УДК 34.01
DOI: https://doi.org/10.17308/law/1995-5502/2025/3/64-71

ПРАВОВОЕ РЕГУЛИРОВАНИЕ ИНДУСТРИИ АННОТАЦИИ ДАННЫХ КАК СПОСОБ ОБЕСПЕЧЕНИЯ КАЧЕСТВА ДАННЫХ[1]

Э. И. Лескина

Национальный исследовательский университет «Высшая школа экономики» (Москва)

LEGAL SUPPORT FOR THE DATA ANNOTATION INDUSTRY AS A WAY TO ENSURE DATA QUALITY

E. I. Leskina

Hational Recearch University «Higher School of Economics» (Moscow)

Аннотация: национальный проект «Экономика данных», направленный на цифровую трансформацию различных сфер, является следующей ступенью для обеспечения научно-технологического суверенитета в Российской Федерации, при этом ключевым моментом для реализации многочисленных направлений, федеральных проектов, мероприятий в рамках национального проекта является повышение качества как данных, так и их наборов, обеспечение роста доступных для использования различными субъектами данных. Вопросы качества данных и особенно управления таким качеством остаются за рамками нормативного обеспечения и регулирования. Между тем одним из способов обеспечения качества данных является индустрия аннотации данных. Многие страны обращают внимание на потенциал, заложенный в развитии данного направления, начинают принимать стратегические документы. В статье обосновывается необходимость нормативного обеспечения развития индустрии аннотации данных в Российской Федерации, рассматриваются зарубежные подходы, которые в настоящее время начинают формироваться в данной сфере, предлагаются направления, которые следует развивать для системного становления индустрии аннотации в России как способа обеспечения качества данных. Обращается внимание на многогранность потенциала аннотации данных, в том числе для обеспечения качества алгоритмов, развития узко специализированных моделей искусственного интеллекта.
Ключевые слова: большие данные, искусственный интеллект, экономика данных, аннотация данных, аннотирование, маркировка, разметка, качество данных.

Abstract: the national project «Data Economy» aimed at digital transformation of various spheres is the next step to ensure scientific and technological sovereignty in the Russian Federation, while the key point for the implementation of numerous areas, federal projects, and activities within the national project is to improve the quality of both data and their sets, ensuring the growth of data available for use by various subjects. Issues of data quality and especially the management of such quality remain beyond the scope of regulatory support and regulation. Meanwhile, one of the ways to ensure data quality is the data annotation industry. Many countries pay attention to the potential inherent in the development of this area and begin to adopt strategic documents. This article substantiates the need for regulatory support for the development of the data annotation industry in the Russian Federation, considers foreign approaches that are currently beginning to form in this area, and suggests directions that should be developed for the systemic formation of the annotation industry in Russia as a way to ensure data quality. Attention is drawn to the versatility of the data annotation potential, including for ensuring the quality of algorithms, the development of highly specialized models of artificial intelligence.
Key words: big data, artificial intelligence, data economics, data annotation, annotation, labeling, marking, data quality.

[1] Исследование выполнено за счет гранта Российского научного фонда № 25-18-00698 (https://rscf.ru/project/25-18-00698/).

Большие данные являются уже не новым феноменом, его введение в научный оборот насчитывает уже более 20 лет, и понимание потенциала ценности массивов данных признается как публичным, так и частным сектором. Большие данные являются ключевым активом для повышения эффективности различных сфер и отраслей экономики. Более того, переход к платформенной экономике приводит к новой реальности, в которой бизнес-модели крупнейших компаний построены на нематериальных активах[2]. Вместе с тем говорить о реализации потенциала, заложенного в данных, их компиляции, аналитике, использовании различными субъектами от стартапов до государств и их объединений невозможно без решения вопроса о качестве данных и управлении им. Вопросы качества данных не только обеспечивают оборот таковых, но и снижают многочисленные риски, присущие эпохе Big Data.

