Информационное общество и право

СТАНДАРТИЗАЦИЯ РАБОТЫ С БОЛЬШИМИ ДАННЫМИ: МЕЖДУНАРОДНЫЕ И НАЦИОНАЛЬНЫЕ СТАНДАРТЫ

Аверкин Алексей Николаевич

Кандидат физико-математических наук, доцент
Российский экономический университет им. Г.В. Плеханова, Учебно-научная лаборатория искусственного интеллекта, нейротехнологий и бизнес-аналитики, ведущий научный сотрудник
Москва, Российская Федерация
averkin2003@inbox.ru

Афанасьев Сергей Дмитриевич

Кандидат юридических наук
МГУ имени М.В.Ломоносова, Национальный центр цифровой экономики, ведущий специалист
Московский государственный областной университет, Институт экономики, управления и права, юридический факультет, кафедра конституционного и муниципального права, доцент
Москва, Российская Федерация
sergei.afanasev@digital.msu.ru

Микрюков Андрей Александрович

Кандидат технических наук, доцент
Российский экономический университет им. Г.В. Плеханова, Институт математики, информационных систем и цифровой экономики, Кафедра прикладной информатики и информационной безопасности, доцент
Москва, Российская Федерация
mikrukov.aa@rea.ru

Паджев Валентин Валентинович

Институт развития информационного общества, руководитель Дирекции правовых программ
Москва, Российская Федерация
vpadzhev@iis.ru

Райков Александр Николаевич

Доктор технических наук, профессор
Институт проблем управления имени В.А. Трапезникова РАН, ведущий научный сотрудник
МГУ имени М.В.Ломоносова, Национальный центр цифровой экономики, руководитель департамента интеллектуальных технологий
Москва, Российская Федерация
Alexander.N.Raikov@gmail.com

Хохлов Юрий Евгеньевич

Кандидат физико-математических наук, доцент
Институт развития информационного общества, председатель Совета директоров
РЭУ имени Г.В. Плеханова, научный руководитель базовой кафедры цифровой экономики ИРИО
Москва, Российская Федерация
yuri.hohlov@iis.ru

Храмцовская Наталья Александровна

Кандидат исторических наук
ООО «Электронные офисные системы», ведущий эксперт по управлению документацией
Москва, Российская Федерация
sspchram@tochka.ru

---

https://doi.org/10.52605/16059921_2021_04_220

Аннотация

Стандарты, устанавливающие требования к технологиям работы с большими данными, призваны повысить эффективность использования этих технологий в различных отраслях экономики. В статье исследуются особенности международных и российских национальных общеотраслевых стандартов в области данных, основанных на стандартах ISO/IEC 20546, серий 20547-X, 9000 и проектов серии 5259-X, в сфере основных понятий, эталонной архитектуры и требований к большим данным и их качеству, а также подходы к стандартизации требований заказчика к действиям, связанным с использованием больших данных. Делается вывод, что разработка документов по стандартизации больших данных в России имеет большое значение и влечёт необходимость наращивания темпов стандартизации.

Ключевые слова

большие данные; данные; стандартизация; национальный стандарт; международный стандарт; ГОСТ; ISO; Росстандарт; эталонная архитектура; качество данных; искусственный интеллект

Введение

«Большие данные» рассматриваются в качестве одной из ключевых технологий, на которые опирается быстро набирающее популярность направление искусственного интеллекта. Хотя споры о том, как наилучшим образом определить понятие «большие данные», по-прежнему продолжаются, к настоящему времени сложился определенный международный консенсус на основе определения, предложенного в международном стандарте ISO/IEC 20546:2019 [1], адаптированном теперь в России.

В ноябре 2021 года вступил в силу национальный стандарт ГОСТ Р ИСО/МЭК 20546–2021 «Информационные технологии. Большие данные. Обзор и словарь» [2], в котором закрепляется следующее определение больших данных (п. 3.1.2):

> Большие данные (big data): Большие массивы данных, отличающиеся главным образом такими характеристиками, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа

Важным уточнением данного термина является содержащееся в примечании к пункту 3.1.2 положение о том, что «термин «большие данные» широко применяется в различных значениях, например, в качестве названия технологии масштабирования, используемой для обработки больших массивов данных».

В готовящемся к публикации международном стандарте по концепциям и терминологии искусственного интеллекта ISO/IEC 22989 [3], в частности, отмечается следующая связь больших данных с системами искусственного интеллекта.

«У больших данных имеется множество применений, в том числе в системах искусственного интеллекта (ИИ). Большие данные делают возможными многие ИИ-системы. Наличие больших коллекций неструктурированных данных в различных областях применения способствует получению новых идей и знаний в результате использования таких методов ИИ, как выявление знаний (knowledge discovery) и распознавание закономерностей (pattern recognition). Доступность огромных объёмов данных для обучения приводит к появлению улучшенных моделей машинного обучения, способных решать задачи в широком спектре приложений» (п. 8.5.1).

Технологии работы с большими данными позволяют обрабатывать большие объёмы данных по сравнению со «стандартными», работать с потоками быстро поступающих данных в очень больших объёмах, а также обрабатывать как структурированные, так и неструктурированные или слабо структурированные данные. Такие особенности данного класса цифровых технологии позволяют выявить скрытые закономерности, ускользающие от ограниченного человеческого восприятия. Это даёт беспрецедентные возможности оптимизации различных сфер деятельности: от системы государственного управления, медицины, образования, финансов или транспорта до промышленного производства или телекоммуникаций.

Большие данные представляют собой, по сути, метод «последней надежды», к которому прибегают там, где невозможны прямые измерения и возникает необходимость опираться на косвенные данные больших объёмов и разнообразия. К таким ситуациям, главным образом,

относятся продвинутые инновационные технологии обработки данных с целью выявления неочевидных закономерностей, часто в ситуациях, когда технические особенности потоков данных (объёмы, скорость, разнообразие и т.д.) таковы, что традиционные, давно разработанные методы обработки «небольших» массивов, как правило, структурированных данных с решением такой задачи не справляются. Однако сводить большие данные только к этим техническим особенностям не совсем верно, поскольку даже за последние десятилетия на наших глазах представления о том, что такое большой объём, изменились многократно – от килобайта в середине 1980-х годов до экзабайтов в настоящее время, и эти изменения в восприятии продолжаются. Вследствие такого изменения восприятия многие технологии, которые в прошлом веке относили к большим данным, сегодня перестали считаться таковыми, поскольку они стали общедоступными и широко распространёнными.

Новые проблемы и растущая вычислительная мощность будут стимулировать разработку новых аналитических методов. Существует также потребность в постоянных инновациях в технологиях и методах, которые помогут отдельным лицам и организациям интегрировать, анализировать, визуализировать и потреблять растущий поток больших данных. Таким образом, использование технологий работы с большими данными в современном мире имеет первостепенное значение. Уровень развития и использования информационных технологий в государственном управлении, бизнесе, социальной сфере, научных исследованиях и иных сферах жизни общества достиг высокого уровня зрелости, что влечёт за собой необходимость выработки стандартизированных подходов к применению таких технологий, включая технологии работы с большими данными.

Современное состояние развития технологий работы с большими данными характеризуется отсутствием устоявшихся подходов не только к определению и основным характеристикам самих больших данных, но и к эталонной архитектуре систем для работы с большими данными. Значительная часть специалистов, если они не являются узкопрофильными специалистами по работе с данными, продолжает понимать под большими данными такие наборы данных, которые отличаются большим объёмом, разнообразием и высокой скоростью обработки, в то время как понятие «большие данные» представляет собой намного более многосложную конструкцию. Вряд ли можно себе представить, например, что данные о температуре тела больного, получаемые миллиард раз в секунду, представляют собой именно большие данные как они должны пониматься для грамотного использования и эффективного применения. Научные работники и профессиональные инженеры хорошо понимают, что когда есть выбор между целевым и контролируемым образом собранными «под задачу» «малыми данными» и обработкой, скрывающейся за красивым титулом «больших данных», – этот выбор очевиден и это не большие данные.

Полезно также вспомнить, что технологии работы с большими данными первоначально были разработаны разведывательными службами в середине прошлого века и использовались для получения сведений, которые по тем или иным причинам нельзя было собрать средствами агентурной разведки и дистанционного зондирования – например, посредством сбора и анализа информации из массовых открытых источников. Корректное понимание больших данных и их характеристик актуализируется при разработке соответствующих информационных систем и технологий, которые должны быть совместимы между собой. Устаревшие системы и несовместимые форматы часто препятствуют интеграции данных и основанной на них аналитике [4], которая и является основной целью и ценностью использования технологий работы с большими данными. Кроме того, как и для любой инновации важно понять, какую отдачу, при каких обстоятельствах и затратах технологии работы с большими данными способны обеспечить, какие риски с ними связаны. Решения об использовании или неиспользовании больших данных желательно принимать на основе баланса отдачи, затрат и рисков.

Решение соответствующих проблем заключается, в том числе, в стандартизации технологий, что позволит обеспечить единообразие подходов к пониманию и применению технологий больших данных. В процессе стандартизации устанавливаются общие характеристики, правила и принципы в отношении объекта стандартизации, что направлено на достижение добровольной и многократно применяемой упорядоченности соответствующих объектов. Целями стандартизации технологий работы с большими данными являются содействие социально-экономическому развитию, а также содействие интеграции России в мировую экономику и в международные системы стандартизации в качестве равноправного партнёра.

Предметом данной статьи является описание и анализ процессов стандартизации технологий работы с большими данными на международном и национальном уровнях. Здесь также приводятся: краткая история возникновения и развития тематики больших данных, принятые и разрабатываемые стандарты в области работы с большими данными в областях: терминология и направления стандартизации больших данных в целом, включая взаимосвязь со смежными стандартами для других информационных технологий; варианты использования больших данных; эталонная архитектура больших данных, её концептуальная схема и обеспечение безопасности; стандартизация требований заказчика к действиям, связанным с оперированием большими данными; обеспечение качества данных.

1 Краткая история стандартизации больших данных

Активная деятельность по разработке и внедрению национальных и международных стандартов в области больших данных началась с 2013 г., в первую очередь – Национальным институтом стандартов и технологий США (NIST), в рамках которого в июне 2013 г. создаётся рабочая группа по большим данным для стандартизации эталонной архитектуры больших данных, в которую входят представители основных заинтересованных сторон – бизнеса, власти и научно-образовательного сообщества. В 2015 г. NIST принимает серию стандартов в сфере терминологии, архитектуры больших данных, безопасности и конфиденциальности персональных данных при использовании соответствующих технологий [5], которые были пересмотрены в 2018 [6] и 2019 [7] гг.

С развитием соответствующих технологий было признано необходимым гармонизировать разнообразные подходы к пониманию и применению технологий больших данных, чему способствовала разработка и принятие ряда международных стандартов в рамках объединенного технического комитета Международной организации по стандартизации и Международной электротехнической комиссии (ISO/IEC; ИСО/МЭК), а также координация действий ИСО/МЭК с Международным союзом электросвязи (ITU; МСЭ). Стандартизацию в области больших данных на международном уровне можно условно разделить на два исторических этапа.

