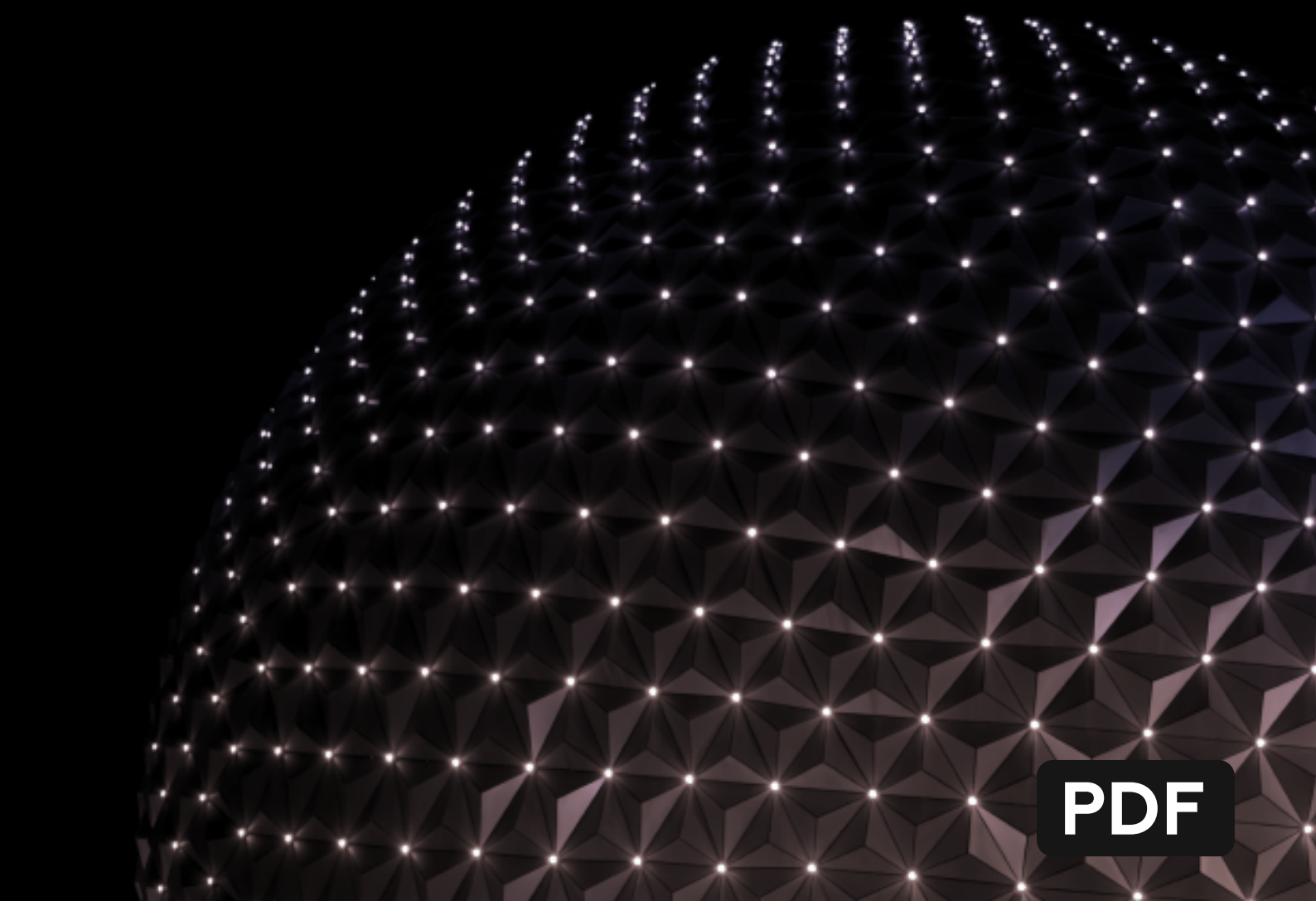
Использование объектов интеллектуальной собственности (базы данных и ее элементов) в машинном обучении



PDF

1. Используемые термины и сокращения

База данных – представленная в объективной форме совокупность самостоятельных материалов (статей, расчетов, нормативных актов, судебных решений и иных подобных материалов), систематизированных таким образом, чтобы эти материалы могли быть найдены и обработаны с помощью электронной вычислительной машины (ЭВМ).

Датасет – набор данных (включая совокупность материалов базы данных), используемый в машинном обучении.

Машинное обучение – процесс, реализующий вычислительные методы, которые предоставляют системам возможность обучаться на данных или на основе опыта.

Скрепинг – технология получения данных путем извлечения их со страниц сайтов в сети Интернет.

Майнинг данных – совокупность методов исследования (включая методы машинного обучения), связанных со сбором и последующей обработкой большого количества данных с помощью автоматизированных программных инструментов с целью обнаружения в данных новых знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Спин-офф – продукт, являющийся побочным результатом осуществления основной деятельности.

РИД и СИ – результаты интеллектуальной деятельности и средства индивидуализации.

TDM (text and data mining) – майнинг текста и данных (в терминологии Директивы ЕС об авторском праве в цифровой среде).

Директива ЕС о базах данных – Директива № 96/9/ЕС Европейского парламента и Совета Европейского Союза «О правовой охране баз данных» (принята в г. Страсбурге 11.03.1996).

Директива ЕС об авторском праве в цифровой среде – Директива № 2019/790 Европейского парламента и Совета Европейского Союза «Об авторском праве и смежных правах на Едином цифровом рынке и о внесении изменений в Директивы № 96/9/ЕС и 2001/29/ЕС» (принята в г. Страсбурге 17.04.2019).

2. Постановка проблемы

В условиях цифровизации широкое развитие получают технологии компьютерного анализа больших массивов данных для получения новых знаний и их применения в разных сферах (государственное управление, научно-исследовательская деятельность, медицина, промышленность, маркетинг и др.).

Правовые аспекты использования данных в целях автоматизированного анализа данных (в т.ч. машинного обучения) включают широкий круг вопросов, затрагивающий не только право интеллектуальной собственности, но и договорное право, конкурентное право, законодательство о персональных данных, законодательство об открытых данных, коммерческой тайне и др. В рамках данного исследования раскрываются аспекты, связанные с правом интеллектуальной собственности.

В случаях, когда датасет (набор данных), используемый в целях машинного обучения, представляет собой объект интеллектуальной собственности либо в его составе имеются объекты интеллектуальной собственности, встают вопросы об основаниях такого использования. Ключевые вопросы сводятся к следующим:

1) **При каких условиях датасет (база данных) признается объектом авторских и (или) смежных прав?**

2) **Является ли использование базы данных и ее элементов в целях машинного обучения использованием объекта (объектов) интеллектуальной собственности?**

3) **В каких случаях использование базы данных и ее элементов в целях машинного обучения может быть свободным, а в каких – осуществляться с согласия правообладателя?**

3. Соотношение категорий база данных и датасет (набор данных)

В машинном обучении анализируются структурированные по каким-либо признакам данные. Такие данные могут составлять как обучающую выборку (в обучении «с учителем» – данные, на которых обучается алгоритм), так и совокупность объектов, в которой находятся зависимости. Данные, используемые при автоматизированном анализе, как правило, называются датасетами (или наборами данных). В ряде правовых и официальных нормативно-технических документов Российской Федерации содержится определение датасета.

В Национальной стратегии развития искусственного интеллекта на период до 2030 года набор данных определяется как совокупность данных, прошедших предварительную подготовку (обработку) в соответствии с требованиями законодательства Российской Федерации об информации, информационных технологиях и о защите информации и необходимых для разработки программного обеспечения на основе искусственного интеллекта[1]. Идентичное определение датасета (в отдельных документах — набора данных) закреплено в ведомственных программах цифровой трансформации[2]. Датасет также определяют как структурированные наборы данных, сформированные для машинного обучения[3].