Качество данных является неотъемлемым атрибутом, требованием и залогом для успешного развития искусственного интеллекта. Поэтому обеспечение качественных данных выступает приоритетной задачей как на уровне бизнеса, так и государства. В феврале 2025 г. в России был презентован национальный проект «Экономика данных и цифровая трансформация государства». Ключевой акцент в развитии IT-отрасли в Российской Федерации отводится развитию решений, работающих на базе накопленных данных[3]. Тем не менее вопрос о том, какие признаки составляют понятие качественных данных, в том числе наборов данных, как обеспечивать такое качество, на законодательном уровне не получил разрешения.

Отдельные нормы отечественного законодательства устанавливают определенные требования к качеству, однако не на системном уровне, что способствовало бы формированию единой экосистемы данных, а применительно к данным, содержащимся в государственных информационных системах, государственных информационных ресурсах. Так, задачей эксперимента по повышению качества и связанности данных, содержащихся в государственных информационных ресурсах, является разработка методик по определению подходов к повышению качества и связанности данных[4]. Документ, который бы фактически определял критерии качества данных, в настоящий момент находится на стадии проекта и содержит такие критерии, как:

– актуальность (свойство данных соответствовать действительности в текущий момент времени);

– достоверность (свойство данных отражать реальное состояние объектов учета на определенный момент времени);

– уникальность (отсутствие повторяющихся записей);

– согласованность и др.

Проект включает также определение качества данных как совокупности свойств данных, обусловливающих их пригодность удовлетворять определенные потребности в соответствии с их назначением[5]. Однако несмотря на достаточно давний срок разработки проекта (2019 г.), он до сих пор не подписан, что препятствует полноценной реализации норм, связанных с обеспечением качества данных, содержащихся в государственных информационных ресурсах.

В качестве другого примера требований к информации (а значит, и к данным), содержащейся в государственных информационных системах, можно отметить достоверность и актуальность[6].

Вопросы, связанные с качеством данных, регулируются и на уровне стандартов[7]. Несмотря

[2] См.: *Макафи Э., Бриньолфсон Э.* Машина, платформа, толпа. Наше цифровое будущее. М., 2019, 368 с.

[3] См.: Дмитрий Григоренко : ИТ-отрасль стала одной из самых быстрорастущих в российской экономике. URL: http://government.ru/news/54354/ (дата обращения: 28.02.2025).

[4] См.: О проведении эксперимента по повышению качества и связанности данных, содержащихся в государственных информационных ресурсах (вместе с «Положением о проведении эксперимента по повышению качества и связанности данных, содержащихся в государственных информационных ресурсах») : постановление Правительства РФ от 3 июня 2019 г. № 710 (ред. от 02.02.2024) // Собр. законодательства Рос. Федерации. 2019. № 23. Ст. 2963.

[5] См.: Проект Приказа Минкомсвязи России «Об утверждении порядка и критериев определения качества сведений, содержащихся в государственных информационных ресурсах, их связанности, включая требования к документированию и разрешению инцидентов при передаче сведений». Доступ из справ.-правовой системы «КонсультантПлюс».

[6] См.: Часть 9 ст. 14 Федерального закона от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации» // Собр. законодательства Рос. Федерации. 2006. № 31 (ч. 1). Ст. 3448.

[7] См., например: ГОСТ Р 71484.2-2024 (ИСО/МЭК 5259-2:2024). Национальный стандарт Российской Федерации. Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 2.

на некоторую терминологическую несогласованность данных стандартов с применяемыми в законодательстве терминами в сфере данных, можно отметить, что такие стандарты содержат общие начала не только выявления критериев качества данных (аккуратность, наполненность, согласованность, достоверность, актуальность, доступность, эффективность, понятность, переносимость и др., причем указанные термины понимаются в ином значении, нежели в рассмотренных ранее актах применительно к данным в государственных информационных ресурсах), но также говорят о способах управления таким качеством, действиях по его обеспечению, включая институт аннотации (по терминологии ГОСТов маркировки, аннотированию) данных, который заявлен в качестве необходимого на стадии подготовки наборов данных. Аннотирование данных здесь включается в план управления качеством данных. Кроме того, указывается, что аннотирование данных может осуществляться автоматически или вручную, а также с помощью метаданных о наборе данных.

