы и библиотеки могут сыграть важную роль в противостоянии могущественным технологическим гигантам, однако подобная возможность предполагает использование открытых данных на принципах уважения к соблюдению конфиденциальности. Расширение доступа не означает полное раскрытие любой информации — законно скрывать те части коллекций, которые содержат не подлежащие распространению сведения. Как обсуждается далее в этой статье, ИИ может быть полезным инструментом в реализации трудоемкой задачи по ручной идентификации конфиденциального контента, при этом сохраняется вероятность совершения ошибок. Более широкое определение конфиденциальной информации включает в себя юридические соглашения, секретные или защищенные требованиями к соблюдению тайны сообщения, а также конфиденциальную или этическую информацию [13]. Следовательно, стремление опубликовать наборы данных и сделать их доступными для поиска контрастирует с необходимостью сохранения прав личности на неприкосновенность частной жизни и, в более общем плане, с вопросами национальной безопасности и сферой международных отношений. Тем не менее закрытие целых коллекций на неопределенный период времени неэтично, поскольку архивы в организациях, финансируемых государством,

должны быть открыты для публики. Наконец, это исключает из процесса и лишает влияния архивные учреждения в исторический момент, когда технологические компании захватывают контроль над нашей цифровой памятью, поскольку именно они хранят огромные объемы личных данных (вспомним, например, объем фотографий, собранных Facebook (принадлежит компании Meta Platforms Inc, которая решением Тверского суда Москвы от 21.03.2022 признана экстремистской организацией, ее деятельность запрещена на территории Российской Федерации) [40].

Авторское право

В дополнение к законам о защите личных данных организации в сфере культуры должны соблюдать законодательство об авторском праве, которое также влияет на доступность материалов. В Великобритании издатели обязаны размещать электронные копии своих публикаций в получающих обязательный экземпляр библиотеках, таких как BL, однако доступ к этим работам строго ограничен Положением о библиотеках, собирающих обязательный экземпляр (непечатные произведения) 2013 года. Так, лишь один читатель может одновременно получить доступ к одной и той же электронной публикации. Требования законов об авторском праве зачастую обуславливают необходимость очного присутствия в учреждении для работы с веб-архивами, поскольку библиотеки и архивы не всегда могут отследить правообладателей для получения разрешений на более широкое распространение материалов.

Во Франции доступ к веб-архивам, содержащим ранее общедоступную информацию, предоставляется в читальных залах BnF для зарегистрированных читателей. Получение читательского билета может стать сложным процессом, особенно для тех, кто не принадлежит к определенным категориям (ученые, аспиранты, журналисты и работники учреждений культуры). Требуется личное собеседование с сотрудником библиотеки, чтобы убедиться, что заявитель проводит исследование и имеет законную причину для посещения BnF. Это ограничивает круг посетителей лишь теми, кто имеет возможность приехать в Париж, умеет уверенно изъясняться на французском языке и может представить необходимое обоснование. Чтобы сделать веб-архивы более доступными, BnF разрешает удаленный доступ к своим коллекциям из нескольких библиотек, получающих обязательный экземпляр, за пределами Парижа. Однако провинциальным пользователям по-прежнему требуется лично посетить библиотеку, чтобы посмотреть документы.

Исследователи часто не могут использовать вычислительные методы для анализа изначально цифровых материалов, заблокированных библиотеками и архивами. По словам Джейн Уинтерс (Jane Winters), существуют предпосылки к завышенным ожиданиям [14] со стороны пользователей, которые думают, что им предоставят свободный доступ к данным. Это относится как к обычным читателям, так и к узким специалистам, которые применяют методы интеллектуального анализа текста и данных (TDM) для работы с ресурсами. В презентации 2020 г. под названием «Нет текста, нет анализа текста» Беатрис Алекс (Beatrice Alex) привела примеры трудностей, с которыми сталкиваются ученые при получении доступа к данным[11]. Речь идет, во-первых, о данных, входящих в обширную коллекцию, когда чрезвычайно трудно получить разрешение от отдельных правообладателей; во-вторых, о данных, защищенных авторским правом, когда владелец не желает делиться ими. Следует отметить, что не всегда требуется разрешение правообладателей. С 2014 г. Великобритания сделала исключение из законодательства об авторском праве для проведения TDM в рамках некоммерческих исследований, что позволяет анализировать материалы, защищенные авторским правом [15]. Если исследователь имеет право читать защищенный авторским правом документ в соответствии с условиями заключенного с поставщиком контента лицензионного соглашения, он также имеет право копировать работу в целях проведения некоммерческого TDM. Однако закон по-прежнему запрещает использовать TDM для неопубликованных материалов, например изначально цифровых документов, в библиотеках, получающих обязательный экземпляр, в частности в BL или в Национальной библиотеке Шотландии (NLS) [16]. По словам Сары Эймс (Sarah Ames) и Стюарта Льюиса (Stuart Lewis), подобное ограничительное законодательство создает проблемы для библиотек, поскольку они стремятся сделать коллекции доступными в крупном масштабе [17].

Учреждения культуры обычно отклоняют запросы на использование вычислительных методов для работы с защищенными авторским правом материалами XX в., поскольку владелец либо не желает делиться ими, либо с ним невозможно связаться[12]. Как утверждает Мелисса Террас (Melissa Terras), авторское право породило цифровой век обскурантизма, когда самые мощные инструменты культурного анализа не используются для освещения материалов, опубликованных между 1910 г. и появлением социальных сетей. Автор добавляет, что библиотеки и архивы зачастую чрезмерно осторожны. Она полагает, что категоричное «нет» следует заменить оценкой рисков, поскольку речь идет об осознании учреждениями своего потенциала пойти на риск. Библиотеки должны задаться вопросом: что самое худшее, что может случиться? В Великобритании только на семь библиотек по-

дали в суд, и, если они больше волнуются о репутационном риске, чем о пользе, которую они могут принести своей аудитории, возникает проблема, действительно ли их так беспокоит авторское право, т. е. именно соблюдение прав автора [18].

Таким образом, авторское право можно использовать в качестве предлога для закрытия целых коллекций — вместо того, чтобы оценивать риски возникновения юридических вопросов с правообладателями (которые, как правило, очень низки).