Первый этап – разработка и утверждение основополагающих международных стандартов. В ноябре 2013 г. Международный союз электросвязи опубликовал аналитический доклад «Большие данные: сегодня большие, завтра нормальные» [8] в котором анализировался феномен больших данных, нашедших применение сразу в нескольких отраслях экономики, в том числе в секторе телекоммуникаций. В начале 2015 г. был опубликован предварительный доклад Объединённого технического комитета № 1 ИСО/МЭК (ISO/IEC JTC 1) «Большие данные» [9], в котором описаны проблемы и направления дальнейшей международной стандартизации технологий работы с большими данными, представлена оценка состояния требований рынка стандартизации больших данных и предложены приоритеты стандартизации. Доклад в целом стал основой плана работ Подкомитета SC42 «Искусственный интеллект» (ISO/IEC JTC 1/SC 42 Artificial intelligence) и его рабочей группы «Большие данные» (ISO/IEC JTC 1/SC 42/WG 2 Big data), переименованной позднее в рабочую группу «Данные» (ISO/IEC JTC 1/SC 42/WG 2 Data).

В ноябре 2015 г. МСЭ утвердил первый международный стандарт в области больших данных «Большие данные – Требования на основе облачных вычислений и их возможности» [10], а уже в июле 2016 г. принимается дорожная карта стандартизации больших данных [11]. Вслед за этим январе-феврале 2018 г. утверждаются два первых международных стандарта ИСО/МЭК из пятикомпонентной серии стандартов ИСО/МЭК 20547-X, посвящённой стандартизации эталонной архитектуры больших данных: часть 2 «Варианты использования и производные требования» [12] и часть 5 «Направления стандартизации» [13] (остальные части данной серии утверждены в марте-сентябре 2020 г. [14–16]). В феврале 2019 года ИСО/МЭК утверждает основополагающий терминологический стандарт по большим данным. Таким образом, первоначальные международные стандарты были ориентированы, в первую очередь, на гармонизацию разрозненных представлений о больших данных и построение их единой стандартизированной концепции, а также утверждение общих положений, требований к их эталонной архитектуре и обеспечению безопасности данных (см. [17]).

Дальнейшим развитием международной стандартизации больших данных стала разработка в рамках ИСО/МЭК серии стандартов 5259-X [18–21], посвящённых качеству данных для аналитики и машинного обучения [см. например, 22–23] и описанию концептуальной схемы жизненного цикла работы с данными в системах искусственного интеллекта [24].

Второй этап – стандартизация требований к работе с большими данными в различных сферах деятельности. Этот этап стандартизации в области больших данных связан с разработкой стандартов для отдельных отраслей экономики и сфер деятельности. В первую очередь следует отметить деятельность МСЭ по стандартизации в телекоммуникационной отрасли, связанной с организаций сетей, обеспечивающих передачу больших данных [см. например, 25–26]. Важно отметить, что в процессах принятия стандартизации помимо трех главных международных организаций стандартизации (ИСО, МЭК, МСЭ) активную и не менее значимую роль играют международные организации, занимающиеся в том числе разработкой и утверждением отраслевых стандартов. К их числу можно отнести Организацию по развитию стандартов структурированной информации (Organization for the Advancement of Structured Information Standards, OASIS), Health Level Seven International, Американское общество по испытанию материалов (ASTM International), Консорциум отрытых геопространственных данных (Open Geospatial Consortium, OGC) и другие.

Стандартизация больших данных в России началась с существенным отставанием от зарубежных стран и международных организаций, в связи с чем на национальном уровне было принято решение ускоренными темпами сократить разрыв, разработать и ввести в действие ряд основополагающих национальных стандартов в области больших данных, гармонизированных с международными. Для решения этой задачи, равно как и для смежной проблемы развития отечественной стандартизации технологий искусственного интеллекта в 2019 г. был создан технический комитет по стандартизации «Искусственный интеллект» (ТК 164) [27], позиционирующийся как зеркальное отражение на национальном уровне профильного международного подкомитета ISO/IEC JTC 1 SC 42 Artificial Intelligence и объединивший более 100 заинтересованных государственных, коммерческих, научных и образовательных организаций и разработчиков в области технологий искусственного интеллекта и больших данных. Как и на международном уровне, стандартизация больших данных была передана в ведение ТК 164, где изначально была создана рабочая группа 02 «Большие данные», которая в 2020 году – вслед за изменениями в структуре профильного международного подкомитета – была преобразована в подкомитет 02 «Данные» (ТК 164/ПК 02) [28]. Функции секретариата данного подкомитета выполняет Национальный центр цифровой экономики Московского государственного университета имени М.В.Ломоносова. Кроме того, отдельные вопросы стандартизации соответствующих технологий входят в компетенцию технических комитетов по стандартизации «Информационные технологии» (ТК 22), «Криптографическая защита информации» (ТК 26), «Кибер-физические системы» (ТК 194), «Каталогизация продукции» (ТК 430).

Становление парадигмы данных как основы развития цифровой экономики привело к необходимости стандартизации технологий работы с большими данными в рамках развития передовых производственных технологий. Разработка и «локализация» документов по стандартизации больших данных в Российской Федерации были обозначены в качестве задач Национальной технологической инициативы и вошли в План мероприятий («дорожную карту») по направлению «Технет» (передовые производственные технологии) 2018 г. [29]. Позднее для реализации соответствующих мероприятий Министерством промышленности и торговли РФ и Федеральным агентством по техническому регулированию и метрологии был утвержден Перспективный план стандартизации в области передовых производственных технологий на 2018–2025 годы [30], а первоначальный перечень соответствующих стандартов по большим данным был включён в Программу национальной стандартизации на 2019 г. [31].

Впоследствии по причинам низкой экспертной оценки эффективности реализации соответствующей дорожной карты стандартизации «Технета» [32] и необходимости обеспечения единства разрабатываемых стандартов со стандартами в области искусственного интеллекта было принято решение объединить разработку и внедрение стандартов для искусственного интеллекта и больших данных в единой Дорожной карте развития «сквозной» цифровой технологии «Нейротехнологии и искусственный интеллект» [33]. Предполагалось, что реализация данной дорожной карты будет осуществляться в виде серии мероприятий федерального проекта «Нормативное регулирование цифровой среды» [34] в составе национальной программы «Цифровая экономика Российской Федерации», однако ни одно из мероприятий по стандартизации в рамках данного проекта так и не было реализовано. Несмотря на это, следует отметить, что работы по стандартизации больших данных ведутся с 2019 г. в рамках проекта «Мониторинг и стандартизация развития и использования технологий хранения и анализа больших данных в цифровой экономике Российской Федерации» по программе Центра

компетенций Национальной технологической инициативы «Центр хранения и анализа больших данных» при Московском государственном университете имени М.В.Ломоносова.

Новый импульс отечественная стандартизация больших данных получила в 2020 г. после принятия Национальной стратегии развития искусственного интеллекта на период до 2030 года [35] и утверждения нового федерального проекта «Искусственный интеллект» [36]. Министерством экономического развития Российской Федерации и Федеральным агентством по техническому регулированию и метрологии была разработана и утверждена Программа стандартизации по приоритетному направлению «Искусственный интеллект» на период 2021–2024 годы [37], которая финансируется из федерального проекта. Мероприятия по стандартизации больших данных являются частью этой программы и направлены на сокращение отставания национальной системы стандартизации от международной.

В настоящее время в России утверждён и вступил в силу гармонизированный ГОСТ Р ИСО/МЭК 20546–2021 «Информационные технологии. Большие данные. Обзор и словарь», положения которого идентичны международному стандарту ISO/IEC 20546:2019 [1]. Кроме того, на различных стадиях разработки находятся проекты национальных стандартов (см. Таблицу 1). Разработка национальных стандартов в сфере работы с большими данными, за исключением одного, закреплена за ТК 164 «Искусственный интеллект».

Таблица 1 — Состояние разработки проектов национальных стандартов Российской Федерации в области больших данных

Наименование проекта национального стандарта	**Международный аналог** (IDT – идентичный стандарт; MOD – модифицированный стандарт)	**Статус разработки**
ГОСТ Р «Информационные технологии. Эталонная архитектура больших данных. Часть 1. Структура и процесс применения» [38]	ISO/IEC TR 20547-1:2020 «Information technology — Big data reference architecture — Part 1: Framework and application process» (IDT) [14]	Доработка окончательного варианта и вынесение на голосование в ТК 164 «Искусственный интеллект»
ГОСТ-Р «Информационные технологии. Эталонная архитектура больших данных. Часть 2. Варианты использования и производные требования» [39]	ISO/IEC TR 20547-2:2018 «Information technology — Big data reference architecture — Part 2: Use cases and derived requirements» (IDT) [12]	Утверждение национального стандарта в Росстандарте
ПНСТ «Информационные технологии. Большие данные. Типовая архитектура»[1] [40]	ISO/IEC 20547-3:2020 Information technology — Big data reference architecture — Part 3: Reference architecture (MOD) [15]	Утверждение предварительного национального стандарта в Росстандарте
ГОСТ Р «Информационные технологии. Эталонная архитектура больших данных. Часть 5. Направления стандартизации» [41]	ISO IEC TR 20547-5:2018 «Information technology - Big data reference architecture - Part: 5: Standards roadmap» (IDT) [13]	Доработка окончательного варианта и вынесение на экспертизу в ТК 164 «Искусственный интеллект»
ГОСТ Р «Информационные технологии. Большие данные. Техническое задание. Требования к содержанию и оформлению» [42]	Разработан впервые, не имеет международных аналогов.	Утверждение национального стандарта в Росстандарте

[1] *ПНСТ «Информационные технологии. Большие данные. Типовая архитектура» закреплён за сферой компетенций ТК 194 «Киберфизические системы».*

Наименование проекта национального стандарта	**Международный аналог** (IDT – идентичный стандарт; MOD – модифицированный стандарт)	**Статус разработки**
ГОСТ Р «Информационные технологии – Искусственный интеллект – Структура управления процессами аналитики больших данных» [43]	ISO/IEC 24668 Information technology — Artificial intelligence — Process management framework for big data analytics (IDT) [44]	Доработка окончательного варианта и вынесение на голосование в ТК 164 «Искусственный интеллект»
ГОСТ Р «Информационные технологии. Искусственный интеллект. Структура жизненного цикла работы с данными»	ISO/IEC WD 8183 Information technology — Artificial intelligence — Data life cycle framework (MOD)	Сформировано предложение по включению разработки национального стандарта в Программу национальной стандартизации на 2022 г. последующие годы с плановым сроком утверждения в июне 2024 г.
ГОСТ Р «Информационные технологии. Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, термины и примеры»	ISO/IEC WD 5259-1 Data quality for analytics and ML – Part 1: Overview, terminology, and examples (MOD) [18]	Включён в Программу национальной стандартизации на 2021 г. и последующие годы с плановым сроком утверждения в декабре 2023 г.
ГОСТ Р «Информационные технологии. Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 2. Меры качества данных»	ISO/IEC AWI 5259-2 Data quality for analytics and ML – Part 2: Data quality measures (MOD) [19]	Включён в Программу национальной стандартизации на 2021 г. и последующие годы с плановым сроком утверждения в декабре 2023 г.
ГОСТ Р «Информационные технологии. Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 3. Требования и руководство по управлению качеством данных»	ISO/IEC WD 5259-3 Data quality for analytics and ML – Part 3: Data quality management requirements and guidelines (MOD) [20]	Включён в Программу национальной стандартизации на 2021 г. и последующие годы с плановым сроком утверждения в декабре 2023 г.
ГОСТ Р «Информационные технологии. Качество данных для аналитики и машинного обучения. Часть 4. Структура процесса повышения качества данных»	ISO/IEC WD 5259-4 Data quality for analytics and ML – Part 4: Data quality process framework (MOD) [21]	Включён в Программу национальной стандартизации на 2021 г. и последующие годы с плановым сроком утверждения в декабре 2023 г.