В литературе качество данных характеризуется с точки зрения их точности, достоверности, полноты, согласованности, своевременности, доступности, уникальности[8], большого объема[9].

Показатели качества данных (утв. и введен в действие приказом Росстандарта от 28 октября 2024 г. № 1551-ст) ; ГОСТ Р 71484.1-2024 (ИСО/МЭК 5259-1:2024). Национальный стандарт Российской Федерации. Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры (утв. и введен в действие приказом Росстандарта от 28 октября 2024 г. № 1537-ст) ; ГОСТ Р 71484.3-2024 (ИСО/МЭК 5259-3:2024). Национальный стандарт Российской Федерации. Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 3. Требования и рекомендации по управлению качеством данных (утв. и введен в действие приказом Росстандарта от 28 октября 2024 г. № 1538-ст) ; ГОСТ Р 71484.4-2024 (ИСО/МЭК 5259-4:2024). Национальный стандарт Российской Федерации. Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 4. Структура процесса управления качеством данных (утв. и введен в действие приказом Росстандарта от 28 октября 2024 г. № 1552-ст) ; и др.

[8] См.: *Чайка М.* Практический подход к валидации рейтинговых моделей при реализации ПВР-подхода : методика 5×5 // Риск-менеджмент в кредитной организации. 2024. № 1. С. 19–34.

[9] См.: Государство, общество и личность : пути преодоления вызовов и угроз в информационной сфере : монография / Н. С. Волкова, А. А. Ефремов, С. М. Зырянов [и др.] ; отв. ред. Л. К. Терещенко. М. : Инфотропик Медиа, 2024. 352 с.

Что касается процедурных вопросов обеспечения данных, то, как правило, речь идет о вопросах регламентации процесса исправления ошибок в данных, контроля за актуальностью, достоверностью, полнотой данных, распределению полномочий субъектов в этой сфере[10].

В отношении требований к качеству самих данных или наборов данных считаем, что эффективными здесь будут подходы, устанавливающие такие требования при обороте, обмене данными с публичным сектором, а также если вопросы оборота могут затрагивать права и законные интересы третьих лиц (применительно, например, к требованию о недискриминационности как критерию качества наборов данных). Что же касается удовлетворения нужд частного оборота, то саморегулирование будет здесь наиболее приемлемым началом. В то же время вопросы системности в регулировании управления качеством данных должны решаться более централизованно, именно это обеспечит стимулирование к росту доступных данных, обеспечению их повторного использования, реализации потенциала, заложенного в больших данных. И одним из направлений управления качества данными является государственная поддержка индустрии аннотации данных.

Индустрия аннотации (аннотирования, маркировки) данных – это сравнительно новая, развивающаяся отрасль, направленная на повышение качества данных. Данная отрасль включает очищение, классификацию, комментирование, иную обработку данных для получения качественных данных и использование их в целях развития инноваций. Сущность института аннотации (маркировки) данных (далее – индустрия маркировки, индустрия аннотации) в том, что данные помечены, на них обозначены признаки, различия, сходства с целевыми данными, и такая обработка производится для целей машинного обучения.

Индустрия аннотации данных является важнейшим звеном для перехода к интеллектуальному производству и необходимым элементом системы по реализации возможностей использования больших данных и развития искусственного интеллекта.

[10] См.: *Назаров Н. А.* Обеспечение качества данных при автоматизированном принятии решений в государственном управлении // Журнал российского права. 2024. № 5. С. 140–155.