Обеспечение доступа к материалам XX в. является комплексной проблемой, в связи с этим специалистам в области цифровых гуманитарных наук часто имеет смысл сосредоточиться на данных, находящихся в открытом доступе, в частности на публикациях XIX в., которые были изданы в большом количестве и имеют стандартизированную форму. Например, в программе «Жизнь с машинами» (Living with Machines — крупный проект, финансируемый Советом по исследованиям в области искусства и гуманитарных наук Великобритании) используются материалы XIX в., включая газеты, данные переписи населения, карты и другие источники с целью изучения последствий промышленной революции. Междисциплинарная проектная группа разрабатывает новые методы в области интеллектуальной обработки данных[13], чтобы помочь исследователям проводить масштабный анализ этих коллекций [19].

Наряду с другими источниками проект «Жизнь с машинами» опирается на оцифрованные коллекции, а также собрания онлайн-карт, предоставленные NLS, учреждением, которое проделало важную работу по публикации открытых данных в формате, подходящем для многократного использования. Новая программа NLS в области цифровых наук ориентирована на представление оцифрованных коллекций в виде наборов данных, коллекций метаданных, аудиовизуальных материалов, картографических и геопространственных данных, а также сведений, касающихся организационной структуры. Коллекции оцифрованных печатных изданий — это лишь часть гораздо более обширного ландшафта, в рамках которого работает цифровая наука [17]. Эти данные находятся в открытом доступе на онлайн-платформе Data Foundry[14].

Каким образом можно организовать доступ к изначально цифровым материалам, хранящимся в архивах?

NLS предлагает перспективную модель предоставления доступа к оцифрованным материалам в виде наборов данных, а также к изначально цифровым материалам — модель, которую можно воспроизвести в других учреждениях. Возьмем в качестве примера набор данных, касающихся Эдинбургского женского дискуссионного клуба, который является частью оцифрованных коллекций. Собрание включает в общей сложности 16 выпусков двух эдинбургских журналов: The Attempt («Попытка», 1865—1874) и его преемника The Ladies' Edinburgh Magazine («Женский журнал Эдинбурга», 1875–1880). Эти издания были созданы ведущим эдинбургским женским клубом, который существовал с 1865 по 1935 год. Во времена, когда жизнь женщин часто была ограничена частной сферой, клуб давал им возможность обменяться мнениями об образовании, избирательном праве, здравоохранении и социальном обеспечении. В журналах публиковались статьи о доступе женщин к высшему образованию и оплачиваемой работе, избирательном праве и других правах женщин, религии, а также литературная критика, художественная литература и поэзия. В 1936 г., т. е. вскоре после роспуска клуба, библиотека получила копии журналов и рукописные протоколы заседаний общества[15].

Набор данных включает в общей сложности 6354 xml-файла в виде страниц и 6354 файла изображений с файлами метаданных METS для каждого элемента[16]. Метаданные содержат контекстную информацию, включая дату оцифровки элемента, технические материалы, использованные для создания изображений, и метод, выбранный NLS для расшифровки текста (технологию оптического распознавания символов, OCR). Всего коллекция насчитывает 259 829 строк и 2 654 641 слово. У пользователей есть возможность загрузить как коллекцию в полном объеме, так и пробный набор данных для первоначальной оценки.

Большинство прочих ресурсов, доступных на платформе Data Foundry, имеют те же характеристики, что и набор данных Эдинбургского женского дискуссионного клуба. Они были созданы в рамках реализуемой NLS программы массовой оцифровки. Отметим, что стратегическая цель библиотеки состоит в том, чтобы к 2025 г. сделать треть всех фондов доступной в цифровой форме. Данная инициатива находится в русле современной концепции «коллекции как данные», согласно которой организациям культурного наследия желательно представлять свои коллекции в машиночитаемых форматах. Поскольку указанные материалы изначально записаны на бумажных носителях, то здесь не возникает никаких проблем с точки зрения конфиденциальности, защиты данных и авторского права. Неудивительно, что материалы, относящиеся к XIX в. и более ранним периодам, представлены в большом объеме, в то время как количество изначально цифровых ресурсов, напротив, часто лимитировано. На момент написания данной работы (апрель 2021 г.) веб-сайт

Data Foundry не размещал коллекции электронной почты или веб-архивы. Среди немногих доступных изначально цифровых записей отметим сведения административного характера, такие как транзакции на сумму более 25 тыс. фунтов стерлингов или информацию о государственных закупках, доступные в файлах формата CSV.

Кажущаяся скрытость изначально цифровых материалов на платформе Data Foundry резко контрастирует с экспоненциальным ростом этих записей в фонде NLS. В 2020 г. библиотека сообщила, что более 5,2 млн электронных журналов и электронных книг было подано по обязательному непечатному экземпляру в рамках общей библиотечной инфраструктуры [17]. Как и в случае веб-архивов, авторское право является основным препятствием на пути к свободной публикации данных. С закрытыми для исследователей «темными» архивами, с одной стороны, и открытыми платформами, с другой стороны, возникает противоречивая ситуация.

Возможно, пришло время разработать более тонкие модели, которые предоставляли бы доступ пользователям, выполняющим определенные условия. Библиотеки специальных коллекций обычно просят исследователей предъявить удостоверение личности, прежде чем получить доступ к бумажным архивам в читальных залах. Иногда требуются дополнительные документы, в том числе рекомендательное письмо (например, для доступа к особо ценным рукописям BL)[17] и подписанное согласие на соблюдение внутренних правил библиотеки и законодательства в целом. В тех случаях, когда исследователи используют цифровые фотокамеры для копирования материалов, им часто необходимо подписать специальный бланк, подтверждая, что они знакомы с законами об авторском праве и защите данных. Подобные меры вводятся для защиты коллекций и во избежание юридических проблем. Так, если пользователь разместит копии конфиденциальных или секретных архивных материалов в социальных сетях, библиотека сможет доказать, что ответственность лежит на пользователе, а не на учреждении. Библиотеки стремятся продемонстрировать, что они чрезвычайно серьезно относятся к своим обязанностям хранителей рукописей и архивов.