2 Анализ публикационной активности в области стандартизации больших данных

Для достижения поставленной цели исследования, помимо анализа международных и национальных инициатив по стандартизации больших данных был проведен анализ научных публикаций и аналитических материалов по данной теме. В качестве источника научных публикаций рассматривались наиболее авторитетные международные библиографические базы данных Web of Science (WoS) и Scopus, для публикации в которых внимательно отбираются рецензируемые научные издания. Первая отмечается как содержащая тщательно отобранные более качественные источники, вторая – как более широкая по охвату публикаций [45].

На основе консультаций с экспертами в области стандартизации и технологий работы с большими данными были сформированы поисковые образы, учитывающие различные варианты написания терминов. Для наиболее полного отражения картины публикационной активности в сфере стандартизации больших данных был выбран период с 2011 г., когда тематика

стандартизации больших данных стала актуальной, по 2020 г.[2]. При анализе публикационной активности наиболее распространённым и эффективным вариантом проведения поиска является поиск по полям «Тема» (поиск по названию, аннотации, автору и ключевым словам) для WoS[3] и «название документа, краткое описание, ключевые слова» (оператор TITLE-ABS-KEY) для Scopus[4]. В соответствии с принятым подходом анализ публикационной активности был осуществлён по следующим этапам.

1. Из «ядерных» коллекций WoS и Scopus был выделен массив публикаций по большим данным, определённый по следующему перечню ключевых слов и словосочетаний: big data; data analytic; data mining; data science; descriptive analytic; diagnostic analytic; Hadoop; large dataset; MapReduce; massive data; predictive analytic; prescriptive analytic; semi-structured data; text mining; unstructured data. Детальное описание процедуры формирования приведенного выше поискового образа описан в статье [46].

За временной период 2011–2020 гг. поисковый запрос выдал 149 349 результатов в базе WoS и 289 514 результатов в базе Scopus.

2. На следующем этапе из сформированного массива публикаций осуществлялся отбор публикаций, имеющих отношение к стандартизации технологий работ с большими данными, путём задания специализированного поискового образа со следующими ключевыми словами и словосочетаниями: committee-based standardization; consortia standard; data quality; data standards; formal standard; government-based standardization; international standard; ISO/IEC; multi-mode standardization; national standard; proprietary standard; reference architecture; standard; standardization union; standardization; standards system.

Отбор указанных ключевых слов и словосочетаний основывался на предварительном исследовании, включающем в себя анализ релевантных публикаций и формирование исходного перечня специфичных для данной области ключевых слов с участием экспертов. Приведённые ключевые слова и словосочетания позволяют получить максимальное покрытие исследуемой области, однако дают значительное количество «шума» – публикаций, слабо либо вообще не связанных с темой исследования, что подтверждается экспериментальной работой по ознакомлению с отобранным массивом публикаций и экспертной оценкой поисковых выдач. Это объясняется, в первую очередь, особенностями и многозначностью использования термина «стандарт» и производных от него в различных отраслях науки. Например, термин «стандарт» и производные от него могут использоваться для обозначения как собственно стандарта как нормативно-технического документа, так и эталонных (стандартизированных) подходов, практик (например, «стандартной» модели качества данных, или «стандартной» практики лечения, или «стандарта» качества услуги, или «золотого стандарта»); общепринятого, обычного подхода, не имеющего формализованной формы (например, «стандартизированные» вопросы; единообразного формата или типа (например, «стандартный» формат файла); этап обработки больших данных (например, «стандартизация» данных как обозначение приведения их к единому формату).

Поисковые запросы первого и второго этапов были объединены логическим оператором «AND», а поиск по ним за временной период 2011–2020 гг. дал следующие результаты: 11 122 публикации (из них 553 обзорные) в базе данных WoS и 21 194 публикации (из них 828 обзорные) в базе данных Scopus. Распределение указанных публикаций по годам в разрезе баз данных представлено на рисунке 1. Интересно при этом отметить, что за временной период с 1998 г. (когда был введён в научный оборот термин «большие данные» (big data)) по 2010 г. было найдено 2 062 публикации (из них 57 обзорные) в WoS и 4 096 публикаций (из них 161 обзорная) в Scopus.

2 Данные о публикациях могут подгружаться и обрабатываться в реферативных базах с задержкой, в связи с чем было принято решение не учитывать публикации 2021 г.

3 Показатели публикационной активности для базы данных Web of Science приведены на основе выгрузок, произведённых 20.08.2021.

4 Показатели публикационной активности для базы данных Scopus приведены на основе выгрузок, произведённых 27.08.2021.

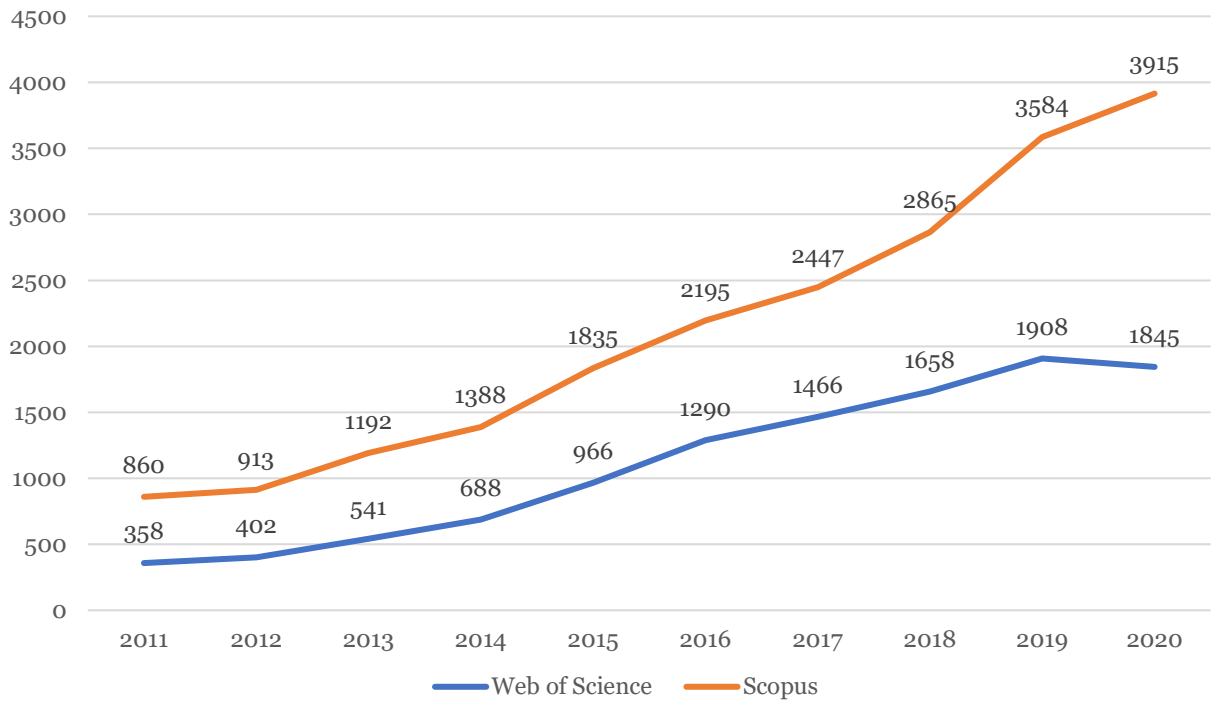

Рисунок 1 — Показатели публикационной активности по стандартизации больших данных, 2011–2020 гг. Источники: WoS, Scopus

Приведённый график демонстрирует устойчивый рост публикационной активности в сфере стандартизации технологий работы с большими данными; при этом можно отметить незначительное падение в WoS за 2020 г. что, скорее всего, объясняется процедурой индексирования статей, которая может продолжаться в течение полутора-двух лет после публикации.

Выборочный анализ наименований и аннотаций ста научных публикаций из найденного массива показал, что большинство публикаций не относится к предмету настоящего исследования, в связи с чем было принято решение для публикаций, имеющих отношение к стандартизации технологий работ с большими данными провести поиск только по полю «название публикации», что признаётся допустимой поисковой стратегией [47]. В связи с этим было принято решение сузить поисковый запрос по следующим параметрам: для публикаций в сфере больших данных оставить поиск в пределах «темы» в базе WoS или «названия документа, краткого описания, ключевых слов» (оператор TITLE-ABS-KEY) в базе Scopus, для публикаций, относящихся к стандартизации больших данных, применить поиск в пределах параметров «заголовок» и «ключевые слова автора» в базе WoS и «название документа» (оператор TITLE) и ключевые слова (оператор KEY) в базе Scopus. Такой поисковый запрос за временной период 2011–2020 гг. выдал 1140 результатов, из них 322 результата (включая 18 обзорных публикаций) в базе WoS и 818 результатов (включая 41 обзорная публикация) в базе Scopus.

Сравнительный анализ наименований, аннотаций и ключевых слов ста случайно выбранных публикаций, найденных на предыдущем этапе, и публикаций, найденных на данном этапе, показал, что наиболее релевантные результаты даёт последний вариант поискового запроса.

На следующем этапе был проведён анализ полученных результатов поиска на предмет соответствия исследуемой теме. Первоначальное выборочное ознакомление с наименованиями, ключевыми словами, аннотациями и полными текстами показало, что в большинстве публикаций предмет исследования либо значительно отличается, либо в малой степени затрагивает вопросы стандартизации технологий работы с большими данными (как правило, на уровне упоминаний в тексте). Исходя из этого был проведён анализ полной выборки публикаций по наименованиям и аннотациям. В результате было отобрано 305 уникальных публикаций за период 2011–2020 гг.,

относящихся к тематике стандартизации больших данных (см. рисунок 2), которые учитывались при анализе публикационной активности.

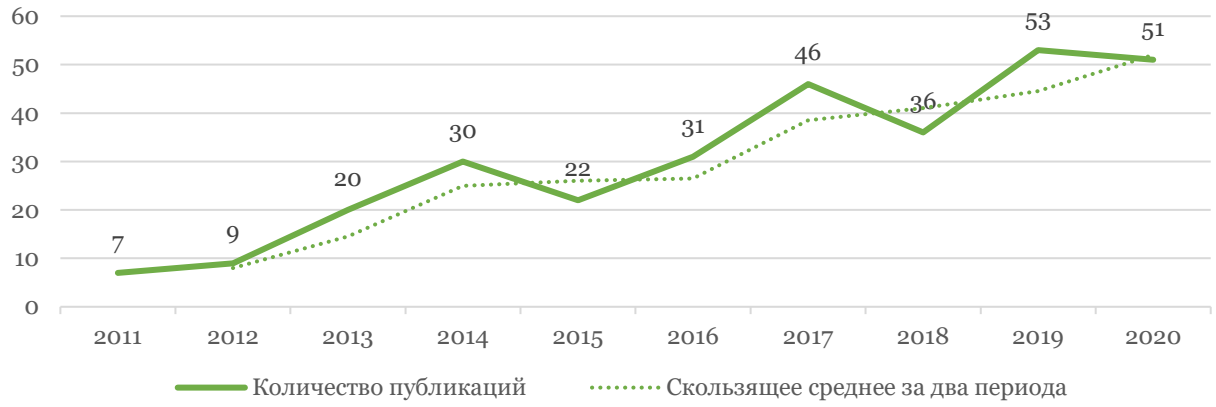

Рисунок 2 — Специализированные публикации по стандартизации больших данных, 2011–2020 гг. Источники: WoS, Scopus

Общее количество специализированных публикаций имеет устойчивую тенденцию постепенного роста с 2011 г. по 2020 г., за исключением периодов снижения публикационной активности в 2015 г. и 2018 г. Представляется, что отмеченный рост имеет прямую зависимость от истории становления и развития сферы стандартизации больших данных (см. раздел 1). Наиболее заметный рост публикационной активности в 2013–2014 гг. связан с принятием первых стандартов в области больших данных Национальным институтом стандартов и технологий США (NIST) [5–7]. Аналогичный всплеск публикационной активности в 2019–2020 гг. связывается с началом работ по международной стандартизации больших данных МСЭ и Объединенным техническим комитетом №1 ИСО/МЭК.