В настоящее время государства устанавливают модели управления данными, которые учитывали бы необходимость использования потенциала данных для обеспечения инновационного развития. Содействие развитию отрасли маркировки данных – одно из направлений в создании экосистемы данных. Так, в КНР по плану до 2027 г. предполагается качественное развитие рассматриваемой отрасли с годовым темпом роста более 20 %[11]. В Китае планируется создать единую сеть государственных служб для маркировки данных по всей стране. Значительное внимание уделено обеспечению индустрии маркировки данных специалистами путем принятия профессиональных стандартов для профессий, связанных с маркировкой данных, поощрения сотрудничества образовательных учреждений и предприятий, строительства учебных баз, формирования кадрового резерва в этой области и т. д., а также обеспечению информационной безопасности на всех этапах маркировки данных, выявлению и предупреждению рисков в сфере обработки данных, подлежащих маркировке.

В ЕС гармонизация аннотаций данных является одним из трех направлений обеспечения обмена, повторного использования данных[12]. В целом аннотация данных рассматривается различными документами[13] в ЕС как улучшение данных, их обогащение, повышение стоимости данных. Кроме того, аннотация данных может быть использована в качестве способа обеспечения конфиденциальности частных данных при их предоставлении государственному сектору[14].

Потенциал использования аннотации может быть реализован и при определении предвзятости алгоритмов в качестве стадий для такого исследования, предваряющей тестирование[15]. В этих целях аннотация может использоваться как определенный контекст для алгоритмов, которого не достает для обеспечения качества аналитики при принятии решений на основании алгоритмов.

Интересно, что до февраля 2024 г. Указом Президента РФ[16] устанавливались определенные способы обеспечения качества данных, которые можно отнести и к маркировке. Например, как атрибуты обеспечения большого объема и качества данных предусматривались их разметка и структурирование, разработка и унификация методологии описания. Тем не менее указанные положения утратили силу 15 февраля 2024 г.

Индустрия маркировки данных включает в себя следующие элементы:

– субъекты (в том числе компании, предприятия), осуществляющие маркировку данных на профессиональной основе, исследователи;

– предмет маркировки – данные, подпадающие, как правило, под общий правовой режим. Часто маркировке подлежат технические, а также научные данные. Однако это не умаляет реализации потенциала маркировки для всех видов данных, включая персональные;

– инфраструктура, обеспечивающая связь между субъектами в сфере маркировки данных, хранение данных, их обработку, в том числе сбор (лаборатории, научные центры, инновационные платформы и т. д.).

[11] См.: 国家发展改革委等部门关于促进数据标注 产业高质量发展的实施意见 发改数据〔2024〕1822号. URL: https://www.gov.cn/zhengce/zhengceku/202501/content_6998194.htm (дата обращения: 07.02.2025).

[12] См.: COMMISSION STAFF WORKING DOCUMENT on the free flow of data and emerging issues of the European data economy Accompanying the document Communication Building a European data economy SWD/2017/02 final. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017SC0002&qid=1740591738305 (дата обращения: 26.02.2025).

[13] См.: COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT REPORT Accompanying the document Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act) SWD/2022/34 final. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022SC0034&qid=1740591738305 ; Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space COM/2022/197 final. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197&qid=1740591738305 (дата обращения: 26.02.2025).

[14] См.: Пункт «c» ст. 4.2. Guidance on sharing private sector data in the European data economy Accompanying the document Communication from the Commission to the European Parliament, the Council, the European economic and social Committee and the Committee of the Regions «Towards a common European data space. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018SC0125&qid=1741855165842 (дата обращения: 13.03.2025).

[15] См.: *Харитонова Ю. С., Савина В. С., Паньини Ф.* Предвзятость алгоритмов искусственного интеллекта : вопросы этики и права // Вестник Перм. ун-та. Юридические науки. 2021. № 3. С. 488–515.

[16] См.: Указ Президента РФ от 10 октября 2019 г. № 490 // Собр. законодательства Рос. Федерации. 2019. № 41. Ст. 5700.