В национальном стандарте ГОСТ Р 59898-2021 «Оценка качества систем искусственного интеллекта. Общие положения» приводится определение набора данных (dataset) — совокупность данных, в том числе соответствующих им метаданных, организованных по определенным правилам и принципам описания. При этом в стандарте указано, что в зависимости от цели применения набор данных может быть представлен следующими типами данных: текстовыми записями, временными рядами, изображениями, видео, сигналами и т.п.[4]

В российском законодательстве широко используется термин «информационная система», под которым понимается совокупность содержащейся в базах данных информации и обеспечивающих ее обработку информационных технологий и технических средств (п. 3 ст. 2 ФЗ от 27.07.2006 № 149-ФЗ «Об информации, информационных технологиях и о защите информации»). При этом законодательство об информации не содержит определения термина «база данных». Предполагается, что это любая совокупность данных, которая может быть обработана

1. Указ Президента РФ от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года») // СПС КонсультантПлюс.
2. «Консолидированная ведомственная программа цифровой трансформации Министерства труда и социальной защиты Российской Федерации на 2022 год и плановый период 2023 и 2024 годов» (утв. Минтрудом России, ПФ РФ, Рострудом, ФСС РФ 01.02.2022) // СПС КонсультантПлюс; «Ведомственная программа цифровой трансформации Федеральной службы по надзору в сфере образования и науки на 2022 год и плановый период 2023 и 2024 годов» (утв. Рособрнадзором 11.03.2022) // СПС КонсультантПлюс; «О ведомственной программе цифровой трансформации Министерства спорта Российской Федерации на 2022–2024 годы» (утв. Минспортом России 31.12.2021) // СПС КонсультантПлюс.
3. «Ведомственная программа цифровой трансформации Министерства юстиции Российской Федерации на 2022 год и плановый период 2023–2024 годов» (утв. Минюстом России) // СПС КонсультантПлюс.
4. ГОСТ Р 59898-2021 «Оценка качества систем искусственного интеллекта. Общие положения» // СПС «Кодекс».

информационными технологиями (процессами, методами поиска, сбора, хранения, обработки, предоставления, распространения информации) и техническими средствами.

Определение базы данных содержится в законодательстве об интеллектуальной собственности. Под базой данных понимается представленная в объективной форме совокупность самостоятельных материалов (статей, расчетов, нормативных актов, судебных решений и иных подобных материалов), систематизированных таким образом, чтобы эти материалы могли быть найдены и обработаны с помощью электронной вычислительной машины. Такое определение вполне могло бы охватить категорию «датасеты» (есть критерии структурирования/систематизации и обработки с помощью ЭВМ). Вызывает вопросы указание на совокупность самостоятельных материалов (статей, расчетов, нормативных актов, судебных решений и иных подобных материалов). Корректнее было бы указание на совокупность информации, данных или информационных элементов.

Несмотря на то, что в российском законодательстве не сформировался единообразный подход к определению базы данных, датасетов и иных схожих категорий, связанных с обработкой массивов данных, **к датасетам (и их элементам), охраняемым в режиме интеллектуальной собственности, условно можно отнести:**

1) **неохраняемые базы данных с охраняемыми элементами – объектами ИС (например, база записей музыкальных произведений, если расположение материалов в базе данных не носит творческий характер и на ее создание не были понесены существенные финансовые, материальные, организационные или иные затраты);**

2) **охраняемые базы данных с неохраняемыми элементами (например, база нормативно-правовых актов, если расположение материалов в базе данных носит творческий характер и (или) на ее создание понесены существенные финансовые, материальные, организационные или иные затраты);**

3) **охраняемые базы данных с охраняемыми элементами – объектами ИС (например, база литературных произведений, если расположение материалов в базе данных носит творческий характер и (или) на ее создание понесены существенные финансовые, материальные, организационные или иные затраты).**

4. База данных как объект авторских и смежных прав

Как было указано выше, в соответствии со статьей 1260 ГК РФ базой данных является представленная в объективной форме совокупность самостоятельных материалов (статей, расчетов, нормативных актов, судебных решений и иных подобных материалов), систематизированных таким образом, чтобы эти материалы могли быть найдены и обработаны с помощью электронной вычислительной машины (ЭВМ). На базу данных может распространяться авторское право автора составного произведения, а также смежное право изготовителя базы данных.

В базе данных как объекте авторского права охраняются расположение и/или выборка материалов, осуществленные создателем базы данных, если они представляют собой результат творческого труда[5]. Например, расположение элементов базы данных в алфавитном или хронологическом порядке не будет охраняться авторским правом, поскольку такое ранжирование не предполагает творческий характер. Если же элементы базы данных структурированы по критериям, предполагающим творческий подход для их классификации (например, по жанру, стилю, сфере интересов, научным направлениям и т.п.), в таком случае создатель базы данных может претендовать на правовую охрану базы данных как составного произведения.

5. Ворожевич А.С. Границы и пределы осуществления авторских и смежных прав. Москва: Статут, 2020. 271 с. // СПС КонсультантПлюс.

В режиме смежных прав охраняется не творческий вклад, а инвестиции в создание базы данных. В соответствии с российским законодательством смежное право на базу данных предоставляется в том случае, если изготовитель базы данных несет на ее создание (включая обработку и представление материалов) **существенные финансовые, материальные, организационные или иные затраты** (ст. 1334 ГК РФ). При отсутствии доказательств иного базой данных, создание которой требует существенных затрат, признается база данных, **содержащая не менее десяти тысяч самостоятельных информационных элементов** (материалов), составляющих содержание базы данных.

Конкретная база данных может охраняться как в каком-то одном из режимов, так и одновременно в режимах авторских и смежных прав (если охраноспособными являются и творческое расположение материалов, и существенные инвестиции в создание базы данных). На практике наибольшие сложности возникают с признанием смежных прав на базу данных, поскольку не всегда очевиден существенный характер затрат на изготовление базы данных.

Проблема квалификации существенности затрат не имеет однозначного решения. Например, в ЕС известна **доктрина spin-off** (побочного продукта), исключающая возможность учета расходов на общую деятельность лица в качестве инвестиций в создание базы данных (например, при определении существенности инвестиций в создание базы данных не будут учитываться расходы на получение информации о клиентах, ее обработку и т.д., так как это является частью общей деятельности компании, а не специфическими расходами на создание базы данных). Данная доктрина получила развитие в судебной практике ЕС[6]. Использование подобной доктрины должно обеспечивать баланс частных и публичных интересов. С одной стороны, сужение понимания существенности затрат ограничивает монополизацию доступа к информации. С другой стороны, слишком узкий подход к существенности затрат может ущемлять интересы изготовителей баз данных, включая организации в сфере образования, науки, культуры (библиотек, архивов, музеев и др.)[7].

Актуальным вопросом в контексте скрепинга данных (извлечения с интернет-страниц) и их последующего использования в машинном обучении является признание смежного права владельцев интернет-сайтов на базу данных, размещаемых пользователями.

Спор о признании смежных прав владельца социальной сети на базу данных пользователей стал предметом в судебном разбирательстве «Вконтакте» vs ООО «Дабл». Постановление Девятого арбитражного апелляционного суда от 08.07.2021 № 09АП-31545/2021-ГК по делу № А40-18827/2017 признало действия ООО «Дабл» по извлечению и последующему использованию информационных элементов из базы данных пользователей социальной сети «ВКонтакте» нарушением исключительных прав общества с ограниченной ответственностью «В Контакте» как изготовителя базы данных[8] пользователей социальной сети «ВКонтакте»[9].