Почему бы не использовать аналогичную систему для предоставления доступа к изначально цифровым материалам, в частности к требующим особых мер защиты вместо того, чтобы скрывать целые коллекции? Пользователи обязаны будут предъявить удостоверение личности и дополнительные документы, а также подписать необходимые формы, прежде чем получить доступ к безопасной онлайн-системе для просмотра записей электронной почты и других изначально цифровых материалов. В идеале исследователям будут выдавать нужные материалы в пользование на несколько дней или недель с возможностью обработки целых массивов данных вместо того, чтобы обращаться к отдельным элементам. Как и в случае с электронными книгами, которые библиотеки выдают во временное пользование через защищенные системы, такие как Acrobat Digital Editions, изначально цифровые записи можно было бы предоставлять для ознакомления через безопасные онлайн-системы. Исследователи в таком случае смогут применять собственные инструменты в рамках системы вместо того, чтобы загружать данные на персональный компьютер. Подобный подход в настоящее время не используется в силу предполагаемых затрат на создание технической инфраструктуры для разработки системы безопасного доступа. Это также потребует трансформации библиотечных операций, которые, как мы видим, по-прежнему сосредоточены на работе с бумажными, а не цифровыми материалами.

Предоставление доступа через безопасную онлайн-систему — это промежуточная модель, которая не всех устроит. Некоторые сторонники движения за открытый доступ скажут, что наименее проблематичные материалы, такие как веб-архивы, должны находиться в свободном доступе без каких-либо ограничений. Национальная библиотека Ирландии уже предоставляет свободный доступ к избранным архивным веб-сайтам. Чтобы защитить себя от возможных претензий со стороны правообладателей и субъектов данных, библиотека сделала следующее заявление о том, что владелец сайта несет ответственность за соблюдение законодательства о защите данных и авторских прав; библиотека архивирует эти материалы в общественных интересах и делает их доступными для исследовательских целей и частного изучения[18]. У сторонников же более ограничительного подхода к доступу возникнут опасения, что ни одна онлайн-система не сможет обеспечить надежную защиту, позволяющую безопасно раскрыть конфиденциальные цифровые сведения. Даже если пользователям не позволят загружать данные на персональные устройства, они все равно смогут делать скриншоты страниц. Некоторые из этих фотографий затем могут быть распространены в Интернете или социальных сетях. Стоимость создания безопасной онлайн-системы также представляет собой трудность, особенно для небольших учреждений, которые и так испытывают сложности с сохранением коллекций изначально цифровых материалов.

Чтобы в некоторой степени разрешить указанные разногласия, можно напомнить заинтересованным сторонам о том, что предоставление ограниченного доступа более предпочтительно по сравнению с полным отсутствием доступа [20]. Движение за открытый доступ сыграло неоцени-

мую роль в продвижении принципов доступности и прозрачности, но оно также вызвало у многих культурных организаций опасения относительно их способности обеспечить наличие соответствующих ресурсов и моделей управления для раскрытия своих коллекций. Оппоненты по обе стороны границы между открытым и закрытым доступом настолько глубоко укоренились в своих убеждениях, что темные архивы, возможно, становятся лишь темнее и темнее. Даже методы, которые считались общепринятыми для печатных собраний (например, указание приблизительной даты, когда коллекция станет доступной), не являются нормой для изначально цифровых коллекций. Значит, пришло время разработать новые подходы.

Для учреждений, которые не могут позволить себе создать онлайн-систему для обеспечения доступа к цифровым материалам, могут быть предусмотрены другие решения, такие как участие в консорциуме, объединяющем библиотеки и архивы (по модели HathiTrust, партнерства академических и исследовательских организаций, предлагающего доступ к миллиону наименований документов, оцифрованных в библиотеках по всему миру)[19]. Создание консорциума с помощью цифровых технологий стало бы долгожданным шагом на пути к более открытым коллекциям. Это также показало бы, что культурные организации могут работать вместе и в общественных интересах создать совместную платформу вместо того, чтобы позволять частным компаниям, таким как Google, доминировать в цифровой среде. Еще в 2004 г. компания Google начала массово оцифровывать книги в рамках проекта «Google Книги», а в случае произведений, уже перешедших в категорию общественного достояния, ввела свои собственные правила загрузки и прочие договорные ограничения. Четыре года спустя HathiTrust был создан как некоммерческий консорциум, призванный сделать оцифрованный контент доступным для максимально широкого круга пользователей. HathiTrust в основном работает с оцифрованными материалами и включает учреждения на территории США. Такое внимание к цифровым материалам, созданным в бумажной форме, объясняет, почему HathiTrust больше заинтересован в решении проблем с авторским правом, нежели с конфиденциальностью и защитой данных (главная проблема в случае непечатных материалов). Деятельность организации регулируется законодательством США об авторском праве и осуществляется под руководством Управления главного юрисконсульта Мичиганского университета[20]. Таким образом, пустует ниша некоммерческого консорциума, занимающегося изначально цифровым контентом и включающего международное сообщество учреждений культуры.

Как может на практике выглядеть работа подобного консорциума, занимающегося изначально цифровыми материалами? Как и в случае с HathiTrust, участники будут делить расходы на операционные услуги и программы[21]. Учреждения будут иметь возможность делиться выбранными цифровыми коллекциями через агрегаторы, т. е. организации, которые собирают данные и делают их доступными на основном веб-сайте консорциума. Подобная модель публикации материалов основана на ведущем проекте Europeana, финансируемом ЕС, этот портал предлагает доступ к миллионам оцифрованных объектов культурного наследия из примерно 4 тыс. учреждений по всей Европе[22]. Пользователи смогут войти в центральную систему консорциума либо в режиме гостя, либо через членство в партнерских организациях. Гости будут иметь возможность осуществлять следующие операции:

- проводить поиск по всему фонду;
- читать, просматривать и загружать контент, выложенный на условиях «полного просмотра»;
- проводить поиск по контенту, выложенному на условиях «ограниченного доступа»;
- получать доступ к контенту, на который нет ограничений, например к работам, доступным на платформе Internet Archive/Wayback Machine, содержащей 475 млрд веб-страниц и другим изначально цифровым материалам[23; 24].

Те, кто входит в систему через партнерское учреждение, получат доступ к более широкому спектру возможностей, включая создание, сохранение и совместное использование коллекций по определенным темам; доступ к контенту, выложенному на условиях «ограниченного доступа», который не может быть свободно распространен в сети Интернет (например, к коллекциям электронной почты). Сначала необходимо будет разработать пилотный проект с целью изучения оптимальных способов предоставления доступа к материалам. Для архивных электронных писем имеет смысл предоставлять доступ к текстовым файлам лишь в формате PDF/A. Отметим, что подобный подход имеет недостатки, поскольку этот формат не сохраняет все характеристики электронных писем (например, тот факт, что электронные письма часто представляют собой цепочки сообщений, а не отдельные тексты). Однако даже небольшой шаг в направлении более широкого доступа к изначально цифровым архивам значительно улучшил бы текущую ситуацию, характеризующуюся отсутствием возможности обнаружения и изучения материалов.