Индустрия аннотации данных активно развивается в различных направлениях, и одним из предметов аннотирования данных являются юридические и нормативные документы. Часто язык таких документов труден и изобилует сложными предложениями. Понимание многих юридических текстов бывает затруднительно и для профессионалов. Тем не менее в связи с распространением справочно-правовых систем и функционала сайтов судов аннотирование юридических данных развивается успешно. Благодаря развитию больших языковых моделей появляются новые возможности, значительно увеличивающие скорость и объем аннотированных данных, в результате создающие большие юридические данные, характеризующиеся высоким качеством. Использование таких инструментов в совокупности с технологиями Big Data позволяет решать в том числе прогностические задачи в области юриспруденции, совершенствование моделей искусственного интеллекта, применяемых для обеспечения деятельности юристов.

Аннотирование здесь будет заключаться не в выявлении ключевых слов или содержания в сжатой форме, а в сборе структурированной детализированной информации, охватывающей максимальные подробности (например, влияние свидетелей на решение суда, мотив преступления и другие). Обучение применения больших языковых моделей осуществляется на тех аннотациях, которые составляли юристы и студенты-юристы[17]. Еще одним подходом к аннотированию больших юридических данных является проставление меток искусственным интеллектом с подтверждением или отклонением таких меток специалистами. Подобные эксперименты с генеративным искусственным интеллектом показали высокий результат – 90 % качества ручного аннотирования[18].

Важной задачей выступает обеспечение качества аннотаций, на которых обучаются модели, а качество таких «тренировочных» аннотаций зависит от различных аспектов: уровня образования, опыта, профессионализма юристов, производящих аннотацию, непонимание контекста, ошибки. В результате в тренировочных аннотациях возникают разногласия, и это затрудняет обучение моделей, которое должно основываться на единой согласованной базе. Вместе с тем существует мнение, что пороки в тренировочных аннотациях будут препятствовать качественным прогностическим функциям, но на само качество данных и последующее обучение разногласия в аннотациях влиять не будут[19].

Какие направления следует развивать для становления индустрии аннотации в России?

Во-первых, это более активное использование общедоступных данных. Необходимо обеспечить юридическими средствами правовой режим общедоступных данных, выявив виды таких данных в различных областях и установив требования к формату и маркировке таких данных. В результате может быть создана база маркировки общедоступных данных, используемая в том числе для развития инновационных технологий. Несмотря на то что общим правовом режимом как для данных, так и для информации является режим свободного использования, открытость доступа[20], простой декларации принципов открытости недостаточно для реализации потенциала общедоступных данных.

Во-вторых, стимулирование субъектов в сфере индустрии маркировки данных к единообразным подходам, стандартизации и обмене опытом и разработками в этой сфере. Это могут быть налоговые льготы. Примеры так называемого мягкого права в европейском законодательстве можно привести пока в отношении содействия переносимости данных пользователя[21] или института совместного создания стоимости дан-

[17] См.: *Subinay A, Procheta S., Dwaipayan R., Kripabandhu Gh*. A case study for automated attribute extraction from legal documents using large language models // Artificial Intelligence and Law. 2024. November.

[18] См.: *Gray M., Savelka J., Oliver W., Ashley K.* Can GPT Alleviate the Burden of Annotation? // Legal Knowledge and Information Systems. 2013. URL: https://www.researchgate.net/publication/376422423_Can_GPT_Alleviate_the_Burden_of_Annotation (дата обращения: 18.02.2025).

[19] См.: *Braun D.* I beg to difer : how disagreement is handled in the annotation of legal machine learning data sets // Artifcial Intelligence and Law. 2024. No. 32. P. 839–862. URL: https://doi.org/10.1007/s10506-023-09369-4

[20] См.: *Терещенко Л. К.* Правовой режим информации. М. : Юриспруденция, 2007. С. 101 ; *Войниканис Е. А.* Регулирование больших данных и право интеллектуальной собственности : общие подходы, проблемы и перспективы развития // Закон. 2020. № 7. С. 135–156.