Анализ тематической направленности публикационного массива показал, что все статьи разбиваются на две категории: общесистемные вопросы стандартизации больших данных (такие как терминологическая база, эталонная архитектура, качество данных, анализ данных, безопасность/конфиденциальность при работе с данными) и стандартизация работы с большими данными в отдельных сферах деятельности (здравоохранение, государственное управление, бизнес, промышленность, наука и другие). Распределение публикаций по тематике приведено ниже на рисунках 3 и 4 соответственно.

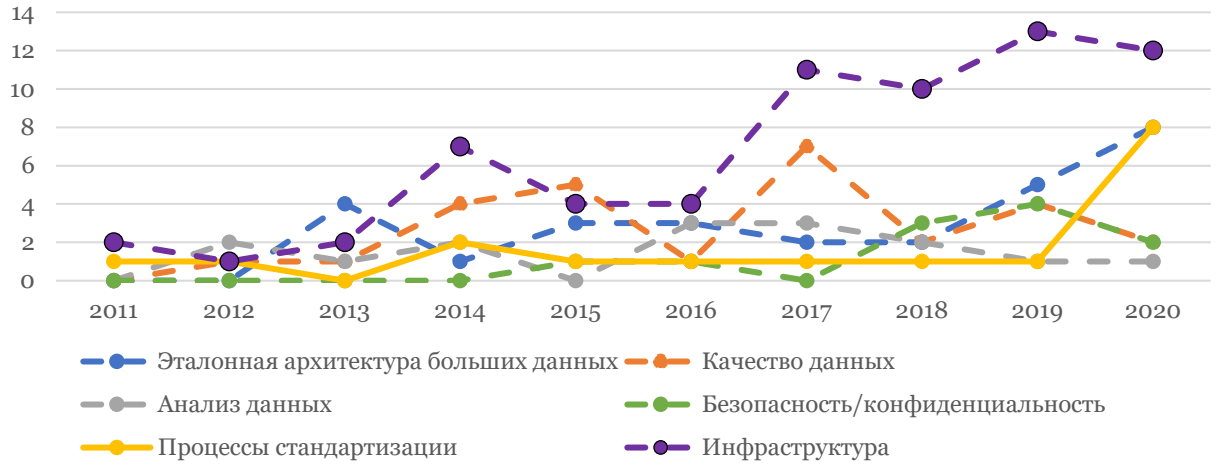

Рисунок 3. Публикационная активность в сфере стандартизации технологий, решений и сервисов работы с большими данными, 2011–2020 Источники: WoS, Scopus

На рисунке 3 представлена динамика публикационной активности, связанной со стандартизацией технологий, решений и сервисов работы с большими данными: инфраструктура работы с большими данными, включая информационные системы (66 статей); эталонная архитектура больших данных (28); качество больших данных (27); аналитика больших данных (15); безопасность/конфиденциальность при работе с большими данными (11); процессы стандартизации в сфере больших данных (17). В категорию «процессы стандартизации» объединены публикации по истории и методологии разработки и ввода в действие стандартов в области больших данных, практике применения стандартов, а также стандартизации процессов взаимодействия и обмена данными.

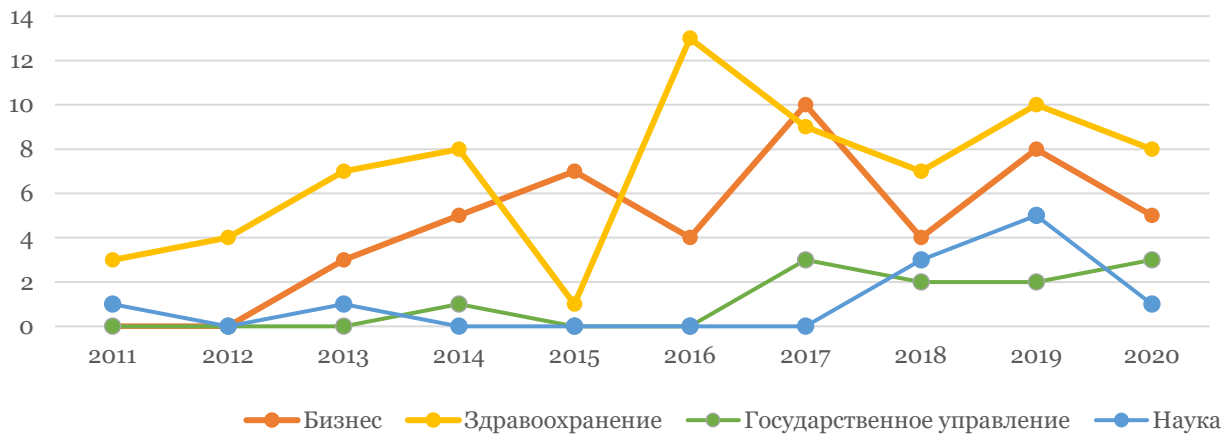

Рисунок 4 – Публикационная активность в сфере стандартизации работы с большими данными в отдельных сферах деятельности, 2011–2020
Источники: WoS, Scopus

На рисунке 4 представлена динамика публикационной активности, связанной со стандартизацией работы с большими данными в четырех сферах деятельности с максимальным числом публикаций. В целом следует отметить тенденцию увеличения числа таких публикаций, что отражает повышение спроса на стандартизацию больших данных в различных прикладных экономических и социальных аспектах. Анализ динамики демонстрирует стабильно высокий интерес экспертов в сфере стандартизации технологий работы с большими данными в здравоохранении, включая вопросы создания и использования медицинских баз данных. При этом постепенно увеличивается, хоть и менее интенсивно, количество публикаций, связанных со стандартизацией использования больших данных в системе государственного управления и в науке.

Подводя итоги, можно сделать несколько выводов о тенденциях публикационной активности в сфере стандартизации больших данных:

- на период 2015–2016 гг. приходится первый пик публикационной активности по стандартам в области работы с большими данными, что связано, в первую очередь, с появлением доклада «Большие данные» [9] и началом разработки международных стандартов; второй пик приходится на 2019-2020 гг. в связи с принятием терминологического стандарта ИСО/МЭК 20546 [1] и серии стандартов ИСО/МЭК 20547-X «Эталонная архитектура больших данных» [12–16];
- в последние пять лет происходит «расщепление» публикационной активности на тематику, связанную с общесистемными вопросами стандартизации больших данных, сопровождающуюся уменьшением числа таких публикаций и увеличении числа публикаций о стандартизации работы с большими данными в отдельных отраслях экономики, сферах социальной жизни и системы государственного управления (исключением сферы здравоохранения, где публикационная активность остается стабильно высокой);
- начиная с 2014 г. возрастает количество публикаций в области стандартов по различным аспектам работы с данными и, прежде всего, стандартов на качество данных, в том числе в различных отраслях (см., например, [48–51]);

- начиная с 2017 г. начинается активная подготовка и публикация результатов исследований по построению эталонных архитектур больших данных для различных сфер деятельности (см. например, [52–54]).

В настоящее время в России ведется активная работа по разработке и гармонизации общесистемных стандартов в области работы с большими данными, поэтому в рамках данной статьи вопросы стандартизации работы с большими данными в конкретных сферах деятельности не рассматриваются. Дальнейшее исследование будет использовать результаты проведенного анализа публикационной активности, действующие международные стандарты и проекты стандартов (прежде всего – ИСО/МЭК), а также проекты национальных стандартов Российской Федерации в области больших данных.

3 Большие данные: терминология и направления стандартизации

Зонтичный характер базового понятия «большие данные» породил две серьезные проблемы для потенциальных разработчиков технологий для работы с большими данными. Первая проблема состоит в многозначности и неопределенности самого термина, о которых шла речь во введении к данной статье, а вторая проблема связана с необходимостью единого подхода к стандартизации архитектуры системы для работы с большими данными.

Именно этими проблемами в первую очередь озаботилась рабочая группа «Большие данные» (впоследствии переименованная в рабочую группу «Данные») подкомитета 42 «Искусственный интеллект» Объединенного технического комитета №1 ИСО/МЭК. В 2019 г. была завершена трехлетняя работа над международным терминологическим стандартом «Большие данные. Обзор и словарь» [1], а в 2020 г.– работа над серией стандартов «Эталонная архитектура больших данных» [12–16], которые подробно рассматриваются далее.

Терминологический стандарт ISO/IEC 20546:2019 «Information technology — Big data — Overview and vocabulary» [1] представляет собой компромисс между несколькими конкурирующими подходами к определению больших данных и связанными с ними технологиями и системами. За основу было взято определение, предложенное в пионерной работе аналитика Meta Group Дага Лэйни [55] для характеристики больших данных через объём, скорость изменений и разнообразие (3V = Volume, Velocity, Variety). Позднее делались попытки добавить еще несколько V (например, V = Volatility, V = Veracity, V = Validity) к характеристикам больших данных – в качестве курьеза можно отметить, что в отдельных статьях насчитывалось более сорока подобных V. Однако в международный стандарт в результате консенсуса были включены только три основные характеристики, где разнообразие могло меняться на вариативность (V = Variability), именно это определение приведено во Введении. В нем и далее определения терминов на русском языке взяты из идентичного национального стандарта «Информационные технологии. Большие данные. Обзор и словарь» [2], вступившего в действие с 1 ноября 2021 года.

Отметим, что наряду с основными характеристиками:

- объем данных (data volume) – количественная характеристика данных (3.1.5), влияющая на выбор ресурсов для вычислений и хранения, а также на управление данными в процессе обработки;
- скорость обработки данных (data velocity) – скорость потока, с которой данные создаются, передаются, сохраняются, анализируются или визуализируются;
- разнообразие данных (data variety) – диапазон форматов, логических моделей, временных шкал и семантики массива данных;
- вариативность данных (data variability) – изменения в скорости передачи, формате или структуре, семантике или качестве массива данных;

в стандарте определены и другие встречающиеся характеристики, такие как:

- достоверность данных (data veracity) – полнота и/или точность данных;
- изменчивость данных (data volatility) – характеристика данных, относящаяся к скорости их изменения с течением времени.

Наряду с характеристиками больших данных в стандарте ISO/IEC 20546:2019 даны определения основных понятий, относящихся к технологиям работы с большими данными, таких как распределенная обработка данных, аналитика данных, облачные вычисления, горизонтальное и вертикальное масштабирование и т.д.

Заканчивается международный стандарт кратким описанием ключевых процессов обработки данных.

С терминологическим стандартом ISO/IEC 20546 тесно связан международный технический отчёт ISO/IEC TR 20547–5: 2018 «Information technology» — Big data reference architecture — Part 5: Standards roadmap» [13] который, прежде всего, содержит сводку основных стандартов, связанных с эталонной архитектурой больших данных.

В отчете также приводится достаточно представительный (хотя и не исчерпывающий) перечень официальных организаций стандартизации, отраслевых консорциумов и организаций-разработчиков программного обеспечения, которые заинтересованы в стандартизации больших данных на международном и национальном уровнях.