В другом деле[10] (Avito vs. Auto.ru) суд не стал признавать смежное право на базу объявлений

6. Например, в деле British Horseracing Board Ltd. v. William Hill Organization Ltd суд пришел к выводу, что расходы на создание списка участников забегов не могут рассматриваться как расходы на создание базы данных (в данном случае имеют место расходы на создание самой информации, а не базы данных).
7. По мнению А.И. Савельева, применительно к признанию смежных прав на базу данных необходимо исходить из широкого подхода к характеру инвестиций, поскольку, если применять узкий подход, схожий с тем, который демонстрируется в европейском праве и поддерживается рядом отечественных экспертов, данный правовой режим приобретет очень ограниченный («нишевый») характер. См. Савельев А.И. Гражданско-правовые аспекты регулирования оборота данных в условиях попыток формирования цифровой экономики // Вестник гражданского права. 2020. № 1. С. 60–92.
8. При рассмотрении дела до итогового решения истец («Вконтакте») настаивал, что на создание базы данных были понесены существенные финансовые, организационные и иные затраты, включая затраты на создание и поддержание инфраструктуры (технологическое оборудование, обеспечивающее функционирование серверов – стойки, источники бесперебойного питания, коммутационное оборудование, кабельные системы), закупку необходимого оборудования и серверов, а также затраты на человеческие ресурсы (заработная плата сотрудникам, выплаты по внешним контрактам на обслуживание оборудования и иное). См. Постановление Девятого арбитражного апелляционного суда от 06.02.2018 № 09АП-61593/2017-ГК по делу № А40-18827/17.
9. Однако ответчик («Дабл») настаивал, что наполнение базы данных осуществляется непосредственно пользователями социальной сети, а общество «В Контакте» не производит затрат на собирание элементов базы данных. В данном контексте, как указывало общество «Дабл», база данных пользователей социальной сети является «побочным продуктом» («spin-off») деятельности общества «В Контакте» по администрированию социальной сети. См. Постановление Суда по интеллектуальным правам от 24 июля 2018 г. № С01-201/2018 по делу № А40-18827/2017.
10. Решение Арбитражного суда города Москвы от 19 декабря 2019 г. по делу № А40-183412/2019 // URL: https://sudact.ru/arbitral/doc/4DI2uKUixYU8/ (дата обращения: 29.03.2022).

сайта avito.ru, поскольку «заявление истца не содержит описания каких-либо существенных затрат в отношении непосредственно интернет-сайта «Авито», за исключением затрат на рекламирование самого ресурса «Авито». Как отмечено в решении суда, «сайт «Авито» (как и любой иной аналогичный ресурс) отражает актуальную в настоящий момент времени информацию о предложениях продавцов. Его наполнение и изменение зависит только и исключительно от воли третьих лиц – продавцов подержанных автомобилей[11]. При этом данные действия продавцы совершают без какого-либо участия или согласования со стороны заявителя [владельца ресурса]».

В российской судебной практике не сформировался единообразный подход к квалификации существенности затрат на изготовление базы данных. Причиной этому, среди прочего, можно назвать отсутствие в законодательстве и в отечественной доктрине каких-либо четких критериев доказывания существенности затрат на изготовление базы данных, равно как и критериев опровержения презумпции существенности затрат, закрепленной в статье 1334 ГК РФ.

Изготовителю базы данных принадлежит исключительное право извлекать из базы данных материалы и осуществлять их последующее использование в любой форме и любым способом. Под извлечением материалов понимается перенос всего содержания базы данных или существенной части составляющих ее материалов на другой информационный носитель с использованием любых технических средств и в любой форме.

В запрете на извлечение и использование материалов из базы данных есть исключения (ст. 1335.1 ГК РФ). В частности, лицо, **правомерно пользующееся обнародованной базой данных**, вправе без разрешения обладателя исключительного права – изготовителя базы данных и в той мере, в которой такие действия не нарушают авторские права изготовителя базы данных и других лиц, извлекать из базы данных материалы и осуществлять их последующее использование:

- в целях, для которых база данных ему предоставлена, в любом объеме, если иное не предусмотрено договором;
- в личных, научных, образовательных целях в объеме, оправданном указанными целями;
- в иных целях в объеме, составляющем **несущественную часть базы данных.**

Сложности применения названных исключений связаны с квалификацией «правомерного использования обнародованной базы данных» и «объема, составляющего несущественную часть базы данных». И если вопрос с оценкой «объема, составляющего несущественную часть базы данных» целесообразно решать в каждом конкретном случае с учетом содержания базы данных и характера использования, то квалификация «правомерного использования обнародованной базы данных» требует однозначного разъяснения.

Под лицом, правомерно использующим обнародованную базу данных, можно понимать любое лицо, которое получило доступ к базе данных без нарушения закона (например, не вследствие «взлома» информационной системы, обхода технических средств защиты и т.п.). Возможен и более узкий подход, предполагающий наличие лицензионного договора с правообладателем либо с другим уполномоченным лицом[12]. Целесообразно внести в законодательство положения, уточняющие критерии правомерности использования обнародованной базы данных. Как отмечается в литературе по вопросу правомерного использования базы данных и программы для ЭВМ, подход к определению правомерности использования должен исходить не из необходимости установления субъективных прав на объект интеллектуальной собственности, а из необходимости обеспечения баланса частных и публич-

11. Однако в юридической литературе отмечается, что сам факт участия пользователей в наполнении базы данных интернет-сайта материалами не должен ставить под сомнение принадлежность исключительных прав на базу данных изготовителя баз данных. См. Донников Ю.Е. Наполнение баз данных сайтов в интернете пользователями: правовые аспекты // Юрист. 2021. № 6. С. 68–74.

12. См. Gaster W.La protection juridique des bases de donnees dans l'Union europeenne // Revue du Marche Unique Europeen. 1996. P. 65, 73.; Vanovermeire V. The Concept of the Lawful User in the Database Directive // International Review of Intellectual Property and Competition Law. 2000. Vol. 31. № 1. P. 72–77.

ных интересов в отношении этого объекта[13]. При таком толковании **под правомерным использованием базы данных разумно понимать не только наличие лицензионного договора между правообладателем и пользователем, но и возможность использования базы данных неограниченному кругу лиц (например, размещение правообладателем базы данных в открытом доступе в сети Интернет)**[14].

Примечательно, что исключения из запрета на извлечение материалов базы данных не могут быть ограничены договором между пользователем и правообладателем. Хотя на практике такое правило может вызывать весьма спорные ситуации. Такая ситуация уже имела место в практике ЕС. Так, в деле Ryanair Ltd v PR Aviation BV, рассмотренном Европейском судом справедливости, компания PR Aviation администрировала интернет-сайт с функцией поиска рейсов авиакомпаний. На сайте Ryanair было размещено уведомление, что содержащаяся на ней информация может использоваться только для частного и некоммерческого использования. Суд признал, что ответчик не может ссылаться на случаи свободного использования баз данных, предусмотренных Директивой № 96/9/ЕС, поскольку данная база данных не являлась охраняемой в силу отсутствия достаточного объема инвестиций в ее создание. Парадоксальность вывода заключается в том, что признание базы данных неохраноспособной привела не к отказу в иске, а к его удовлетворению, поскольку имеет место нарушение договора. Если бы база данных была признана охраноспособной, то действия ответчика не являлись бы нарушением, поскольку запрет, установленный договором, был бы недействительным в соответствии со статьей 15 Директивы ЕС о базах данных. В российском законодательстве хоть и не содержится аналогичный прямой запрет на установление договорных ограничений, но положения статьи 1335.1 ГК РФ, закрепляющие действия, не являющиеся нарушением исключительного права изготовителя базы данных, очевидно носят императивный характер.