[21] См.: Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union. URL: https://eur-lex.europa.eu/eli/reg/2018/1807/oj/eng (дата обращения: 13.03.2025).

ных[22], однако подобные методики, включая поощрение саморегулирования и особенно этического регулирования, должно распространяться и на индустрию аннотации данных. В это же направление должно включаться обеспечение разработки и принятия стандартов аннотации данных, возможно, создание единой системы аннотации, которая бы включала разбивку данных по различным показателям. Например, это может быть разбивка данных по полу, возрасту и другим атрибутам, что в итоге позволит понять характеристику наборов данных, способы из создания и т. д.[23] Отсутствие единых подходов и требований к аннотации данных снижает качество данных, негативным образом влияет на обучение моделей искусственного интеллекта, делает общедоступные наборы данных непригодными для использования[24]. В литературе отмечается эффективность краудсорсинга в сфере аннотации, привлечения различных платформ, что позволит снизить затраты на аннотацию данных[25]. При наличии стандартизации можно преодолеть проблемы этого подхода, связанные с обеспечением качества услуг.

В-третьих, необходимо обеспечить взаимодействие различных субъектов в сфере аннотации данных, включая государственный сектор, частные компании, образовательные организации. Российские образовательные организации высшего образования разрабатывают различные дисциплины по системам аннотации данных, при этом практической составляющей обучения является, в частности, создание разметки, датасетов, на которых впоследствии будут обучаться нейросети. Взаимодействие с образовательными организациями на системной основе позволит привлекать студентов различных специальностей для увеличения объема аннотируемых данных, учитывая специфику самих сведений (например, применительно к большим юридическим, большим медицинским и другим типам данных).

В-четвертых, решение кадрового вопроса в сфере индустрии аннотации данных. Данное направление должно являться частным случаем федерального проекта «Кадры для цифровой трансформации» в рамках национального проекта «Экономика данных»[26]. Однако пока в рамках стратегического планирования говорится лишь о количественных показателях обучающихся по IT-специальностям, а также лицах, прошедших практико-ориентированное обучение и т. п. Вместе с тем считаем, что нужно отдельно выделить приоритетные направления, в том числе касающиеся сферы развития индустрии аннотации данных.

В-пятых, формирование на государственном уровне аннотированных наборов данных в качестве эталонных коллекций. В настоящее время активно решается вопрос о формировании различных наборов данных государством с предоставлением различным субъектам возможности обрабатывать такие данные в рамках государственных информационных систем. Например, это касается правового режима составов данных, полученных в результате обезличивания[27]. Считаем, что как в рамках собственной деятельности, так и по результатам государственно-частного партнерства актуальным будет и формирование аннотируемых коллекций.

[22] См.: Guidance on sharing private sector data in the European data economy Accompanying the document Communication from the Commission to the European Parliament, the Council, the European economic and social Committee and the Committee of the Regions «Towards a common European data space» SWD/2018/125 final. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52018SC0125 (дата обращения: 13.03.2025).

[23] См.: Глобальный атлас регулирования искусственного интеллекта. Вектор БРИКС / Е. О. Васин, А. Д. Гвоздырева, А. П. Дейнека [и др.] ; под ред. А. В. Незнамова. 3-е изд., перераб. и доп. Доступ из справ.-правовой системы «КонсультантПлюс».

[24] См.: *Алейникова Д. В.* Аннотирование данных как объект обучения студентов социальногуманитарной направленности // Вестник Моск. гос. лингв. ун-та. Образование и педагогические науки. 2022. Вып. 4 (845). С. 15–19.

[25] См.: *Гилязев Р. А., Турдаков Д. Ю.* Активное обучение и краудсорсинг : обзор методов оптимизации разметки данных // Труды ИСП РАН. 2018. Т. 30, вып. 2. С. 215–250.

[26] См.: Основные показатели и мероприятия национального проекта «Экономика данных и цифровая трансформация государства» . URL: http://government.ru/info/54314/ (дата обращения: 13.03.2025).