Как можно задействовать возможности ИИ в архивах?

ИИ также можно использовать для поиска релевантного контента. Например, участвующие в судебных процессах юридические фирмы сегодня полагаются на возможности ИИ и регулярно ис-

пользуют инструменты eDiscovery, являющиеся более эффективными по сравнению с традиционными методами поиска по ключевым словам. В дополнение к поиску подтверждающих доказательств, которые используются в ходе судебного процесса с целью доказательства или опровержения версии по делу, eDiscovery также может выявить, были ли доказательства уничтожены или отсутствуют. Ожидается, что объем глобального рынка eDiscovery вырастет с 9,3 млрд долл. США в 2020 г. до 12,9 млрд долл. США к 2025 году[25]. Эти инструменты основаны на прогнозирующем кодировании, форме машинного обучения, которая учится на подмножестве документов, отобранных юристами и адвокатами, а затем применяет полученные знания к значительно более широкому набору документов. Таким образом, разработанные для отбора документов алгоритмы могут далее применяться в работе с огромными наборами данных, что делает процесс проверки более быстрым, дешевым и простым. Использование программного обеспечения для обнаружения электронных данных не требует наличия продвинутых компьютерных навыков, что объясняет популярность указанных инструментов.

Несмотря на то что подобные инструменты могли бы использоваться учеными для определения релевантного контента, не стоит всецело полагаться на готовое коммерческое программное обеспечение. В своем отчете о машинном обучении и его применении в библиотеках Райан Корделл (Ryan Cordell) отмечает, что он глубоко не затрагивает тему инструментов машинного обучения, предоставляемых поставщиками, в первую очередь потому, что не считает, что они соответствуют стандартам открытости, прозрачности и адаптируемости, выработанным в соответствии с методическими рекомендациями. Абигейл Поттер (Abigail Potter) из Библиотеки Конгресса США сделала аналогичное заявление о том, что существует несоответствие между тем, что предлагается, т. е. комплексными решениями или инструментами «черного ящика», и потребностями в области работы с предметами культурного наследия, а именно, прозрачность, оценка, проверка и, возможно, повторная обработка информации [21]. Подобно хранителям архивов, которым необходимо участвовать в процессе вспомогательного просмотра архивных документов, ученым следует лично взаимодействовать с программами машинного обучения таким образом, чтобы понимать, как и почему машина выбрала определенные документы в большом наборе данных.

Использование технологий машинного обучения не обязательно требует специальной подготовки в области теории вычислительных систем. Обратимся снова к веб-сайту Data Foundry, который доступен для широкого круга пользователей, включая тех, у кого нет опыта программирования. Он предлагает не только наборы данных, но и различные инструменты для анализа коллекций. Так, Люси Хейвенс (Lucy Havens) из NLS создала практическое руководство по изучению коллекции Эдинбургского женского дискуссионного клуба с использованием возможностей анализа текста и данных, а также технологий обработки естественного языка (NLP) на базе языка программирования Python[26]. Разработки в области NLP как одно из направлений развития ИИ нацелены на то, чтобы помочь машинам понимать текст и произносимые слова по аналогии с тем, как это делают люди. Примеры задач NLP включают распознавание речи, анализ эмоциональной окраски высказываний и выделение именованных сущностей, которое идентифицирует слова и фразы как полезные объекты (например, географические названия или имена собственные). Л. Хейвенс использовала выделение именованных сущностей для автоматического определения мужских и женских имен. Затем она визуализировала набор данных, чтобы продемонстрировать количество случаев использования имени «Мэри» в течение определенного периода времени. Пошаговый подход позволяет освоить применение NLP для работы с библиотечными коллекциям даже тем пользователям, которые не обладают специальными техническими навыками. Л. Хейвенс также предоставляет консультации по дальнейшему обучению на платформе Library Carpentry и опубликовала книгу по NLP, доступ к которой открыт на условиях лицензии Creative Commons [22; 23]. Стоит отметить, что рассмотренный выше набор данных, касающихся Эдинбургского женского дискуссионного клуба, не защищен авторским правом и не представляет никаких проблем с точки зрения защиты данных и конфиденциальности, при этом анализ других коллекций может оказаться намного сложнее. Например, в электронных письмах используются лексические единицы или термины, понимание которых часто зависит от общего контекста коммуникации. Подобный неформальный и разговорный стиль общения помешает машине корректно идентифицировать релевантные слова или фразы, поэтому процесс всегда должен включать человека, проверяющего достоверность результатов.

ИИ и проверка информации на конфиденциальность

Как было указано выше, проблема конфиденциальной информации или не подлежащих публикации материалов является одной из ключевых причин, по которой многие изначально цифровые коллекции оказываются недоступны и не могут быть обнаружены в ходе поиска. При этом можно использовать возможности ИИ и машинного

обучения для просмотра огромного количества цифровых файлов и выявления проблемных материалов. В презентации 2020 г. Стив Ригден (Steve Rigden), специалист по архивированию цифровых материалов NLS, рассказал о роли ИИ в идентификации конфиденциальных материалов в цифровых коллекциях библиотеки. Он подчеркнул роль хранителей в изучении данных и принятии окончательных решений. Так, именно сотрудники архивов должны идентифицировать наборы данных, определять наиболее эффективные алгоритмы для работы с ними, тестировать и улучшать модели обработки данных, чтобы далее позволить эффективно обучать машины, а также дорабатывать и тестировать внедренные технологии. По словам С. Ригдена, хранителям архивов не требуется разбираться в технических аспектах ИИ, скорее они должны проявлять интерес к разработке подобных инструментов в качестве консультантов и тестировщиков. Иными словами, им необходимо вплотную участвовать в процессе «вспомогательной проверки» архивных документов[27; 28].

Способность обрабатывать и автоматически классифицировать большие объемы данных представляет собой одну из наиболее перспективных возможностей для использования ИИ в работе с архивами изначально цифровых материалов. Вступившие в действие различные законы о защите персональных данных (от Общего регламента ЕС по защите персональных данных до Закона о защите данных Великобритании 2018 г.), конфликтуют с более чем 100 принятыми по всему миру законами и актами о свободе информации, которые были разработаны с целью обеспечения доступа к государственным документам. Следовательно, раскрытие изначально цифровых архивов предполагает, прежде всего, умение правильно идентифицировать и классифицировать информацию, позволяющую определить лицо (PII).