Таким образом, для формирования сбалансированной системы регулирования в сфере использования объектов интеллектуальной собственности в целях машинного обучения необходимо решение следующих задач в контексте правовой охраны баз данных:

- **уточнить критерии существенности/несущественности затрат на изготовление баз данных с учетом баланса частных и публичных интересов (не ущемляя законных интересов изготовителей баз данных и не способствуя чрезмерной монополизации доступа к данным);**
- **уточнить критерии правомерного использования обнародованной базы данных;**
- **уточнить возможность установления договорных ограничений на использование в целях машинного обучения правомерно обнародованной базы данных.**

5. Майнинг данных и использование объектов интеллектуальной собственности

Под майнингом данных понимается совокупность методов исследования (включая методы машинного обучения), связанных со сбором и последующей обработкой большого количества данных с помощью автоматизированных программных инструментов с целью обнаружения в данных новых знаний, необходимых для принятия решений в различных сферах человеческой деятельности. В контексте сопоставления с использованием объектов интеллектуальной собственности майнинг данных предполагает[15]:

13. Ворожевич А.С. Границы и пределы осуществления авторских и смежных прав. Москва: Статут, 2020. 271 с. // СПС КонсультантПлюс.
14. В Директиве ЕС о базах данных аналогичные исключения, закрепленные в статье 8, распространяются на базы данных, сделанные любым способом общедоступными (database which is made available to the public in whatever manner). См. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases // Режим доступа: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009 (дата обращения: 29.03.2022).
15. Geiger, C., Frosio, G. & Bulayenko, O. Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?. IIC 49, 814–844 (2018). https://doi.org/10.1007/s40319-018-0722-2 (дата обращения: 22.03.2022).

1) идентификацию входных материалов, таких как произведения, отдельные данные или предварительно систематизированные массивы данных;

2) копирование значительного количества материалов для перевода их в определенный машиночитаемый формат, совместимый с технологиями машинного обучения, а в отдельных случаях – для загрузки материалов в отдельную информационную систему (платформу), используемую в машинном обучении;

3) извлечение материалов из баз данных;

4) перестановку материалов для выявления закономерностей и получения выходных данных.

Очевидно, что многие действия, сопровождающие машинное обучение, связаны со способами использования объектов интеллектуальной собственности, которые охраняются в режиме исключительных прав. В зависимости от того, какой метод машинного обучения применяется в каждом конкретном случае, может осуществляться[16]:

- воспроизведение (при загрузке материалов в информационную систему для последующей обработки в рамках машинного обучения);
- извлечение материалов из базы данных (при формировании обучающей выборки из разных источников);
- переработка (преобразование материалов в специальный машиночитаемый формат, перестановка материалов в базе данных и т.п.);
- доведение до всеобщего сведения или распространение (например, при передаче исходных материалов для верификации результатов исследования);
- иные способы использования (например, перевод текста).

В некоторых случаях применение машинного обучения предполагает лишь «сканирование» материалов в открытых источниках с целью поиска конкретной информации или закономерностей в массиве данных. Такой метод машинного обучения может обойтись без использования объектов ИС в тех пределах, которые охраняются исключительным правом. Также в отдельных ситуациях использование объектов интеллектуальной собственности в машинном обучении может подпадать под ограничения ИС, случаи свободного использования (такие как краткосрочная запись произведения, носящая временный и случайный характер[17]; воспроизведение в личных целях; извлечение несущественной части правомерно обнародованной базы данных; и др.)[18]. Тем не менее, такие исключения носят фрагментарный характер и не позволяют сделать вывод о том, что майнинг данных полностью охватывается ограничениями исключительных прав и свободным использованием.

Для создания единообразного подхода к легализации майнинга РИД и СИ целесообразно предусмотреть специальный режим использования объектов интеллектуальной собственности в целях машинного обучения. При этом использование РИД и СИ в машинном обучении не должно рассматриваться как новый способ использования объектов ИС (поскольку технологические этапы машинного обучения охватываются предусмотренными способами использования), а должно характеризоваться как отдельная цель использования (наряду с цитированием, созданием пародии, использованием в личных целях и т.п.).

16. Подробное сопоставление майнинга данных с различными видами использования объектов интеллектуальной собственности проведено в работах как российских, так и зарубежных авторов. См. Кольздорф М.А. Свободное использование объектов авторских и смежных прав при обработке больших данных (Big Data) // Закон. 2021. № 5. С. 142–164; Geiger, C., Frosio, G. & Bulayenko, O. Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?. IIC 49, 814–844 (2018).

17. Как отмечается в литературе, в зависимости от обстоятельств на различных этапах майнинга текста и данных могут создаваться как временные, так и постоянные копии. См. Кольздорф М.А. Свободное использование объектов авторских и смежных прав при обработке больших данных (Big Data) // Закон. 2021. № 5. С. 142–164.

18. Подробнее см. Geiger, C., Frosio, G. & Bulayenko, O. Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?. IIC 49, 814–844 (2018). https://doi.org/10.1007/s40319-018-0722-2 (дата обращения: 22.03.2022).

6. Основания использования объектов интеллектуальной собственности в целях машинного обучения в зарубежных странах

В настоящее время в российском законодательстве об интеллектуальной собственности отсутствуют положения, регулирующие использование РИД и СИ в целях машинного обучения. Однако в некоторых зарубежных юрисдикциях уже существуют специальные нормы о майнинге объектов интеллектуальной собственности.

Наиболее обсуждаемым документом в контексте данной проблематики является Директива ЕС 2019 года об авторском праве и смежных правах на едином цифровом рынке[19], закрепившая условиях т.н. «майнинга текста и данных» (text and data mining – TDM).

Под майнингом текста и данных в Директиве понимается любой автоматизированный аналитический метод, предназначенный для анализа текста и данных в цифровой форме с целью получения информации, включающей, помимо прочего, паттерны, тенденции и корреляции. Технические аспекты применения Директивы в контексте майнинга данных раскрыты в рекомендациях ЕС[20].

В зависимости от целей майнинга текста и данных Директива закрепляет две модели регулирования.

Статья 3 предусматривает исключения для исследовательских организаций и институтов культурного наследия (research organisations and cultural heritage institutions) в отношении воспроизведения и извлечения материалов, к которым есть правомерный доступ, при майнинге текста и данных в научных исследованиях. Директива определяет исследовательские организации и институты культурного наследия достаточно узко. Например, коммерческие исследовательские организации исключены из-под действия данной статьи. В п. 14 преамбулы правомерный доступ к материалам определен как доступ в соответствии с договором (лицензией), а также доступ к общедоступному контенту в сети Интернет (content that is freely available online).

Статья 4 предусматривает общие исключения для воспроизведения и извлечения в целях майнинга материалов, к которым есть правомерный доступ (lawfully accessible works). Однако данные исключения носят диспозитивный характер и могут быть ограничены правообладателем. Такое ограничение (запрет) должно быть явно выражено в соответствующем формате (например, путем размещения прямого запрета на майнинг в машиночитаемом формате, указанием в соглашении с пользователем). Подход, закрепленный в статье 4, применяется к использованию данных в любых целях, включая коммерческие.

Таким образом, для исследовательских целей Директива ЕС предусматривает прямое исключение при использовании объектов авторских и смежных прав в машинном обучении, а во всех остальных случаях закрепляется модель opt-out, разрешающая использование материалов без согласия правообладателя, если последний не выразил явный запрет на такое использование.