Последние разделы отчета посвящены потенциальным пробелам в стандартизации больших данных в таких областях как сценарии использования больших данных; спецификации и стандартизация метаданных, включая происхождение данных; методы обработки (например, пакетные, потоковые); семантика соответствия больших данных предметным областям; расширенные сетевые протоколы для эффективной передачи данных; общие и предметные онтологии и таксономии для описания семантики данных, включая взаимосвязь между онтологиями;

Стандарты обычно создаются с учетом лучших практик и подходов, которые проверены на реальных приложениях и в теории. В случае больших данных многие стандарты также развиваются на основе существующих стандартов, которые модифицируются для учета уникальных особенностей больших данных. Это направление деятельности обозначено в техническом отчете как подход, который принят и реализуется на международном уровне.

4 Большие данные: структура и процесс применения

Как уже отмечалось выше, еще одним барьером к развитию и использованию технологий работы с большими данными было отсутствие референсной архитектуры информационных систем для оперирования большими данными. В этом направлении в ИСО/МЭК с 2018 года началась активная разработка единой серии стандартов ISO/IEC 20547-X [12–16], посвящённых эталонной архитектуре больших данных, которая была успешно завершена в 2020 г.

Стандарты серии ISO/IEC 20547-X предназначены в первую очередь для обеспечения однозначного описания архитектуры системы работы с большими данными. Состав и взаимосвязи между отдельными частями серии стандартов представлен в ISO/IEC 20547–1 [14] (см. рисунок 5).

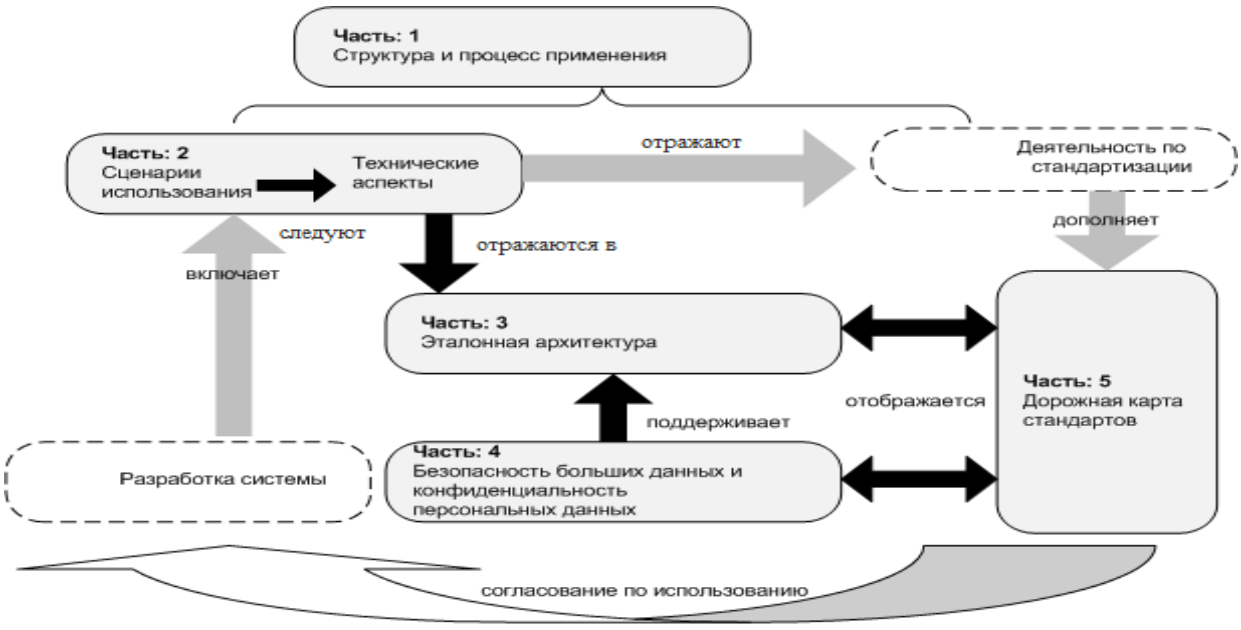

Рисунок 5 — Взаимосвязи между частями стандарта ИСО/МЭК 20547-X
Источник: [14]

В ***части 1*** «Структура и процесс применения» представлена логическая взаимосвязь между всеми частями стандарта ISO/IEC 20547-X, дано описание (концептуальной схемы) эталонной архитектуры больших данных и процесса применения стандарта в конкретной предметной области. В ***части 2*** «Сценарии использования и производные требования» представлены примеры описания сценариев (вариантов) использования больших данных в различных предметных областях, а таже выделены вытекающие из них технические требования к системам для работы с большими данными. В ***части 3*** «Эталонная архитектура» содержится описание эталонной архитектуры больших данных, включающей в себя основные архитектурные понятия, в том числе пользовательское и функциональное представление. В ***части 4*** «Безопасность больших данных и конфиденциальность персональных данных» представлено описание аспектов доверия и безопасности при работе с большими данными, а также обеспечения конфиденциальности персональных данных применительно к эталонной архитектуре больших данных. Обоснованы причины снижения уровня безопасности и конфиденциальности при работе с большими данными, а также представлены подходы, направленные на повышение уровня информационной безопасности систем больших данных. С учетом особенностей работы с большими данными и их высокой ценности сформулированы задачи обеспечения безопасности и конфиденциальность персональных данных, а также представлены рекомендации по обеспечению безопасности и защите персональных данных при работе с большими данными как на уровне организации, так и на уровне экосистемы больших данных. В ***части 5*** «Дорожная карта стандартов» содержится перечень стандартов (как существующих, так и разрабатываемых), организаций, активно участвующих в процессах стандартизации, а также представлен анализ существующих проблем и направлений разработки будущих стандартов, относящихся к большим данным.

Общее описание системы для работы с большими данными, содержащееся в ISO/IEC 20547–1 позволяет сформулировать ключевые требования к их функционированию:

- обеспечение эффективной обработки, хранения, управления и анализа больших массивов данных, характеризующихся объёмом, разнообразием, скоростью обработки, а также их вариативностью; с этой целью должны использоваться различные технологии масштабирования;
- реализация перспективных методик построения масштабируемых систем данных на основе независимых ресурсов в ситуациях, когда характеристики массивов данных требуют разработки специальных архитектур для эффективного хранения, обработки и анализа;
- реализация парадигмы распределения массивов данных по горизонтально связанным и независимым ресурсам с целью достижения масштабируемости, необходимой для эффективной обработки больших массивов данных.

Разнообразие систем больших данных и технологий работы с ними определяют набор необходимых и достаточных требований, предъявляемых к эталонной архитектуре, позволяющих реализовывать широкий спектр потенциальных сценариев использования больших данных.

Вторая половина стандарта ISO/IEC 20547–1 посвящена описанию пошагового процесса применения эталонной архитектуры для разработки архитектуры конкретной системы больших данных. Эталонная архитектура больших данных, описанная в стандарте ISO/IEC 20547–3 (см. далее раздел 6), является достаточно общей и предназначена для различных сценариев использования, однако для учета потенциального разнообразия систем больших данных (и их компонентов) процесс применения предоставляет возможность расширения эталонной архитектуры. Основной особенностью этого расширения является идентификация дополнительных действий, связанных с ролями, и/или назначение действий различным ролям/подролям.

Применение эталонной архитектуры к построению конкретной системы больших данных сводится к выполнению ряда взаимосвязанных шагов.

1. Идентификация заинтересованных сторон и их требований.
2. Отображение в ролях и подролях заинтересованных сторон и их требований.
3. Разработка подробных описаний деятельности и её соответствие интересам (требованиям).
4. Определение функциональных компонентов для реализации деятельности.
5. Определение соответствия сквозных действий/функциональных компонентов интересам (требованиям).

Каждый из перечисленных шагов достаточно детально описывается в стандарте ISO/IEC 20547–1 с отсылкой к другим действующим стандартам.

На первом шаге в процессе применения эталонной архитектуры осуществляется выявление заинтересованных сторон и определение их интересов, связанных с разработкой системы больших данных.

На следующем шаге решается задача отображения заинтересованных сторон и их требований в общей структуре понятий и представлений о системе больших данных в ролях и подролях, под которыми понимаются виды деятельностей с большими данными.

На третьем шаге выполняется подробное описание деятельности для конкретной системы больших данных, представленной в категориях ролей и подролей с учетом требований заинтересованных сторон.

На следующем шаге, представляющем этап высокоуровневого проектирования системы больших данных, выполняется идентификация функциональных компонентов, предназначенных для осуществления деятельности системы больших данных.

На последнем шаге процесса разработки выполняется валидация соответствия сформированных функциональных компонентов системы больших данных заданным требованиям путём трассировки каждого требования до функционального компонента и наоборот, трассировки каждого функционального компонента до конкретного требования в рамках реализуемой деятельности.

5 Сценарии использования больших данных

Варианты использования технологий работы с большими данными, а также вытекающие из них требования к эталонной архитектуре создаваемых систем, зафиксированы в международном техническом отчёте ISO/IEC 20547–2 [12], который содержит описания 51 варианта использования больших данных, классифицированных по следующим 9 категориям:

- деятельность государственных органов;
- коммерческая деятельность;
- оборона;
- здравоохранение и медико-биологические науки;
- глубокое обучение (Deep Learning) и социальные сети;
- экосистема для исследований;
- астрономия и физика;
- науки о Земле, экологические науки и полярные исследования;
- энергетика.

Появление любой инновационной технологии влечёт за собой желание собирать и распространять интересные варианты её использования. Этим занимаются многочисленные профессиональные ассоциации, научные учреждения и консультационные фирмы, а в последнее время – международные организации по стандартизации. Усилия по сбору, описанию и анализу вариантов использования можно разделить на две части: 1) «коллекционирование» кейсов и 2) попытки проанализировать собранные варианты и обнаружить какие-либо высокоуровневые закономерности. При этом социально-экономические эффекты от «коллекционирования» кейсов проявляются лишь в первые несколько лет с момента появления технологии в пилотных проектах, но в дальнейшем сходят на нет. Высокоуровневый анализ кейсов редок ввиду нехватки сильных аналитиков.

К сожалению, все проблемы, связанные с «коллекционированием» вариантов использования, можно наблюдать и в техническом отчёте ISO/IEC 20547–2. В этом документе недостаёт высокоуровневой аналитики, доведённой до чётко изложенных результатов, пригодных для использования лицами, принимающими решения. В этом есть потенциал для разработки в будущем небольшого по объёму отечественного стандарта, закрывающего данный пробел.

Вместе с тем адаптация технического отчёта ISO/IEC 20547–2 принесёт пользу отечественным специалистам и организациям, ввиду следующего:

- российские специалисты смогут на родном языке познакомиться с материалами, отражающими внедрение и развитие технологий работы с большими данными за рубежом;

- описания многих вариантов использования представляют самостоятельный интерес, особенно для специалистов из «родственных» сфер деятельности – здесь можно почерпнуть полезные идеи для собственных проектов;
- в документе сделана попытка выделить типичные проблемы, пути их решения и варианты дальнейшего развития технологий работы с большими данными, которые могут стать полезной отправной точкой для последующей аналитики накопленного опыта;
- в документе нашли отражение вопросы обеспечения сохранности, доступности для конечных пользователей и эффективного повторного использования разнородных данных, накапливаемых в рамках научных исследований и экспериментов – то есть, по сути, речь часто идёт о стратегическом управлении информацией и данными с целью оптимизации расходов и увеличения полезной отдачи;
- документ отражает хороший мировой опыт подготовки и обработки описаний вариантов использования – так, ИСО на основе данного документа подготовила технический отчёт о вариантах использования технологий распределённых реестров (блокчейна), известно об интересе к этому опыту и других международных технических комитетов.