Термин PII относится к любой информации, с помощью которой можно однозначно идентифицировать человека (от имени, номера телефона или паспорта до даты рождения и медицинских данных). PII включает секретную и юридическую информацию, а также данные, полученные в результате исследовательских экспериментов. Определение понятия PII является расплывчатым, так как зачастую речь идет об информации, которую нельзя напрямую приписать физическому лицу, например журналы поиска в сети Интернет и IP-адреса, поскольку даже это может привести к идентификации людей при глубоком подходе к добыче и анализу сведений. Например, Латания Суини (Latanya Sweeney) продемонстрировала, как можно идентифицировать лицо путем установления перекрестных связей между наборами ранее обезличенных данных лишь на основе общих атрибутов, таких как почтовый индекс, дата рождения и пол [24].

Типичные информационно-поисковые системы обрабатывают информацию, пытаясь оптимизировать ее точность и полноту. При этом они руководствуются таким подходом: если предмет является доступным в сети Интернет, то его также можно обнаружить с помощью инструментов поиска [25]. Правильный баланс между открытостью и безопасностью может быть достигнут путем пересмотра этого подхода к обработке и потреблению данных. «Защита и поиск» и «поиск и защита» — это две возможные парадигмы решения существующей проблемы. Однако, как отмечают специалисты [25], следует защитить конфиденциальную информацию не только от человеческих глаз, но и скрыть ее от поисковой системы. Следовательно, новый взгляд на проблему предполагает разработку структур, которые будут включать в себя принципы релевантности и защиты конфиденциальности информации.

Независимо от выбранной парадигмы, мы рассматриваем два аспекта проблемы конфиденциальности при обеспечении доступа к изначально цифровым архивам: идентификация и количественная оценка конфиденциальной информации.

Идентификация

Во-первых, необходимо определить, какая именно информация является конфиденциальной. Эта задача может быть широкой, например общая классификация текстов, или очень конкретной, т. е. точечное выделение частей текстов, которые могут раскрывать конфиденциальную или личную информацию. Классификация текстов — это классическая и хорошо изученная область применения технологий машинного обучения [26], в рамках которой набор извлеченных из документов отличительных признаков используется для прогнозирования классов или категорий документа. Наподобие работы с учителем, здесь модель обучается, просматривая предварительно аннотированный набор документов, которым был присвоен правильный ярлык того или иного класса. Обучение модели обычно предполагает изучение функции отображения между входными данными (признаками, представляющими документы) и выходными данными (классами). Подобная функция может быть либо простой линейной комбинацией плотности признаков, либо сложной моделью, применяемой в работе глубоких нейронных сетей, с несколькими вложенными слоями функций активации и тысячами параметров. Одним из важнейших аспектов этого класса алгоритмов является определение набора признаков, адекватно фиксирующих корреляцию между входными и выходными данными. В целях классификации текстов признаки обычно извлекаются непосредственно из текста в виде ключевых слов, т. е. отдельных терминов, составляющих текст.

Однако, как отмечает ряд исследователей [27], проверка на конфиденциальность — это задача, ориентированная не на выявление конкретной тематики, где ключевые слова позволяют определить тему документа, а скорее на определение «кто сказал, что и про кого», где для раскрытия конфиденциальной информации, в дополнение к отдельным ключевым словам, важны отношения между терминами и объектами в дискурсе.

Таким образом, набор базовых признаков часто обогащается более сложными признаками, полученными с помощью технологий обработки естественного языка и извлечения информации. Синтаксическая информация (например, обозначающие части речи ярлыки), структурная информация (например, заголовки разделов и таблиц) и целые последовательности слов (также называемые N-граммы) часто используются для выявления композитной конфиденциальной информации, что может быть результатом сочетания нескольких категорий такой информации. Кроме того, чтобы преодолеть двусмысленность языка, где полисемия и синонимия приводят к смещению темы и расхождениям, и в то же время выявить контекстуальную семантическую информацию, можно использовать методы векторного представления или вложения слов на основе моделей дистрибутивной семантики, что позволяет заменить простые ключевые слова или применять два похода параллельно. Дж. Макдональд и др. объединили все указанные выше подходы (термины, ярлыки частей речи, N-граммы и вложения слов) в классификатор SVM и провели анализ коллекции, включающей 3801 государственный документ, отметив каждый из них как «конфиденциальный» или «неконфиденциальный» [27]. Авторы показали, что включение семантических признаков (вложения слов) повышает точность классификатора на 9,99% по сравнению с базовыми подходами.

Классификация не всегда является бинарной (конфиденциальная или не-конфиденциальная информация), и иногда ранжированный подход к определению критериев конфиденциальности лучше соответствует основной задаче. В эмпирическом исследовании, направленном на создание «более четкого понимания концепции служебной тайны, которое может как учитывать различия в методах классификации, так и способствовать более эффективному регулированию» [28], была проведена работа примерно с миллионом дипломатических телеграмм 1970-х гг., классифицированных как «секретно», «конфиденциально», «для ограниченного служебного использования» или «несекретно». При этом «совершенно секретные» документы, количество которых было ограничено, в коллекцию не вошли. Текст документа был обработан стандартными методами разметки и нормализации. Однако в представление была встроена дополнительная информация, отражающая структуру документов и исходные поля признаков (отправитель/получатель, тема, основной текст и т. д.). Авторы провели серию экспериментов с некоторыми стандартными алгоритмами классификации, построенными на взвешенных векторных признаках. Изучение полученных результатов позволило сделать вывод о том, что полезность дат в целях классификации оказалась ограниченной, в то время как слова в поле основного текста были наиболее ценными характеристиками для выявления конфиденциальной информации. В целом наилучшие результаты были достигнуты в тех случаях, когда все признаки использовались в комбинации. Работа с ранжированными классами конфиденциальности также привела еще к одному выводу. Рассмотрение секретных и конфиденциальных документов позволило провести точную классификацию, однако «категория ограниченного служебного использования» выявила менее четкие результаты, что отражает специфику подобных документов, не поддающуюся формальному определению в силу широкого толкования.