Как отмечают эксперты[21], с принятием Директивы 2019 года в ЕС создаются условия для формирования производного рынка датасетов для майнинга данных. Правообладатели получили возможность контролировать, лицензировать или вовсе запрещать майнинг своих объектов интеллектуальной собственности.

19. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC // Режим доступа: https://eur-lex.europa.eu/eli/dir/2019/790/oj (дата обращения: 29.03.2022).
20. The Exception for Text and Data Mining (TDM)in the Proposed Directive on Copyright in the Digital Single Market – Technical Aspects. – URL: https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf (дата обращения: 22.03.2022).
21. P. Bernt Hugenholtz. The New Copyright Directive: Text and Data Mining (Articles 3 and 4) / Kluwer Copyright Blog. July 24, 2019. – URL: http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/ (дата обращения: 22.03.2022).

Из текста Директивы не вполне ясно, могут ли правообладатели взимать плату за использование объектов интеллектуальной собственности при майнинге данных в коммерческих целях. Пункт 17 преамбулы к Директиве напрямую исключает взимание платы только для случаев использования данных в научных (исследовательских) целях. Некоторые эксперты считают, что вопрос о взимании вознаграждения во всех остальных случаях остается на усмотрении стран – членов ЕС[22]. В таких случаях допустимо установление платы за использование объектов интеллектуальной собственности в машинном обучении в коммерческих целях[23].

Положения Директивы ЕС о майнинге текста и данных уже имплементированы в законодательстве многих стран ЕС (Франции, Германии, Эстонии и др.)[24]. Так, согласно ст. L. 122-5-3 Кодекса интеллектуальной собственности Франции (Code de la Propriété Intellectuelle, CPI)[25] под интеллектуальным анализом текста и данных подразумевается использование автоматизированного метода анализа текстов и данных в цифровой форме для извлечения информации, в частности закономерностей, тенденций и корреляций. Цифровые копии или репродукции произведений, к которым есть правомерный доступ, могут быть созданы без разрешения авторов (правообладателей) для извлечения текстов и данных, осуществляемого исключительно с целью проведения научных исследований научно-исследовательскими организациями, общедоступными библиотеками, музеями, архивами или учреждениями культурного наследия либо от их имени и по их просьбе другими лицами, в том числе в рамках некоммерческого партнерства с частными субъектами. Такие копии хранятся с надлежащим уровнем безопасности и могут использоваться исключительно для целей научных исследований, в том числе для проверки результатов исследований. Правообладатели могут принимать соразмерные и необходимые меры для обеспечения безопасности и целостности сетей и баз данных, в которых размещены произведения. Помимо этого, копии или цифровые репродукции произведений, к которым был получен доступ на законных основаниях, могут быть сделаны любым лицом для анализа текста и данных в любых целях, если автор не выразил надлежащим образом запрет, в частности, с помощью машиночитаемых процессов в отношении общедоступного контента, размещенного в Интернете. Копии и репродукции в таком случае хранятся с соответствующим уровнем безопасности, а затем уничтожаются после осуществления анализа текста и данных. Таким образом, во Франции в полноценном виде имплементированы положения Директивы ЕС как в отношении некоммерческого использования в исследовательских целях, так и в отношении иных целей.

Специальные нормы об использовании объектов интеллектуальной собственности в машинном анализе данных содержатся не только в законодательстве стран ЕС, но и в некоторых иных юрисдикциях.

Япония стала первой страной, в национальном праве которой появились положения, касающиеся анализа текста и данных (в 2009 г.). В 2018 г. были внесены новые поправки в законодательство об интеллектуальной собственности. Согласно ст. 30-4 Закона об авторском праве Японии[26] разрешается использовать произведение любым способом и в той мере, в какой это считается необходимым, в любых случаях, когда целью не является «личное наслаждение или принесение наслаждения другому лицу» мыслями или чувствами, выраженными в работе, если это не ущемляет необоснованным образом интересы правообладателя… если это делается для использования в анализе данных (имеется в виду извлечение, сравнение,

22. См. Stieper M. Das Verhältnis der verpflichtenden Schranken der DSM-RL zu den optionalen Schranken der InfoSoc-RL, GRUR 2020, 1 (цит. по Кольздорф М.А. Указ. соч.).
23. Возможность взимания платы за майнинг текста и данных отмечается также другими авторами. См. Séverine Dusollier. The 2019 Directive on Copyright in the Digital Single Market: Some progress, a few bad choices, and an overall failed ambition. Common Market Law Review, Kluwer Law International, 2020, 57 (4), P. 987. Режим доступа: https://hal-sciencespo.archives-ouvertes.fr/hal-03230170/document (дата обращения: 29.03.2022).
24. Papadopoulou, M. , Kolotourou, K. and Bottis, M. (2021) The Exception of Text and Data Mining from the Academic Libraries Standpoint. Open Journal of Social Sciences, 9, 502-539 // Режим доступа: https://www.scirp.org/journal/paperinformation.aspx?paperid=109293 (дата обращения: 29.03.2022).
25. Code de la propriété intellectuelle // Режим доступа: https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006069414/LEGISCTA000006161637/?anchor=LEGIARTI000044365551#LEGIARTI000044365551 (дата обращения: 29.03.2022).
26. Copyright Law of Japan // Режим доступа: https://www.cric.or.jp/english/clj/cl2.html (дата обращения: 29.03.2022).

классификация или другой статистический анализ текста, звуков, изображений или других элементарных данных из большого количества произведений или большого объема других таких данных). Анализ текста и данных допускается также в коммерческих целях, а ограничения, устанавливаемые правообладателем, не имеют значения. Также в законе отсутствуют положения, касающиеся невозможности передавать созданные копии произведений. По мнению отдельных исследователей, исключения в области TDM в законодательстве Японии являются, вероятно, самыми широкими в мире[27].

В **Великобритании** еще в 2014 году были внесены поправки в Закон об авторском праве, дизайнах и патентах применительно к созданию копий для анализа текста и данных для некоммерческих исследований[28]. Согласно статье 29A создание копии произведения лицом, имеющим законный доступ к произведению, не нарушает авторских прав на произведение при соблюдении нескольких условий: 1) машинный (вычислительный) анализ содержания произведения осуществляется исключительно с целью некоммерческого исследования; 2) копия сопровождается указанием на правообладателя или автора (за исключением случаев, когда это невозможно по соображениям практичности или по иным объективным причинам). При этом является нарушением авторского права передача такой копии любому другому лицу и использование копии для любых целей, кроме некоммерческого исследования, за исключением случаев, когда передача и использование копии разрешены правообладателем. В настоящее время в Великобритании ведутся дискуссии о возможном изменении правового режима майнинга объектов интеллектуальной собственности с учетом широкого распространения цифровых технологий обработки данных[29].

В законодательстве **США** не содержится специальных норм об анализе текста и данных. Обоснование законности TDM приведено в судебных решениях и основано на доктрине fair use (свободного «добросовестного» использования). Два решения апелляционного суда США — Authors Guild v. Google[30] и Authors Guild v. HathiTrust[31] — определили, что копирование текстов, защищенных авторскими правами, в рамках TDM для исследовательских целей является добросовестным использованием, а не нарушением.

В **Сингапуре** в ноябре 2021 г. вступил в силу новый Закон об авторском праве[32]. Среди прочего, в сингапурском законодательстве появился раздел, посвященный вычислительному (компьютерному) анализу данных. Вычислительный анализ данных в отношении произведения или записи охраняемого исполнения включает использование компьютерной программы для идентификации, извлечения и анализа данных из произведения или записи; и использование произведения или записи в качестве примера для улучшения функционирования компьютерной программы по отношению к этому типу информации или данных. Примером такого анализа данных является использование изображений для обучения компьютерной программы распознаванию изображений. Лицо (X) может использовать копию произведения или записи исполнения при соблюдении следующих условий: 1) копия сделана с целью анализа данных или подготовки произведения или записи для анализа данных; 2) лицо не использует копию для каких-либо других целей; 3) лицо не предоставляет копию другим лицам, кроме как для целей проверки результатов анализа, совместного исследования или исследования, относящегося к цели анализа, проведенного лицом (X); 4) лицо имеет законный доступ к материалу, с которого сделана копия.