Если говорить о связанных с большими данными проблемах и рисках, то из описаний вариантов использования можно увидеть, в том числе, следующие.

«Хрупкость» результатов – даже небольшие изменения в составе данных и в алгоритмах обработки способны привести кардинальным отличиям в результатах. Это особенно критично при использовании больших данных для принятия юридически значимых решений, поскольку в случае спора часто невозможно эти результаты воспроизвести.

Риски для информационной безопасности и персональных данных. В последнее время ужесточаются требования к информационной безопасности и особенно к защите персональных данных. При этом именно технологии работы с большими данными и искусственного интеллекта создают новые риски, заставляя ограничивать доступ к разнородным массивам данных и их объединение или совместную обработку, а также проводить анонимизацию. Всегда следует помнить, что в сфере больших данных значительное конкурентное преимущество (в том числе при обеспечении национальной безопасности) получает сторона, обладающая наиболее мощной инфраструктурой для обработки данных, ресурсами, кадрами и т.д.

Проблемы документирования и долговременной сохранности. Для решений на основе больших данных существует проблема объяснимости принятых решений и их полноценного документирования, а также обеспечения целости, аутентичности, надёжности, конфиденциальности и пригодности к использованию ценных данных в долговременной перспективе. Многие проблемы связаны с быстрым устареванием программной, аппаратной и сетевой инфраструктуры для работы с данными, а также с необходимостью постоянно инвестировать существенные ресурсы в её поддержание и развитие.

Правовые риски. Постоянно появляются противоречия между требованиями к эффективности архитектуры больших данных и требованиями законодательства соответствующих юрисдикций. Глобальное распределённое хранение больших данных может, например, противоречить требованиям законодательства к локализации данных, особенно персональных. Проблематичным может оказаться исполнение повышенных требований к импортозамещению.

Проблемы интерпретации. Данные могут быть правильно интерпретированы только в правильном контексте. При объединении разнородных наборов данных нередко имеет место частичная или даже полная потеря ключевого контекста (например, сведений о том, кто, когда, каким образом и для чего собрал данные, какова их точность, насколько они актуальны и т.д.). В результате даже сами по себе вполне корректные данные могут быть неверно интерпретированы (в том числе умышленно), что может приводить к принятию ошибочных деловых и политических решений.

Проблемы контролируемого сбора данных. Общий принцип обработки данных заключается в том, что лишние данные мешают эффективной работе алгоритмов и «поедают» ресурсы. Во многих проектах большое внимание уделяется оперативной предобработке поступающих данных, результаты которой используются для оперативного управления процессом сбора данных.

В техническом отчёте ISO/IEC 20547–2 также имеется раздел с перечислением технических проблем, выявленных по результатам анализа вариантов использования. Таким образом, выборка

и анализ собранных вариантов использования позволили – через производные требования из них – построить эталонную архитектуру больших данных (см ниже раздел 6). Обобщённые требования и анализ вариантов использования также позволили разработать проект национального стандарта о требованиях заказчика к действиям, связанным с использованием больших данных (см. ниже раздел 8), в котором отражено, что соответствующие требования укладываются в производные требования, зафиксированные стандартом ISO/IEC 20547–2, и гарантируют качественный результат.

6 Большие данные: эталонная архитектура

Эталонные архитектуры разрабатываются для решения широкого круга задач, и их основное предназначение состоит в ориентации на будущее и использование в качестве основы для будущих реализаций информационных систем [56]. Архитектура систем для работы с большими данными позволяет принимать, обрабатывать и анализировать данные, которые являются слишком объёмными или слишком сложными для традиционных информационных систем, таких как реляционные базы данных. Эталонная архитектура предоставляет заинтересованным сторонам универсальный язык для описания больших данных, обеспечивает поддержку общих стандартов, спецификаций и шаблонов, а таже демонстрирует последовательность действий при реализации технологий для решения однотипных задач.

Результаты применения эталонной архитектуры позволяют снять или существенно уменьшить возникающие проблемы при эксплуатации систем для работы с большими данными и дают инструмент для описания, обсуждения и развития специализированных архитектур. В данном разделе рассматривается эталонная архитектура больших данных, описанная в третьей части международного стандарта ISO/IEC 20547–3 [15].

Архитектура больших данных обычно используется для реализации следующих сценариев [57]:

- хранение и обработка данных, в том числе неструктурированных, в объёмах, слишком больших для традиционной базы данных;
- преобразование неструктурированных или слабо структурированных данных для анализа и создания отчетов;
- запись, обработка и анализ непривязанных потоков данных в режиме реального времени или с низкой задержкой;
- прогнозная аналитика и машинное обучение.

Обычно разработчики решений для работы с большими данными стремятся объединить возможности обработки разнородных данных. Одним из распространённых подходов является использование платформы Hadoop, которая работает по принципу перемещения вычислений ближе к месту хранения данных: обработка обычно выполняется на больших кластерах серверов, созданных с помощью стандартного аппаратного обеспечения [58]. Сочетание платформы Hadoop со стандартными серверами – основа для экономичной и высокопроизводительной аналитической платформы для параллельной работы приложений. Среди сформировавшихся подходов к работе с большим данными можно отметить модель распределённых параллельных вычислений MapReduce, которая находит применение в компьютерных кластерах [59]. Согласно этой модели, приложение разделяется на большое количество одинаковых элементарных заданий, выполняемых на узлах кластера и затем естественным образом формирующих конечный результат. Для формирования сложных и гибко построенных запросов к большим данным используются языки NoSQL (от англ. Not Only SQL, не только SQL). Обычные реляционные базы данных обеспечивают формирование для достаточно быстрых и однотипных запросов. В случае больших данных нагрузка может существенно превышать разумные пределы и использование СУБД становится неэффективным.

Разнообразная природа вариантов использования больших данных (см. выше раздел 5) определяет необходимость того, что эталонная архитектура больших данных должна быть достаточно общей и охватывать многообразие потенциальных архитектур систем работы с большими данными. С объектно-ориентированной точки зрения она может быть представлена как абстрактный класс, определяющий структуру и атрибуты конкретных вариантов архитектур. Таким образом, эталонная архитектура должна включать структуру и взаимосвязь компонентов, правила и ограничения, общие для всех систем больших данных, а также ряд соглашений,

принципов и практик для описания архитектур систем больших данных, в том числе отражать взаимосвязи с окружающей средой, областью применения, заинтересованными сторонами и др.

На рисунке 6 представлена концептуальная схема взаимосвязей между базовыми понятиями эталонной архитектуры больших данных. Эталонная архитектура формируется для конкретной области применения, связанной с большими данными и определяющей окружающую среду. В случае больших данных окружающая среда описывается и определяется ключевыми характеристиками больших данных: объёмом, скоростью обработки, разнообразием, а также их вариативностью. Окружающая среда включает заинтересованные стороны и их интересы. Под заинтересованными сторонами подразумеваются пользователи, владельцы, архитекторы и другие субъекты любой системы, у которых имеется интерес, связанный с большими данными.

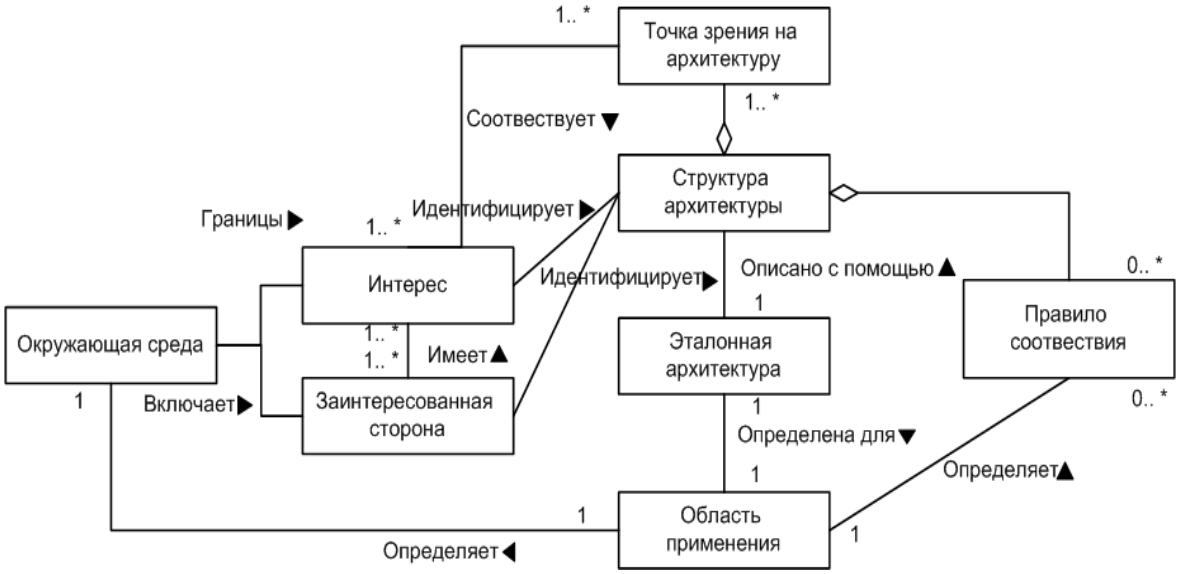

Рисунок 6 — Концептуальная схема взаимосвязей между базовыми понятиями эталонной архитектуры Источник: [14]

Эталонной архитектуре соответствует концептуальная модель экосистемы больших данных, определяющая роли/подроли и их отношения в экосистеме, а также описание типов деятельностей ролей и подролей в экосистеме больших данных.

В основе архитектуры систем больших данных лежат логические отношения между такими сущностями, как: стороны, сквозные аспекты, роли/подроли, деятельности и функциональные компоненты, которые составляют архитектуру системы больших данных (рисунок 7).

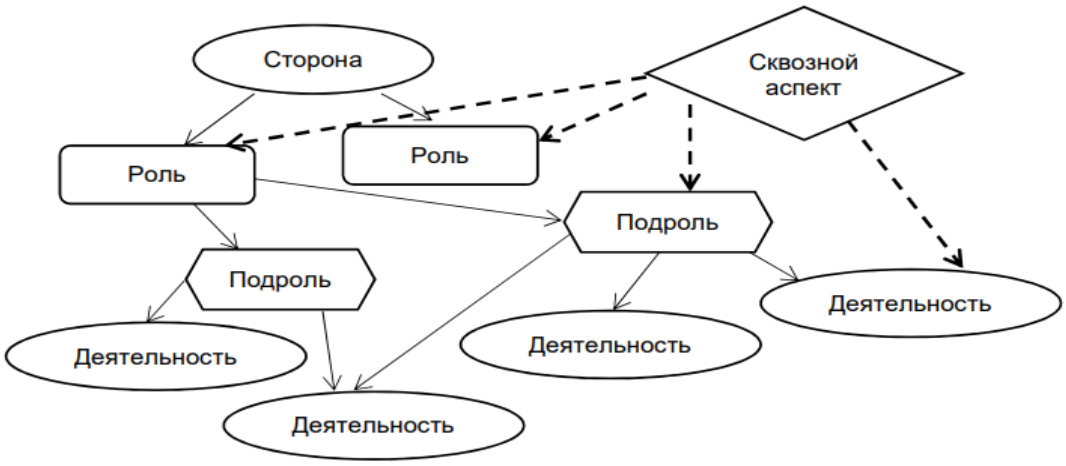

Рисунок 7 — Логические отношения между сущностями архитектуры больших данных в представлении пользователя. Источник: [15]

Сторона представляет собой физическое или юридическое лицо, или группу лиц, которые в экосистеме больших данных представляются её заинтересованными сторонами.