Другой подход к проблеме конфиденциальности заключается в редакции конфиденциальной/личной информации [13]. С этой целью можно использовать инструменты криминалистической информатики для автоматического отбора элементов, подлежащих изъятию. Программа с открытым исходным кодом BitCurator, являющаяся результатом разработок по автоматическому редактированию конфиденциальных документов, предлагает функцию массового извлечения, которая лексически анализирует текст в поисках конфиденциальных признаков (адреса электронной почты, номера телефонов и др.).

Несмотря на то что были проведены обширные исследования в области публикации данных на условиях защиты содержащейся в них конфиденциальной информации, большая часть этой работы была направлена на анализ реляционных и статистических данных, текстовые же данные остаются относительно малоисследованной областью [29]. Сосредоточившись конкретно на этом типе данных, Давид Санчес (David Sánchez) и Монтсеррат Батет (Montserrat Batet) предлагают алгоритм обезличивания информации в документах, который имитирует человеческое суждение при оценке параметров конфиденциальности [30]. Инструмент количественно определяет риск раскрытия конфиденциальной информации, представленной набором терминов, посредством логического вывода на основе анализа базы знаний, содержащей конфиденциальную информацию. Как отмечают авторы, идентификация надлежащей базы знаний имеет решающее значение для достижения правильного баланса между спецификой предметной области и обобщенной моделью.

Количественная оценка

Невзирая на впечатляющую точность получаемых результатов, алгоритмы ИИ для проверки информации на конфиденциальность, работающие по принципу как классификации, так и редактирования, не застрахованы от сбоев. Во многих случаях, когда речь идет об общей классификации, возможные ошибки не представляют серьезную проблему. Однако в силу требований законодательства и соответствующих тяжких последствий потенциальных нарушений, возникающих в результате непреднамеренной утечки секретных и персональных данных, необходимо проявлять большую осторожность при классификации конфиденциальной информации.

Оценочный анализ извлекаемой информации характеризуется сложившейся традицией, в рамках которой были выработаны парадигма и показатели, направленные на количественную оценку возможностей систем в области поиска и извлечения релевантной информации. Показатели качества извлеченной информации обычно оцениваются с точки зрения ее точности и полноты. Однако проверка на конфиденциальность предполагает учет иных соображений, помимо релевантности документов, и подразумевает необходимость найти баланс между обеспечением доступа к информации и рисками, связанными с раскрытием конфиденциальной информации, вплоть до оценки последствий наиболее пессимистичных сценариев [25].

Достичь идеальной точности невозможно. Существует ошибочное представление о том, что выполненное человеком аннотирование представляет собой золотой стандарт и верхнюю границу того, чего могут достичь алгоритмы ИИ. Однако эта идея может быть оспорена в силу лимитированного объема информации, которую специалисты способны просмотреть, а также вероятности человеческой предвзятости в отношении репрезентации информации и когнитивного анализа, что может привести к нерациональной классификации конфиденциальных материалов. Предпочтительным подходом может стать сотрудничество человека и машины, при котором алгоритмы не заменяют экспертов, а лишь повышают эффективность их работы. Так, в 2020 г. было проведено исследование [27] по проблеме влияния алгоритмов автоматической классификации на работу специалистов в области проверки информации на конфиденциальность. Авторы проанализировали влияние точности и уровня достоверности прогноза на количество документов, которые оценивают люди, и на время, которое они тратят на формулировку суждения. Выводы авторов подчеркивают ценность проверки информации на конфиденциальность с помощью цифровых технологий для повышения как скорости проверки, так и количества обработанных документов. Во-первых, конфиденциальные документы требовали больше времени для анализа по сравнению с не-конфиденциальными. Во-вторых, точность и скорость работы специалиста значительно увеличились при использовании методов автоматического прогнозирования (+37,9 и +72,2% соответственно). В-третьих, уровни достоверности классификации оказывали влияние на достижение согласия между специалистом в области проверки информации на конфиденциальность и классификатором текстов: консенсус приводил к более быстрому принятию решения, а несогласие сказывалось на производительности труда специалиста.

Альтернативный подход состоит в том, чтобы включить анализ, выполняемый специалистами, в цикл обучения машин, т. е. рассматривать разработки в области машинного обучения не как одноразовый продукт, а как цифровых помощников, которые учатся бок о бок с людьми. Подобная адаптивность особенно актуальна в ситуациях, когда типы конфиденциальной информации заранее неизвестны [31]. Реализация концепции проверки с применением цифровых технологий может быть достигнута за счет внедрения стратегий активного обучения. Отправной точкой является начальный набор, полученный путем ручного аннотирования фонда документов, удовлетворяющих заданному запросу. Подобный начальный набор используется для обучения алгоритма на базовой стадии. Далее система запускает цикличный процесс, в ходе которого генерируются новые прогнозы для нового набора документов без меток, которые затем передаются специалистам, чтобы они вручную присвоили им новые метки. Таким образом, исходный обучающий набор расширяется за счет новых примеров, что запускает новый цикл обучения. Конкурентное преимущество данного подхода заключается в уменьшении количества маркированных документов, необходимых для достижения той же эффективности классификации [31], что особенно важно, когда подобные технологии внедряют в новых коллекциях.

В целом изложенные результаты говорят о том, что использование технологий стало нормой. Даже если полностью автоматизированного решения не существует и на всех этапах по-прежнему требуется участие человека, Национальный архив Великобритании указывает на важность проверок с помощью технологий как способа понять, оценить и расставить приоритеты в отношении изначально цифровых документов, а также уменьшить количество материалов, которое необходимо просматривать вручную [32].

Заключение

В данном документе подчеркивается идея о том, что ИИ может реализовать свой потенциал и сделать цифровые архивы более доступными, но он также создает потенциальные этические проблемы.

Несмотря на то что ИИ помог добиться существенного прогресса в таких областях, как обработка естественного языка, машинное распознавание образов, машинный перевод и пр., что невозможно было бы сделать без использования масштабных наборов данных, на основе которых обучают и совершенствуют эти модели, существуют неоспоримые риски, связанные с подобными слепыми источниками данных. Наиболее ярким примером является предвзятость в репрезентации концепций, в силу которой такие связки, как «мужчина — программист, а женщина — домохозяйка», могут лишь усиливать стереотипы, содержащиеся в собранных данных [33]. Кроме того, искажение взглядов меньшинств и представителей различных социальных движений может привести к неверным логическим заключениям и в итоге негативно повлияет на процесс принятия решений. И.С. Джо (E.S. Jo) и Т. Гебру (T. Gebru) обращаются именно к архивам и библиотекам как институтам, использующим четкий терминологический аппарат и устоявшиеся процедуры сбора данных, бросая вызов историческим и репрезентативным предубеждениям [34]. Риск слепого подхода к ИИ может свести на нет все эти усилия. Стимулировать внедрение технологий ИИ должна четкая структура управления, основанная на выверенном языке и процедурах получения согласия, авторитете, инклюзивности, прозрачности, принципах этики и конфиденциальности, а также опирающаяся на обобщенный опыт сотрудников архивов, социологов, историков и антропологов. Раскрытие цифровых архивов требует междисциплинарного сотрудничества и пристального внимания к соблюдению этических норм.