27. Tatsuhiro UENO. TDM Exception in Japan. Possible implication for Europe? // Режим доступа: https://ipwi.uj.edu.pl/documents/122195199/144296432/Ueno_2019_Cracow_Comment.pdf/adf57b6a-ad86-4bd6-94b0-4bbac61b049c (дата обращения: 29.03.2022).
28. Copyright, Designs and Patents Act 1988 // Режим доступа: https://www.legislation.gov.uk/ukpga/1988/48/section/29A (дата обращения: 29.03.2022).
29. AI developers to learn UK position on text and data mining rights // Режим доступа: https://www.pinsentmasons.com/out-law/analysis/ai-developers-to-learn-uk-position-on-text-and-data-mining-rights (дата обращения: (29.03.2022).
30. Authors Guild v. Google, Inc. – 804 F.3d 202 (2d Cir. 2015) // Режим доступа: https://www.lexisnexis.com/community/casebrief/p/casebrief-authors-guild-v-google-inc (дата обращения: 29.03.2022).
31. Authors Guild, Inc. v. HathiTrust – 755 F.3d 87 (2d Cir. 2014) // Режим доступа: https://www.lexisnexis.com/community/casebrief/p/casebrief-authors-guild-inc-v-hathitrust (дата обращения: 29.03.2022).
32. Copyright Act 2021 // Режим доступа: https://sso.agc.gov.sg/Acts-Supp/22-2021/Published/ (дата обращения: 29.03.2022).

Для наглядности в законе Сингапура раскрываются примеры законного и незаконного доступа к исходным материалам. Примерами незаконного доступа считаются получение доступа к материалам в обход платного доступа или получение доступа в нарушение условий использования базы данных. Законным доступ будет считаться, если первая копия (исходный материал) не является копией, нарушающей авторские права. В случае если исходный материал все-таки нарушает авторские права, лицо не должно знать об этом и не должно разумно этого предполагать.

7. Примеры сервисов, предоставляющих доступ к объектам ИС для целей интеллектуального анализа (машинного обучения)

Доступ к датасетам с научными публикациями

Издательство Springer Nature[33] предоставляет сервис, облегчающий доступ исследователей к текстам научных публикаций для интеллектуального анализа данных (TDM)[34]. Немалая часть статей издательства Springer Nature публикуется в открытом доступе. Для этих публикаций TDM обычно разрешается без ограничений, поскольку они, как правило, распространяются по свободной лицензии CC-BY. Доступ к остальным (закрытым) материалам предоставляется по подписке, оформляемой академическими учреждениями. В отношении журналов и книг, на которые распространяется подписка, Springer Nature предоставляет исследователям права на интеллектуальный анализ текстов и данных при условии, что целью является некоммерческое исследование. Исследователи обязаны применять разумные меры для защиты загружаемого контента, хранить контент на защищенном внутреннем сервере без доступа третьих лиц и только на время реализации исследовательского проекта. Для коммерческих TDM исследований предусмотрена возможность доступа к закрытым материалам за плату.

Аналогичный сервис реализован другим крупным издательством Elsevier[35]. Лицензионные условия издательства позволяют исследователям в учреждениях-подписчиках получать доступ к полному текстовому содержимому в машиночитаемом формате и осуществлять интеллектуальный анализ текстов[36]. Доступ к контенту для интеллектуального анализа текста предоставляется подписчикам в некоммерческих исследовательских целях. Иные запросы для осуществления TDM (в том числе в коммерческих целях) рассматриваются в индивидуальном порядке.

Американская организация по управлению правами на коллективной основе Copyright Clearance Center (CCC) предоставляет пользователям сервис RightFind XML for Mining для осуществления интеллектуального анализа текста и данных[37]. Обычно доступ к полным текстам статей возможен только по подписке на ресурсы издательств, а документы в них, как правило, имеют не машиночитаемый формат. RightFind XML for Mining предоставляет доступ к большому количеству материалов различных издательств в формате, пригодном для интеллектуального анализа текста и данных (XML).

33. Springer Nature — академическая издательская компания, выпускающая передовые научные журналы (более 3000) и книги (более 300000 книг, более 200 тематических серий).
34. Text and Data Mining at Springer Nature // Режим доступа: https://www.springernature.com/gp/researchers/text-and-data-mining (дата обращения: 22.04.2022).
35. Elsevier — один из крупнейших научных издательских домов мира наряду с Springer (выпускает более 2000 научных журналов).
36. Text and data mining. Elsevier // Режим доступа: https://www.elsevier.com/about/policies/text-and-data-mining (дата обращения: 22.04.2022).
37. RightFind XML for Mining // Режим доступа: https://www.copyright.com/businesses/xmlformining/ (дата обращения: 22.04.2022).

Доступ к датасетам с музыкальными произведениями

Датасеты с музыкальными произведениями могут использоваться в машинном обучении в целях создания новой музыки[38], разработки автоматизированных систем классификации музыкальных произведений (например, по жанру[39]), систем пользовательских рекомендаций[40] и т.п.

Например, сервис Jukebox, предназначенный для создания музыки с помощью искусственного интеллекта, использует датасет с более чем 1,2 млн музыкальных произведений и сопутствующих данных (тексты песен и метаданные записей)[41]. Аналогичный сервис MuseNet[42] собирает обучающие данные из разных источников (как открытых, так и проприетарных). Некоторые организации (ClassicalArchives, BitMidi) для реализации проекта передали датасеты на безвозмездной основе.

Как правило, разработчики для обучения систем искусственного интеллекта используют открытые музыкальные датасеты, распространяемые на условиях свободных лицензий[43]. Так, датасет MAESTRO[44] с фортепьянными записями доступен любому разработчику на условиях лицензии Creative Commons Attribution Non-Commercial Share-Alike 4.0 (CC BY-NC-SA 4.0). Некоторые сервисы (например, крупнейший стриминговый сервис Spotify[45]) по запросу предоставляют свои датасеты исключительно для исследовательских и некоммерческих целей.

Доступ к датасетам с изображениями

Изображения в машинном обучении используются для генерации новых изображений, разработки систем компьютерного зрения (распознавания объектов), сжатия файлов, в редакторах изображений и т.п. Например, сервис DALL·E 2 на основе машинного обучения позволяет создавать и преобразовывать изображения по текстовому описанию[46]. Разработка подобных сервисов невозможна без доступа к обучающим датасетам, состоящим из большого количества изображений и их описаний.

Как и в случае с музыкальными произведениями, разработчикам доступны открытые датасеты с изображениями на условиях свободной лицензии[47]. Например, компания Google предоставляет доступ к миллионам изображений в рамках проекта Open Images Dataset[48]. Изображения «подкреплены» описанием объектов из более 6000 категорий.

Разработчикам также доступны платные датасеты с изображениями[49]. Такие датасеты, как правило, проходят предварительную обработку, разметку объектов. Некоторые организации предлагают услуги подготовки датасетов заказчика к машинному обучению (аннотирование, сегментирование изображений и т.п.)[50].