Под ролью понимается набор деятельностей с большими данными. Подмножество деятельностей с большими данными для конкретной роли носят название подроли, при этом деятельности данной роли могут совместно использовать различные подроли.

Деятельность представляет собой определенное исполнение одной задачи или набора задач. Деятельности с большими данными должны иметь цели и обеспечивать получение одного или нескольких результатов, которые достигаются с использованием функциональных компонентов.

В экосистеме больших данных реализуются так называемые сквозные аспекты, которые обеспечивают возможность координации между ролями и функциональными компонентами. Они влияют на выполнение нескольких ролей, действия с большими данными и функциональные компоненты, а также учитываются при совместном выполнении конкретных ролей или использовании функциональных компонентов.

Концепция эталонной архитектуры больших данных, представленная в стандарте ISO/IEC 20547–3, является основой для общего описания систем работы с большими данными и рассматривается с двух точек зрения:

- ***пользовательского представления***, включающего роли, подроли, деятельности и сквозные аспекты функционирования системы работы с большими данными, обеспечивающие удовлетворение потребностей заинтересованных сторон;
- ***функционального представления***, включающего функциональные уровни, компоненты и многоуровневые функции, обеспечивающие реализацию действий и сквозных аспектов, указанных в представлении пользователя.

Каждая из точек зрения, в свою очередь, затрагивает один или несколько интересов (проблем, требований). В рамках указанных архитектурных представлений интересы могут быть воплощены в одной или нескольких ролях, деятельностях и функциональных компонентах. Схема взаимодействия заинтересованных сторон, а также интересы (проблемы, требования), связанные с указанными представлениями, представлены на рисунке 8.

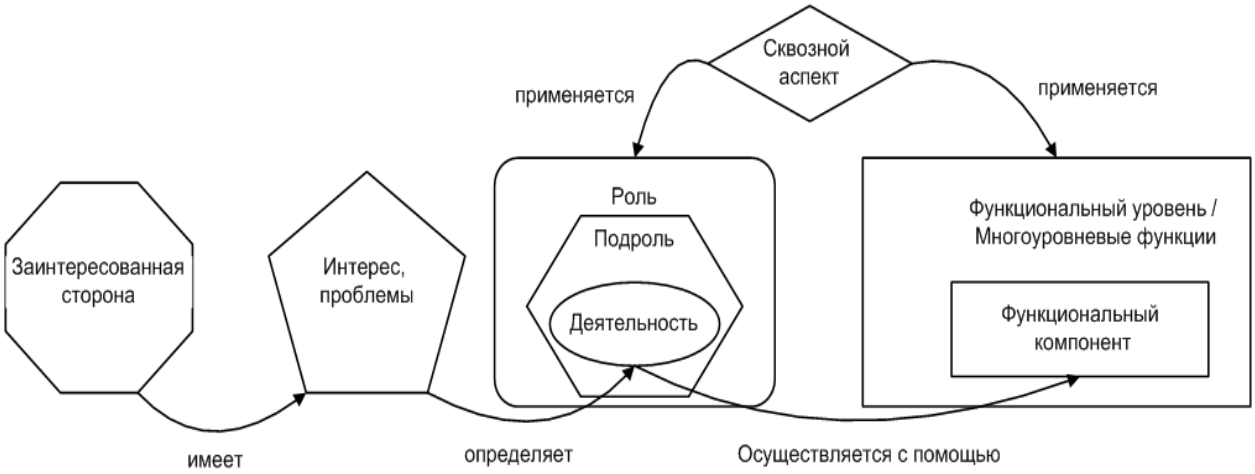

Рисунок 8 – Схема взаимодействия между компонентами представлений эталонной архитектуры больших данных
Источник: [15]

Сквозные аспекты, координирующие совместное выполнение отдельных ролей или использование функциональных компонентов, существуют как в пользовательском, так и функциональном представлениях эталонной архитектуры больших данных. Примером сквозного аспекта может служить обеспечение безопасности и конфиденциальности персональных данных. В отличие от пользовательского функциональное представление является технологически нейтральным представлением о функциях, необходимых для формирования системы работы с большими данными.

Функциональное представление описывает распределение функций, обеспечивающих поддержание деятельности с большими данными. Зависимости между функциями определяет функциональная архитектура. Функциональное представление включает следующие понятия в архитектуре больших данных (рисунок 9):

- ***функциональные компоненты***, которые являются функциональными строительными блоками, необходимыми для участия в деятельности;
- ***интерфейсы функциональных компонентов***, выполняющие функции границы между двумя функциональными компонентами для поддержки коммуникации и обмена данными;
- ***функциональные уровни,*** включающие набор функциональных компонентов со схожими возможностями или служащими общей цели;
- ***многоуровневые функции,*** представляющие декомпозицию сложной функции на совокупность более простых; многоуровневые функции включают в себя функциональные компоненты с возможностями, используемыми на нескольких функциональных уровнях, и сгруппированы в подмножества.

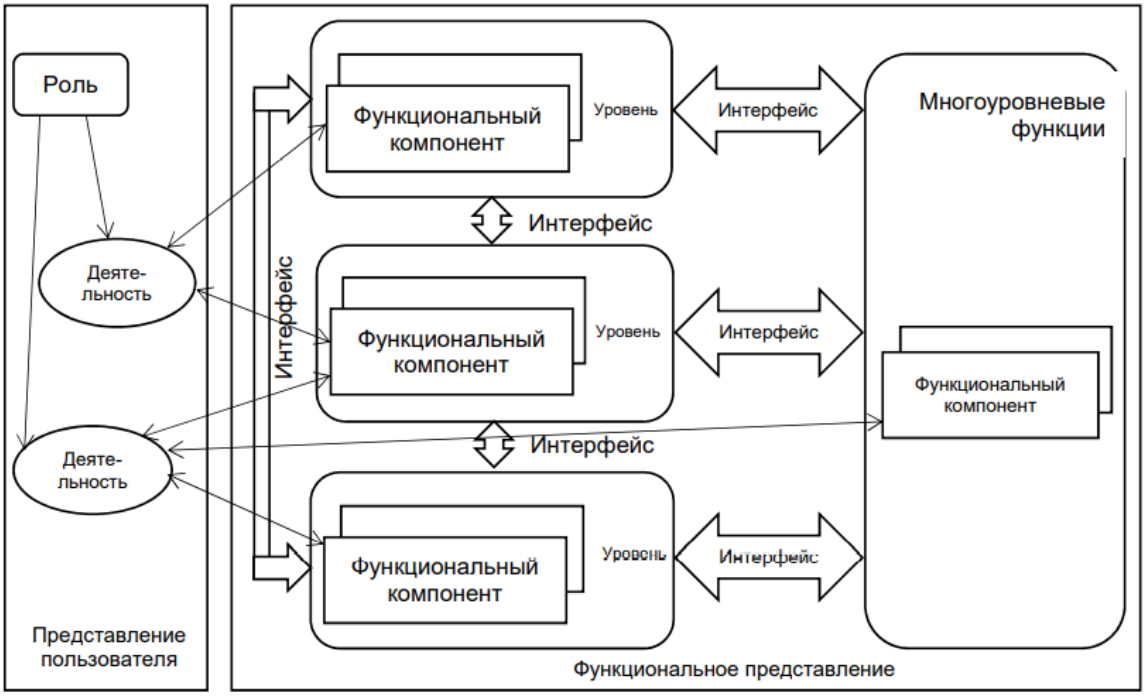

Рисунок 9 — Схема взаимодействия сущностей пользовательского и функционального представлений
Источник: [15]

Что касается сквозных аспектов, то они включают как архитектурные, так и эксплуатационные аспекты и применяются к компонентам и функциям эталонной архитектуры больших данных.

Сквозные аспекты влияют на деятельности с большими данными, выполняемые ролями. Роли, в свою очередь, могут координировать между собой функции поддержки сквозного аспекта. Каждому сквозному аспекту соответствуют функциональные компоненты для поддержки деятельностей и набор деятельностей с большими данными.

Сквозные аспекты являются общими для ролей, деятельностей и функциональных компонентов. Например, безопасность является сквозным аспектом, поскольку она применяется к сервис-провайдерам (доступа к большим данным, приложений больших данных, среды обработки больших данных), к потребителям больших данных и к партнёрам сервиса больших данных.

Применение эталонной архитектуры больших данных основано на реализации пошагового процесса, приведенного в разделе 4.

Таким образом, эталонная архитектура больших данных позволяет обеспечивать решение следующих задач стандартизации:

- выполнение формального описания различных компонентов, процессов и систем больших данных в контексте общей концептуальной модели больших данных;
- выполнение функциональной классификации, которая позволит заинтересованным государственным ведомствам, агентствам и другим потребителям классифицировать решения для больших данных и проводить их сравнительный анализ;
- проведение анализа стандартов по их функциональной совместимости.

7 Безопасность больших данных и конфиденциальность персональных данных

При применении эталонной архитектуры больших данных важное место занимает проблема обеспечения информационной безопасности, которая затрагивает все роли и подроли в её экосистеме и функциональных компонентах. Задачи обеспечения информационной безопасности отражены в четвертой части международного стандарта ISO/IEC 20547–4 [16], в котором представлено описание аспектов безопасности больших данных и их конфиденциальности в рамках эталонной архитектуры больших данных, включая роли, деятельности и функциональные компоненты.

Требуемый уровень информационной безопасности обеспечивается ***оркестратором системы больших данных*** при формировании политики информационной безопасности, обосновании требований к системе больших данных и проведении аудита информационной безопасности, а также ***сервис-провайдерами*** приложений больших данных и среды обработки больших данных при разработке системы, её развёртывании (вводе в эксплуатацию) и в процессе эксплуатации.

В стандарте ISO/IEC 20547–4 представлены результаты анализа проблем информационной безопасности, вытекающих из ключевых свойств больших данных и ключевых свойств процесса обработки данных, которые создают дополнительные риски и являются причинами снижения уровня безопасности и конфиденциальности при работе с большими данными, сформулированы рекомендации по обеспечению безопасности и конфиденциальности при выполнении различных операций с большими данными.

Парадигма больших данных привела к стиранию границы безопасности между системами сбора, хранения и доступа к данным — областями, которые традиционно рассматривались как независимые. В связи с этим первостепенное значение для создания атмосферы взаимного доверия и сотрудничества между заинтересованными сторонами в сфере сбора, хранения и обработки больших данных имеет стандартизация требований к безопасности и конфиденциальности персональных данных.

Ключевые свойства больших данных (объём, скорость обработки, разнообразие и изменчивость), а также ключевые свойства процесса обработки данных (волатильность, достоверность и ценность) определяют дополнительные риски и, следовательно, являются причинами снижения уровня безопасности и конфиденциальности при работе с большими данными:

- скорость обработки данных порождает риск, связанный с существенным повышением скорости потока, в рамках которого данные создаются, хранятся, анализируются или визуализируются; в этом случае средства управления безопасностью могут приводить к снижению скорости передачи данных, поэтому от них нередко отказываются;
- разнообразие данных значительно увеличивает их сложность, поскольку множество различных источников данных находится под контролем различных субъектов; повышение сложности данных неизбежно приводит к появлению новых уязвимостей; новые возможности, вызванные разнообразием больших данных, позволяют получить персональные данные из обезличенных наборов путем их сопоставления с доступными публичными базами данных;
- изменчивость данных влечет за собой риски, связанные с более быстрыми изменениями скорости передачи, формата/структуры, семантики и (или) качества данных; всё это приводит к необходимости совершенствования средств управления безопасностью в целях защиты персональных данных;

- волатильность данных негативно влияет на эффективность ведения журналов безопасности и усложняет управление безопасностью;
- необходимость обеспечения достоверности данных предъявляет более высокие требования к таким показателям, как целостность, согласованность и точность; возникающие при этом взаимосвязанные риски могут агрегироваться и значительно возрастать;
- ценность данных порождает большее число атак, преследующих различные цели и интересы.