[27] См.: О внесении изменений в Федеральный закон «О персональных данных» и Федеральный закон «О проведении эксперимента по установлению специального регулирования в целях создания необходимых условий для разработки и внедрения технологий искусственного интеллекта в субъекте Российской Федерации – городе федерального значения Москве и внесении изменений в статьи 6 и 10 Федерального закона “О персональных данных”» : федер. закон от 8 августа 2024 г. № 233-ФЗ // Собр. законодательства Рос. Федерации. 2024. № 33 (ч. I). Ст. 4929.

В-шестых, развитие технологий по автоматизированной аннотации данных. Так, задача по оптимизации процесса аннотации может быть решена путем поддержки технологических решений в сфере аннотации, новых разработок по созданию специализированных моделей для аннотации данных различных отраслей.

Итак, индустрия аннотации данных обладает огромным потенциалом для реализации повторного использования данных и особенно Big Data, обеспечения взаимодействия между различными субъектами, создающими, обрабатывающими и использующими данные. Аннотация данных повышает качество наборов данных, обеспечивает развитие искусственного интеллекта и иных смежных технологий. На этом основании полагаем, что в Российской Федерации будет актуальным как минимум стратегическое регулирование, которое определит направления для обеспечения развития индустрии аннотации данных, включая государственно-частное партнерство, взаимодействие с образовательными организациями, развитие кадрового потенциала, обеспечение стандартизации, будет стимулировать саморегулирование.

Библиографический список

Алейникова Д. В. Аннотирование данных как объект обучения студентов социальногуманитарной направленности // Вестник Моск. гос. лингв. ун-та. Образование и педагогические науки. 2022. Вып. 4 (845). С. 15–19.

Войниканис Е. А. Регулирование больших данных и право интеллектуальной собственности : общие подходы, проблемы и перспективы развития // Закон. 2020. № 7. С. 135–156.

Гилязев Р. А., Турдаков Д. Ю. Активное обучение и краудсорсинг : обзор методов оптимизации разметки данных // Труды ИСП РАН. 2018. Т. 30, вып. 2. С. 215–250.

Глобальный атлас регулирования искусственного интеллекта. Вектор БРИКС / Е. О. Васин, А. Д. Гвоздырева, А. П. Дейнека [и др.] ; под ред. А. В. Незнамова. 3-е изд., перераб. и доп. Доступ из справ.-правовой системы «КонсультантПлюс».

Государство, общество и личность : пути преодоления вызовов и угроз в информационной сфере : монография / Н. С. Волкова, А. А. Ефремов, С. М. Зырянов [и др.] ; отв. ред. Л. К. Терещенко. М. : Инфотропик Медиа, 2024. 352 с.

Макафи Э., Бриньолфсон Э. Машина, платформа, толпа. Наше цифровое будущее. М., 2019, 368 с.

Назаров Н. А. Обеспечение качества данных при автоматизированном принятии решений в государственном управлении // Журнал российского права. 2024. № 5. С. 140–155.

Терещенко Л. К. Правовой режим информации. М. : Юриспруденция, 2007. 192 с.

Харитонова Ю. С., Савина В. С., Паньини Ф. Предвзятость алгоритмов искусственного интеллекта : вопросы этики и права // Вестник Перм. ун-та. Юридические науки. 2021. № 3. С. 488–515.

Чайка М. Практический подход к валидации рейтинговых моделей при реализации ПВР-подхода : методика 5×5 // Риск-менеджмент в кредитной организации. 2024. № 1. С. 19–34.

Braun D. I beg to difer: how disagreement is handled in the annotation of legal machine learning data sets // Artifcial Intelligence and Law. 2024. No. 32. P. 839–862. URL: https://doi.org/10.1007/s10506-023-09369-4

Gray M., Savelka J., Oliver W., Ashley K. Can GPT Alleviate the Burden of Annotation? // Legal Knowledge and Information Systems. 2013. URL: https://www.researchgate.net/publication/376422423_Can_GPT_Alleviate_the_Burden_of_Annotation

Subinay A, Procheta S., Dwaipayan R., Kripabandhu Gh. A case study for automated attribute extraction from legal documents using large language models // Artificial Intelligence and Law. 2024. November.