---

38. Magenta. Make Music and Art Using Machine Learning // Режим доступа: https://magenta.tensorflow.org/ (дата обращения: 22.04.2022).
39. Lansdown, Bryn. (2019). Machine Learning for Music Genre Classification // Режим доступа: https://www.researchgate.net/publication/337001430_Machine_Learning_for_Music_Genre_Classification (дата обращения: 22.04.2022).
40. How Do AI Music Recommendation Systems Work // Режим доступа: https://cyanite.ai/2021/09/02/how-do-ai-music-recommendation-systems-work/ (дата обращения: 22.04.2022).
41. Jukebox // Режим доступа: https://openai.com/blog/jukebox/ (дата обращения: 22.04.2022).
42. MuseNet // https://openai.com/blog/musenet/ (дата обращения: 22.04.2022).
43. Подборка открытых датасетов с музыкой доступна здесь: https://paperswithcode.com/datasets?mod=music (дата обращения: 22.04.2022).
44. The MAESTRO Dataset // Режим доступа: https://magenta.tensorflow.org/datasets/maestro (дата обращения: 22.04.2022).
45. Datasets. Dive into datasets for everything from podcasts to music recommendation // Режим доступа: https://research.atspotify.com/datasets/ (дата обращения: 22.04.2022).
46. DALL·E 2 // Режим доступа: https://openai.com/dall-e-2/ (дата обращения: 22.04.22).
47. См. напр. 50 free Machine Learning Datasets: Image Datasets // Режим доступа: https://blog.cambridgespark.com/50-free-machine-learning-datasets-image-datasets-241852b03b49 (дата обращения: 22.04.2022).
48. Open Images Dataset // Режим доступа: https://storage.googleapis.com/openimages/web/index.html (дата обращения: 22.04.2022).
49. См. напр. Make ML Dataset Store // Режим доступа: https://makeml.app/dataset-store (дата обращения: 22.04.2022).
50. См. iMERIT. ML DATA SOLUTIONS // Режим обращения: https://imerit.net/ (дата обращения: 22.04.2022).

Рынок датасетов в Российской Федерации

В Российской Федерации рынок датасетов находится на раннем этапе формирования, поэтому нет устоявшихся правил по их предоставлению и использованию в целях машинного обучения. Проекты в области искусственного интеллекта зачастую используют зарубежные датасеты из открытых источников. Русскоязычные датасеты для интеллектуального анализа текста и их подборки зачастую являются результатом труда отдельных энтузиастов (например, коллекция Corus[51]).

Некоторые крупные IT-компании предоставляют доступ к созданным и размеченным ими датасетам. К примеру, Сбербанк открыл доступ к датасету Golos — самому большому набору речевых данных на русском языке, размеченному вручную[52]. Сервис Яндекс.Толока предоставляет возможность заказчикам использовать краудсорсинг для разметки данных. С помощью Яндекс.Толоки можно собирать и размечать данные для задач компьютерного зрения, обработки естественного языка, речевых технологий, алгоритмов информационного поиска и других задач машинного обучения. Яндекс предоставляет датасеты для проведения академических исследований безвозмездно только для некоммерческого использования со ссылкой на Толоку как источник датасета[53]. Если датасеты планируется использовать в коммерческих целях, требуется отдельное согласие.

8. Возможные модели использования и коммерциализации объектов интеллектуальной собственности в целях машинного обучения в России

С учетом опыта зарубежных стран и национальных особенностей правового регулирования в сфере интеллектуальной собственности возможно рассмотреть три условные модели использования и коммерциализации интеллектуальной собственности в целях машинного обучения в Российской Федерации.

Первая модель, основанная на принципах свободного использования, предполагает использование объектов интеллектуальной собственности в машинном обучении без согласия правообладателя и без выплаты вознаграждения, если существенным образом не ущемляются интересы правообладателя. Такая модель не представляется оптимальной для правовой системы Российской Федерации, поскольку российское право (в отличие от права США и других стран, где используется доктрина fair use) не обеспечивает достаточную гибкость в применении критериев свободного использования к каждой конкретной ситуации, в связи с чем существуют высокие риски нарушения законных интересов правообладателей.

Другая, противоположная модель предполагает запрет на использование объектов интеллектуальной собственности в машинном обучении без согласия правообладателя и без выплаты вознаграждения вне зависимости от цели использования. Такая модель является жесткой фиксацией исключительных правомочий правообладателя по предоставлению согласия на интеллектуальный анализ охраняемых объектов интеллектуальной собственности. Представляется, что такой подход, с одной стороны, будет препятствовать доступу к данным для разработки и обучения систем искусственного интеллекта (особенно для исследовательских целей), а с другой – сохранит коллизии при квалификации действий пользователей

51. Corus — коллекция русскоязычных NLP-датасетов // Режим доступа: https://natasha.github.io/corus/ (режим доступа: 02.05.2022).
52. Сбер открывает доступ к датасету Golos // Режим доступа: https://press.sber.ru/publications/sber-otkryvaet-dostup-k-datasetu-golos-samomu-bolshomu-naboru-rechevykh-dannykh-na-russkom-razmechennomu-vruchnuiu (дата обращения: 02.05.2022),
53. Яндекс. Толока. Открытые датасеты // Режим доступа: https://toloka.ai/ru/datasets (дата обращения: 03.05.2022).

между использованием объектов интеллектуальной собственности в целях машинного обучения и отдельными случаями свободного использования.

Третья, компромиссная модель, по аналогии с законодательством ЕС предполагает **дифференциацию подходов к использованию объектов интеллектуальной собственности в машинном обучении в зависимости от целей использования**. Для некоммерческого использования в исследовательских целях предусматривается режим свободного использования правомерно обнародованных объектов[54]. Для всех других целей может быть предусмотрена либо модель opt-in (допускается только с согласия правообладателя), либо модель opt-out (допускается, если правообладатель не запретил использование, как в ЕС). Модель opt-out представляется более предпочтительной, поскольку, во-первых, позволяет вовлечь в сферу свободного доступа большой объем объектов интеллектуальной собственности, в отношении которого у правообладателя нет возражений по использованию в машинном обучении, во-вторых, не лишает правообладателя возможности коммерциализации своих объектов интеллектуальной собственности при их использовании в машинном обучении (таким образом обеспечивается гибкий баланс частных и публичных интересов). Реализация подобного подхода будет способствовать развитию исследований с применением технологий искусственного интеллекта, а также создаст предпринимательские стимулы и условия для формирования рынка датасетов.

Посредническая деятельность по сбору, подготовке и последующей передаче третьим лицам датасетов, предназначенных для использования в машинном обучении, в рамках предложенной модели возможна, по общему правилу, с согласия правообладателей первичных материалов. Однако в тех случаях, когда правообладатели предоставили правомерный доступ к материалам неограниченному кругу лиц и не установили запрет на использование объектов ИС в машинном обучении, такое посредничество возможно без согласия правообладателей (с выплатой или без выплаты вознаграждения – в зависимости от условий правообладателя). В тех случаях, когда создание датасета требует от посредника творческого вклада и (или) существенных финансовых, материальных, организационных или иных затрат, на подготовленный датасет будет распространяться авторское и (или) смежное право. Использование такой базы данных также должно осуществляться с учетом правил, установленных для использования объектов ИС в машинном обучении.

Для обеспечения взаимодействия правообладателей и пользователей (разработчиков, дата-сайнтистов) возможно создание платформ (маркетплейсов), где на коммерческой основе будут доступны подборки датасетов, представленные в машиночитаемом формате и структурированные для использования в машинном обучении. Побочным позитивным эффектом от развития таких платформ будет более высокое качество и количество доступных датасетов, в том числе пригодных для свободного использования в некоммерческих исследовательских целях.