Примечания

1 Интервью с Каллумом Маккином, 28 мая 2021 г. в рамках Проекта AURA: www.aura-network.net (дата обращения: 20.03.2023).

2 Существует много исследований в области ИИ и этики. Полезно начать с определения этики, включающего ряд принципов (прозрачность, справедливость и честность, непричинение вреда, ответственность и конфиденциальность) [35], см. также [36; 37; 38].

3 www.aura-network.net (дата обращения: 20.03.2023).

4 www.aeolian-network.net (дата обращения: 20.03.2023).

5 https://sites.google.com/view/ai4lam (дата обращения: 20.03.2023).

6 Письмо автору от 22 апреля 2021 г. Для получения дополнительной информации о программном обеспечении ePADD см.: https://library.stanford.edu/projects/epadd (дата обращения: 20.03.2023).

7 C. McKean Processing E-mail at the British Library, презентация, семинар «Изначально цифровой архив и цифровая криминалистика — где мы сейчас?». 15 марта 2019 г. Школа перспективных исследований, Лондонский университет, Великобритания.

8 Там же.

9 Délibération SAN-2020–012 (2020), https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000042635706 (дата обращения: 20.03.2023).

10 Райан Корделл (Ryan Cordell) утверждает, что осторожный и постепенный подход к работе с данными позволяет библиотекам избежать деструктивных идеологий технологической экспансии, отдавая приоритет созиданию, а не разрушению [21].

11 B. Alex. No Text, No Text Mining. Семинар AURA, 1 ноября 2020 г., онлайн: https://www.aura-network.net/2020/12/21/workshop-1-bea-alex/ (дата обращения: 20.03.2023).

12 Там же.

13 Интеллектуальная обработка данных предполагает извлечение ценной аналитической информации из данных. Это междисциплинарная область, объединяющая информатику, статистику и другие дисциплины.

14 https://data.nls.uk/ (дата обращения: 20.03.2023).

15 https://data.nls.uk/data/digitised-collections/edinburgh-ladies-debating-society/ (дата обращения: 20.03.2023).

16 METS расшифровывается как Стандарт кодирования и передачи метаданных (разработан Библиотекой Конгресса США).

17 https://www.bl.uk/help/access-manuscripts-and-archives (дата обращения: 20.03.2023).

18 https://www.nli.ie/en/web_archive.aspx (дата обращения: 20.03.2023).

19 https://www.hathitrust.org/ (дата обращения: 20.03.2023).

20 https://www.hathitrust.org/copyright (дата обращения: 20.03.2023).

21 https://www.hathitrust.org/Cost (дата обращения: 20.03.2023).

22 https://pro.europeana.eu/share-your-data/process (дата обращения: 20.03.2023).

23 https://pro.europeana.eu/about-us/mission (дата обращения: 20.03.2023).

24 https://archive.org/about/ (дата обращения: 20.03.2023).

25 https://www.marketsandmarkets.com/Market-Reports/e-discovery-market-11881863.html (дата обращения: 20.03.2023).

26 https://data.nls.uk/wp-content/uploads/2020/10/Exploring_Ladies_Edinburgh_Debating_Society.html (дата обращения: 20.03.2023).

27 S. Rigden. Sensitivity Review and Access to Digital Materials at the National Library of Scotland. Семинар AURA, 1 ноября 2020 г., онлайн: https://www.aura-network.net/2020/12/21/workshop-1-steve-rigden-sensitivity-review-and-access-to-digital-materials-at-the-national-library-of-scotland/ (дата обращения: 20.03.2023).

28 См. также: R. Oliva Understanding Sensitivity: A First Step Towards Automating Sensitivity Review. Конференция «Архивы, доступ и ИИ», январь 2021 г., URL: https://www.poetrysurvival.com/presentation-slides-archives-access-and-ai-conference/ (дата обращения: 20.03.2023).

Список источников

1. *Dumont-Johnson M.* Peut-on faire l'histoire de la femme? // Revue D'histoire De L'amérique Française. 1975. Vol. 29, № 3. P. 421—428.
2. *Quinn P.M.* The Archivist as Activist // Georgia Archive. 1977. Vol. 5, № 1. P. 25—35.