References

Aleynikova D. V. Annotation of data as an object of education for students of socio-humanitarian orientation // Bulletin of the Moscow State Linguistic University. un-ta. Education and pedagogical sciences. 2022. Issue 4 (845). P. 15–19.

Voynikanis E. A. Regulation of big data and intellectual property law : general approaches, problems and development prospects // Law. 2020. No. 7. P. 135–156.

Gilyazev R. A., Turdakov D. Yu. Active learning and crowdsourcing : a review of data markup optimization methods // Proceedings of the ISP RAS. 2018. Vol. 30, issue. 2. P. 215–250.

Global Atlas of Artificial Intelligence regulation. Vector BRICS / E. O. Vasin, A. D. Gvozdyreva, A. P. Deineka [et al.] ; ed. by A. V. Neznamov. 3rd ed., revised and add. Access from the legal reference system «ConsultantPlus».

The state, society and personality : ways to overcome challenges and threats in the information sphere : a monograph / N. S. Volkova, A. A. Efremov, S. M. Zyryanov [et al.] ; ed. by L. K. Tereshchenko. Moscow : Infotropik Media, 2024. 352 p.

Mcafee E., Brynjolfson E. Car, platform, crowd. Our digital Future. Moscow, 2019. 368 p.

Nazarov N. A. Data quality assurance in automated decision-making in public administration // Journal of Russian Law. 2024. No. 5. P. 140–155.

Tereshchenko L. K. The legal regime of information. Moscow : Jurisprudence, 2007. 192 p.

Kharitonova Yu. S., Savina V. S., Panini F. Bias of artificial intelligence algorithms : issues of ethics and law // Bulletin of Perm. un-ta. Legal sciences. 2021. No. 3. P. 488–515.

Chaika M. A practical approach to the validation of rating models in the implementation of the PVR approach : the 5×5 methodology // Risk management in a credit institution. 2024. No. 1. P. 19–34.

Braun D. I beg to difer: how disagreement is handled in the annotation of legal machine learning data sets // Artifcial Intelligence and Law. 2024. No. 32. P. 839–862. URL: https://doi.org/10.1007/s10506-023-09369-4

Gray M., Savelka J., Oliver W., Ashley K. Can GPT Alleviate the Burden of Annotation? // Legal Knowledge and Information Systems. 2013. URL: https://www.researchgate.net/publication/376422423_Can_GPT_Alleviate_the_Burden_of_Annotation

Subinay A, Procheta S., Dwaipayan R., Kripabandhu Gh. A case study for automated attribute extraction from legal documents using large language models // Artificial Intelligence and Law. 2024. November.

Национальный исследовательский университет «Высшая школа экономики» (Москва)

Лескина Э. И., кандидат юридических наук, доцент департамента права цифровых технологий и биоправа факультета права

E-mail: elli-m@mail.ru

Поступила в редакцию: 14.03.2025

Для цитирования:

Лескина Э. И. Правовое регулирование индустрии аннотации данных как способ обеспечения качества данных // Вестник Воронежского государственного университета. Серия: Право. 2025. № 3 (62). С. 64–71. DOI: https://doi.org/10.17308/law/1995-5502/2025/3/64-71

Hational Recearch University «Higher School of Economics» (Moscow)

Leskina E. I., Candidate of Legal Sciences, Associate Professor of the Department of Digital Technology Law and Biolaw

E-mail: elli-m@mail.ru

Received: 14.03.2025

For citation:

Leskina E. I. Legal support for the data annotation industry as a way to ensure data quality // Proceedings of Voronezh State University. Series: Law. 2025. № 3 (62). P. 64–71. DOI: https://doi.org/10.17308/law/1995-5502/2025/3/64-71