9. Выводы

1) Интеллектуальная собственность в машинном обучении

Правовые аспекты использования объектов интеллектуальной собственности в машинном обучении охватывают круг вопросов, связанных с признанием исключительных прав на датасеты (базы данных) и их составные элементы, а также вопросы о законных основаниях такого использования. Исключительные права могут признаваться на следующие материалы, используемые в машинном обучении (отдельно или в совокупности):

54. Для недопущения расширительного толкования исследовательской цели возможно ограничить круг субъектов, которые имеют право свободно использовать объекты авторского права в машинном обучении. Таким образом, условиями свободного использования по аналогии с нормами законодательства ЕС будут: 1) исследовательская цель; 2) субъекты использования – учреждения науки и культуры.

- составные элементы базы данных (датасета), являющиеся самостоятельными объектам ИС (например, музыкальные произведения, статьи, фотографии и др.);
- базу данных (датасет), охраняемую в режиме авторского и (или) смежного права, если имеет место творческий вклад и (или) существенные инвестиции в создание базы данных (например, крупная база научных публикаций, собранных из большого количества издательств);
- базу данных (датасет), созданную в результате посреднической деятельности при подготовке материалов непосредственно к машинному обучению, если также есть творческий вклад и (или) существенные инвестиции в создание базы данных (например, подборка материалов из открытых источников в сети Интернет, полученная в результате веб-скрепинга).

2) Правовая охрана базы данных (датасета) как объекта интеллектуальной собственности

На базу данных может распространяться авторское право автора составного произведения, а также смежное право изготовителя базы данных. В базе данных как объекте авторского права охраняются расположение и/или выборка материалов, осуществленные создателем базы данных, если они представляют собой результат творческого труда (например, расположение материалов по жанру, стилю, сфере научных интересов и иным критериям, предполагающим творческий подход к оценке и структурированию материалов).

В режиме смежных прав охраняется не творческий вклад, а инвестиции в создание базы данных. В соответствии с российским законодательством смежное право на базу данных предоставляется в том случае, если изготовитель базы данных несет на ее создание (включая обработку и представление материалов) существенные финансовые, материальные, организационные или иные затраты (например, создание справочной правовой системы из миллионов правовых актов). На уровне судебной практики в России не сформировался единообразный подход к квалификации существенности затрат на изготовление базы данных.

3) Машинное обучение как использование объектов интеллектуальной собственности

При использовании объектов интеллектуальной собственности в машинном обучении возможны воспроизведение, переработка, извлечение материалов из базы данных, доведение до всеобщего сведения и осуществление иных способов использования, охраняемых исключительным правом.

В некоторых случаях применение машинного обучения предполагает лишь «сканирование» материалов в открытых источниках с целью поиска конкретной информации или закономерностей в массиве данных. Такой метод машинного обучения может обойтись без использования объектов ИС в тех пределах, которые охраняются исключительным правом. Также в отдельных ситуациях использование объектов интеллектуальной собственности в машинном обучении может подпадать под ограничения ИС, случаи свободного использования (такие как краткосрочная запись произведения, носящая временный и случайный характер; воспроизведение в личных целях; извлечение несущественной части правомерно обнародованной базы данных; и др.).

Предусмотренные в законодательстве основания для свободного использования объектов ИС лишь фрагментарно могут легализовывать отдельные случаи майнинга данных. В целях создания благоприятных правовых условий для использования объектов ИС в машинном обучении необходимо законодательно закрепить основания такого использования. При этом использование РИД в машинном обучении не должно рассматриваться как новый способ использования объектов ИС (поскольку технологические этапы машинного обучения охватываются предусмотренными способами использования), а должно характеризоваться как отдельная цель использования (наряду с цитированием, созданием пародии, использованием в личных целях и т.п.)

4) Исключения для майнинга текста и данных в зарубежных правопорядках

В настоящее время в российском законодательстве об интеллектуальной собственности отсутствуют положения, регулирующие использование РИД в целях машинного обучения. Однако в некоторых зарубежных юрисдикциях уже существуют специальные нормы о майнинге объектов интеллектуальной собственности. Наиболее обсуждаемым документом в контексте данной проблематики является Директива ЕС 2019 года об авторском праве и смежных правах на едином цифровом рынке, закрепившая условиях т.н. «майнинга текста и данных» (text and data mining – TDM). Для исследовательских целей Директива ЕС предусматривает прямое исключение при использовании объектов авторских и смежных прав в машинном обучении (то есть допускается свободное использование учреждениями науки и культуры), а во всех остальных случаях (в том числе при коммерческом использовании) закрепляется модель opt-out, разрешающая использование материалов без согласия правообладателя, если последний не выразил явный запрет на такое использование.

5) Направление развития регулирования майнинга объектов ИС в России

Оптимальной моделью использования и коммерциализации объектов ИС в машинном обучении для Российской Федерации представляется подход, близкий к законодательству ЕС. В рамках данной модели может быть предусмотрена дифференциация условий использования объектов ИС в зависимости от цели и субъектов. Для использования в исследовательских целях учреждениями науки и культуры – свободное использование правомерно обнародованных РИД. Для всех иных случаев – по умолчанию без согласия правообладателя и без выплаты вознаграждения, если иное не установил правообладатель. Реализация подобного подхода будет способствовать развитию исследований с применением технологий искусственного интеллекта, а также создаст предпринимательские стимулы и условия для формирования рынка датасетов.

6) Рекомендации по изменению российского законодательства

- Дополнить ст. 1274 ГК РФ правом научных организаций и учреждений культуры без согласия правообладателя и без выплаты вознаграждения использовать правомерно обнародованные произведения для автоматизированной обработки с помощью ЭВМ при проведении исследований без цели извлечения прибыли (данное положение также будет распространяться на объекты смежных прав в соответствии со статьей 1306 ГК РФ);
- Дополнить главу 70 ГК РФ новой статьей «Использование правомерно обнародованных произведений при автоматизированной обработке с помощью ЭВМ» следующего содержания:

«Использование правомерно обнародованных произведений при автоматизированной обработке с помощью ЭВМ, если такое использование не противоречит обычному использованию произведений и не ущемляет необоснованным образом законные интересы правообладателей, допускается без согласия правообладателя и без выплаты вознаграждения, кроме случаев, когда правообладатель установил запрет или ограничения на такое использование (в том числе посредством технических средств защиты авторских прав)». Аналогичную статью следует предусмотреть в отношении объектов смежных прав в главе 71 ГК РФ «Права, смежные с авторскими»;

- В целях защиты интересов авторов (правообладателей) оригинальных произведений, используемых в машинном обучении, в новую статью об использовании произведений при автоматизированной обработке с помощью ЭВМ целесообразно добавить следующее положение:
- «Создание производных произведений в результате автоматизированной обработки оригинальных произведений, не связанное исключительно с технической переработ-

кой оригинального произведения в машиночитаемый формат, допускается с соблюдением прав авторов оригинальных произведений»;

- Для внесения определенности в вопросы правовой охраны баз данных как объектов смежных прав следует внести в статью 1334 ГК РФ уточнение о характере финансовых, материальных, организационных или иных затрат, понесенных на изготовление баз данных. Рекомендуется учитывать существенные затраты не только непосредственно на создание базы данных, но и на деятельность, способствующую созданию базы данных. Такой широкий подход к распространению правовой охраны на базы данных будет компенсирован исключениями, предусматривающими более свободное использование объектов интеллектуальной собственности в машинном обучении;
- Для внесения определенности в вопрос квалификации правомерно обнародованного произведения и базы данных, используемых в целях машинного обучения, целесообразно добавить уточнение в соответствующие статьи ГК РФ (ст. 1274 и другие), что под правомерно обнародованным произведением (базой данных) понимается произведение (база данных), доступ к которому возможен без нарушения закона (включая доступ неограниченного круга лиц в сети Интернет).