3. *Mason K.M., Zanish-Belcher T.* Raising the Archival Consciousness: How Women's Archives Challenge Traditional Approaches to Collecting and Use, or, What's in a Name? // Library Trends. 2007. Vol. 56. P. 344—359.
4. *Zuboff S.* The Age of Surveillance Capitalism: the Fight for a Human Future at the New Frontier of Power. New York : PublicAffairs, 2019.
5. *Verborgh R.* Getting My Personal Data Out of Facebook. 2019. URL: https://ruben.verborgh.org/facebook/ (дата обращения: 20.03.2023).
6. *Rolan G., Humphries G., Jeffrey L. et al.* More Human than Human? Artificial Intelligence in the Archive // Archives and Manuscripts. 2019. Vol. 47, № 2. P. 179—203.
7. *Jaillant L.* After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives // Archives and Manuscripts. 2019. Vol. 47, № 3. P. 285—304.
8. *Baron J.R., Payne N.* Dark Archives and E-Democracy: Strategies for Overcoming Access Barriers to the Public Record Archives of the Future // Conference for E-Democracy and Open Government, 2017. P. 3—11.
9. *Flood A.* Wendy Cope's Archive Sold to British Library // Guardian. 2011. 20 April. URL: https://www.theguardian.com/books/2011/apr/20/wendy-cope-archive-british-library (дата обращения: 20.03.2023).
10. Some Sort of Record Seemed Vital: British Library Acquires the Archive of Wendy Cope // British Library. 2011. https://www.bl.uk/press-releases/2011/april/some-sort-of-record-seemed-vital-british-library-acquires-the-archive-of-wendy-cope (дата обращения: 20.03.2023).
11. *Pledge J., Dickens E.* Process and Progress: Working with Borndigital Material in the Wendy Cope Archive at the British Library // Archives and Manuscripts. 2018. Vol. 46, № 1. P. 59—69.
12. *Cook T.* Electronic Records, Paper Minds: the Revolution in Information Management and Archives in the Post-Custodial and Postmodernist Era // Archives & Manuscripts. 1994. Vol. 22, № 2. P. 300—328.
13. *Woods K., Lee C.A.* Redacting Private and Sensitive Information in Born-Digital Collections // Archiving Conference. 2015. № 1. P. 2—7.
14. *Winters J.* Coda: Web Archives for Humanities Research — Some Reflections // The Web as History / ed. by Brügger N., Schroeder R. London : UCL Press, 2017. P. 238—248. URL: http://discovery.ucl.ac.uk/1542998/1/The-Web-as-History.pdf (дата обращения: 20.03.2023).
15. Exceptions to Copyright: Research // Intellectual Property Office. 2014. 16 p. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf (дата обращения: 20.03.2023).
16. *Gooding P., Terras M., Berube L.* Towards User-Centric Evaluation of UK Non-Print Legal Deposit: A Digital Library Futures White Paper. 2019 // University of Nebraska-Lincoln. URL: https://digitalcommons.unl.edu/scholcom/180/ (дата обращения: 20.03.2023).
17. *Ames S., Lewis S.* Disrupting the Library: Digital Scholarship and Big Data at the National Library of Scotland // Big Data & Society. 2020. Vol. 7, № 2. P. 1—7.
18. *Mackinlay R.* Why Is Most of the 20th Century Invisible to AI? // Information Professional — CILIP: the Library and Information Association. 19 March 2021. URL: https://www.cilip.org.uk/news/557160/Why-is-most-of-the-20th-Century-invisible-to-AI.htm (дата обращения: 20.03.2023).
19. Living with Machines : Corporate report. Arts and Humanities Research Council, 2020. 7 p. URL: https://www.ukri.org/publications/living-with-machines/ (дата обращения: 20.03.2023).
20. *Jaillant L.* User Experience and Access to Born-Digital Data Produced by Publishers : The Case of Carcanet Press // Books.Files: Preservation of Digital Assets in the Contemporary Publishing Industry / by ed. M. Kirschenbaum et al. College Park, MD, USA : University of Maryland and the Book Industry Study Group, 2020. P. 38—39.
21. *Cordell R.* Machine Learning + Libraries : A Report on the State of the Field. 2020. 91 p. URL: https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig (дата обращения: 20.03.2023).
22. *Alex B., Llewellyn C.* Library Carpentry: Text and Data Mining // Centre for Data, Culture and Society. University of Edinburgh, 2020. URL: http://librarycarpentry.org/lc-tdm/ (дата обращения: 20.03.2023).
23. *Bird S., Klein E., Loper E.* Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit // O'Reilly Media. 2019. URL: https://www.nltk.org/book/ (дата обращения: 20.03.2023).
24. *Sweeney L.* K-anonymity: a Model for Protecting Privacy // International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002. Vol. 10, № 5. P. 557—570.
25. *Olteanu A., Garcia-Gathright J., de Rijke M. et al.* FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval // ACM SIGIR Forum. 2021. Vol. 53, № 2. P. 20—43.
26. *Sebastiani F.* Machine Learning in Automated Text Categorization // ACM Computing Surveys. 2002. Vol. 34, № 1. P. 1—47.
27. *Mcdonald G., Macdonald C., Ounis I.* How the Accuracy and Confidence of Sensitivity Classification Affects Digital Sensitivity Review // ACM Transactions on Information Systems. 2020. Vol. 39, № 1. P. 1—34.
28. *Souza R.R., Coelho F.C., Shah R., Connelly M.* Using Artificial Intelligence to Identify State Secrets. 2016. P. 1—48. URL: https://arxiv.org/ftp/arxiv/papers/1611/1611.00356.pdf (дата обращения: 20.03.2023).
29. *Fung B.C.M., Wang K., Chen R., Yu P.S.* Privacy-Preserving Data Publishing: a Survey of Recent Developments // ACM Computing Surveys. 2010. Vol. 42, № 4. P. 1—53.
30. *Sánchez D., Batet M.* C-Sanitized: a Privacy Model for Document Redaction and Sanitization // Journal of the Association for Information Science and Technology. 2016. Vol. 67, № 1. P. 148—163.
31. *McDonald G., Macdonald C., Ounis I.* Active Learning Stopping Strategies for Technology — Assisted Sensitivity Review // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA : Association for Computing Machinery, 2020. P. 2053—2056.
32. The Application of Technology-assisted Review to Born-digital Records Transfer, Inquiries and Beyond // The National Archives. 2016. URL: https://www.nationalarchives.gov.uk/documents/technology-assisted-review-

to-born-digital-records-transfer.pdf (дата обращения: 20.03.2023).
33. *Bolukbasi T., Chang K.-W., Zou J. et al* Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings // In Proceedings of the 30th International Conference on Neural Information Processing Systems. NY, USA : Curran Associates Inc., Red Hook, 2016. P. 4356—4364.
34. *Jo E.S., Gebru T.* Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY, USA : Association for Computing Machinery, 2020. P. 306—316.
35. *Jobin A., Ienca M., Vayena E.* The Global Landscape of AI Ethics Guidelines // Nature Machine Intelligence. 2019. Vol. 1. P. 389—399.
36. *Hagendorff T.* The Ethics of AI Ethics: an Evaluation of Guidelines // Minds & Machines. 2020. Vol. 30, № 1. P. 99—120.
37. *Coeckelbergh M.* AI Ethics. Cambridge : MIT Press, 2020.
38. *Véliz C.* Privacy Is Power: Why and How You Should Take Back Control of Your Data. London : Bantam Press, 2020.
39. *Schneider J., Adams C., DeBauche S. et al.* Appraising, Processing, and Providing Access to email in Contemporary Literary Archives // Archives and Manuscripts. 2019. Vol. 47, № 3. P. 305—326.
40. *Ovenden R.* Burning the Books: a History of Knowledge under Attack. Cambridge : Harvard University Press, 2020.

Перевод **Марии Федотовой,**
Российская государственная библиотека