Активное распространение приложений для работы большими данными вызывает всё более серьезные проблемы с точки зрения безопасности и конфиденциальности персональных данных, включая случаи их потери и утечки, а также скрытую неконтролируемую передачу данных. Это порождает злоупотребление данными, а также ставит под угрозу социальную стабильность и национальную безопасность.

В стандарте ISO/IEC 20547–4 значительное место отведено рассмотрению вопросов обеспечения безопасности и конфиденциальности персональных данных при использовании технологических платформ, в которых применяются разнообразные технологии обработки больших данных, новая техническая архитектура и вспомогательные платформы, а также специализированное программное обеспечение.

Выявлены недостатки существующих способов обеспечения информационной безопасности при применении в системах работы с большими данными и сформулированы следующие задачи совершенствования подходов к обеспечению их защиты.

А) Совершенствование традиционных средств управления безопасностью при работе с большими данными. Приложения для больших данных обычно используют открытую распределённую архитектуру вычислений и хранения со сложной базовой поддержкой для организации распределённых хранилищ больших данных и высокопроизводительных вычислительных сервисов. Благодаря этим новым технологиям и архитектурам границы приложений для больших данных размываются, поэтому традиционные методы защиты на основе границ перестают работать. Кроме того, целенаправленные устойчивые угрозы, распределённые атаки «отказ в обслуживании», интеллектуальный анализ данных на основе машинного обучения и способы обнаружения персональных данных, а также другие типы атак выявляют серьёзные недостатки традиционных инструментов защиты и обнаружения, а также других мер обеспечения безопасности. Необходимы новые технологические подходы для обеспечения конфиденциальности персональных данных, а также реализации методов машинного обучения, криптографических механизмов безопасности, ориентированных на данные, и средства контроля доступа.

Б) Обеспечение безопасности и конфиденциальности персональных данных при использовании инфраструктуры распределённых вычислений и хранения больших данных. С этой целью требуется разработка и внедрение инструментов безопасных распределённых вычислений и безопасных систем хранения данных. Для обработки больших данных необходимы масштабируемые и распределённые решения для безопасного хранения данных, проведения аудита и выявления источников происхождения данных. Аналитика в реальном времени для анализа угроз требует обработки больших объёмов данных, связанных с безопасностью.

В) Совершенствование механизмов обеспечения безопасности технологической платформы в контексте безопасности и конфиденциальности персональных данных при работе с большими данными. Многие существующие приложения для работы с большими данными используют платформы и технологии управления большими данными на базе платформы Hadoop и технологии обработки и программной модели для распределенных вычислений MapReduce.

На начальном этапе проектирования эти платформы и технологии в основном рассматриваются как решения для доверенной внутренней сети; вместе с тем функции аутентификации, авторизации, ключевые сервисы, а также возможности проведения аудита безопасности практически не учитываются. При этом общая эффективность средств обеспечения безопасности является недостаточной. Между тем в приложениях для больших данных часто используются сторонние компоненты с открытым исходным кодом. Из-за отсутствия надлежащего управления тестированием и сертификацией безопасности этих компонентов во многих случаях не

удается предотвратить появление уязвимостей и вредоносного программного обеспечения в приложениях для больших данных.

Г) Создание средств управления доступом к приложениям для работы с большими данными. Из-за большого разнообразия типов данных и широкого спектра приложений для работы с большими данными они часто используются для предоставления множества услуг пользователям с разными учётными данными и целями из разных организаций или разных подразделений одной организации. В общем случае контроль доступа является эффективным средством обеспечения контролируемого использования данных, однако из-за большого числа неизвестных пользователей и данных, к которым необходимо получить доступ, становится сложным предварительно определить роли и разрешения на доступ к данным. Права пользователя на доступ к данным могут быть классифицированы заранее, тем не менее многочисленность ролей снижает достоверность контроля фактических разрешений для каждой роли. Поэтому становится затруднительным для каждого пользователя точно определить диапазон данных, к которому ему необходимо получать доступ, без развёртывания более современной модели управления доступом.

Д) Создание масштабируемых механизмов обеспечения безопасности и конфиденциальности персональных данных. При разработке и применении механизмов обеспечения безопасности больших данных и конфиденциальности персональных данных, таких как управление ключами, управление идентификацией и доступом, обезличивание и т. д., в среде больших данных необходимо учитывать не только функции безопасности и конфиденциальности, но и масштабируемость этих механизмов, чтобы гарантировать обработку большого объёма данных, поступающих с высокой скоростью.

С точки зрения приложений для работы с данными с учётом их характеристик (объёма, разнообразия, вариативности, скорости обработки, достоверности, изменчивости), а также огромной ценности больших данных требуется решение следующих задач для обеспечения безопасности и конфиденциальность персональных данных.

- Создание средств защиты для обработки персональных данных в рамках системы больших данных, которые должны учитывать особенности распределенных систем, открытой сетевой среды, сложных приложений для обработки данных.
- Создание средств обеспечения конфиденциальности персональных данных при работе с большими данными с учетом таких инцидентов безопасности, как злоупотребление данными, внутренняя кража и сетевые атаки, ведущие к утечке персональных данных, которые будут иметь более серьёзные последствия, чем в обычных информационных системах.
- Создание средств контроля аутентичности данных и источников больших данных с учетом их широкого спектра, включая различные датчики, активные загрузки и общедоступные веб-сайты, а также большое количество ненадёжных источников, через которые злоумышленники имеют возможность фальсифицировать данные. При проверке аутентичности данных возникает множество проблем из-за ограниченной производительности терминалов сбора данных, отсутствия необходимых технологий, лимитированного объёма информации, а также разнообразия и сложности источников.
- Создание средств защиты прав владельцев данных при работе с большими данными с учетом того, что во время работы с большими данными к ним могут обращаться различные пользователи, их могут передавать от одного специалиста к другому, а также анализировать для получения новых данных. То есть в процессе обмена данными и общего доступа к ним возникает обстоятельство, при котором право владения данными для владельца данных и право использования данных разделяются. Иными словами, данные могут находиться вне контроля владельца, что влечёт за собой риски злоупотребления данными, неопределённого владения ими и неясных требований по надзору за их безопасностью, что может нанести серьёзный ущерб правам и интересам владельцев данных.

Поскольку экосистемы больших данных и сети организаций обмениваются данными с целью сбора, обработки, хранения и анализа, должны быть предусмотрены следующие варианты взаимодействия между заинтересованными сторонами:

- для согласования общих требований к безопасности и защите персональных данных в экосистеме и соответствующих требований в отдельной организации;

- для согласования вопросов общего управления рисками на уровне экосистемы и уровне отдельной организации;
- для обеспечения того, чтобы отдельные организации гарантировали согласованную работу с защищаемыми активами.

В стандарте представлены аспекты безопасности и защиты персональных данных с точки зрения пользователя и с точки зрения мероприятий по верхнеуровневому управлению организацией в пространствах проблем и решений. При этом пространство проблем относится к миру проблем и мотиваций конечного пользователя, а пространство решений – к миру продуктов, услуг и технологий.

В стандарте ISO/IEC 20547–4 также представлены рекомендации по обеспечению безопасности и защите персональных данных при работе с большими данными как на ***уровне организации, так*** и на ***уровне экосистемы больших данных***.

Взаимосвязи между организациями, экосистемой больших данных, а также их эталонной архитектурой при решении задач обеспечения безопасности и конфиденциальности персональных данных показаны на рисунке 10.

Рисунок 10 — Структурная схема взаимосвязи между организациями, экосистемой больших данных и эталонной архитектурой больших данных
Источник: [16]

Организация может быть частью экосистемы, которая придерживается принципов эталонной архитектуры больших данных. Эталонная архитектура больших данных включает роли и подроли, деятельности и функциональные компоненты.

Организация осуществляет операции, связанные с ролями, деятельностями и функциональными компонентами. Например, организация играет роль сервис-провайдера приложений для больших данных. Организация осуществляет операции, ориентированные на безопасность и конфиденциальность персональных данных. Они применяются для защиты активов от уязвимостей. Некоторые активы могут быть специфическими для организации, например сведения, содержащие коммерческую тайну, в то время как другие могут совместно использоваться в экосистемах, например наборы данных.

Подход, представленный в стандарте ISO/IEC 20547–4, предусматривает описание проблем информационной безопасности с точки зрения двух типов операций в области обеспечения безопасности и конфиденциальности персональных данных: операции по организации деятельности по обеспечению информационной безопасности и защите персональных данных, а также операции по сотрудничеству в экосистеме.

В стандарте представлены рекомендации по обеспечению безопасности и защите персональных данных при работе с большими данными на уровне организации и на уровне экосистемы, которые соответствуют этапам: разработки требований; анализа рисков;

проектирования средств управления; разработки функционала системы для работы с большими данными.

Применительно к каждому этапу сформулированы особенности решения задач обеспечения безопасности и защиты персональных данных при работе с большими данными. В основу рекомендаций положено описание операционной деятельности, представленное в ряде стандартов, посвященных управлению информационной безопасностью, защите персональных данных и проектированию систем защиты персональных данных [ИСО/МЭК 27001 Системы управления информационной безопасность. Требования; ISO/IEC 27000 Применение стандарта ISO/IEC 27001 в конкретных секторах. требования; ИСО/МЭК 29100 Основы защиты персональных данных; ИСО/МЭК 27701 Дополнение к ISO/IEC 27001 и ISO/IEC 27002 для управления информацией о защите персональных данных; ИСО/МЭК 27005 Управление рисками, связанными с информационной безопасностью; ИСО/МЭК 27550 Проектирование систем защиты персональных данных и др.]

Стандарт ISO/IEC 20547–4 содержит описание ***функциональных компонентов*** безопасности и конфиденциальности персональных данных, под которыми понимаются функциональные элементы эталонной архитектуры, используемые для выполнения определенной процедуры или совокупности процедур безопасности в конкретной реализации архитектуры. В качестве функциональных компонентов могут использоваться категории управления безопасностью, представленные в национальном стандарте ГОСТ Р ИСО/МЭК 27002 [60].

Таким образом, содержание стандарта ISO/IEC 20547–4 включает описание различных аспектов обеспечения безопасности больших данных и конфиденциальности персональных данных применительно к эталонной архитектуре, в том числе задач совершенствования существующих подходов, а также рекомендации по обеспечению безопасности и конфиденциальности при выполнении операций с большими данными в отдельной организации, а также в экосистеме больших данных.

8 Стандартизация требований заказчика к действиям, связанным с оперированием большими данными

Система международных стандартов ISO/IEC 20546 / 20547-X формирует системно-технологическое окружение оперирования большими данными – эталонную архитектуру, виды данных и технологии их обработки, хранения, анализа, многое другое.

Однако парадигма «больших данных» является значительно более ёмкой и не ограничивается большими массивами данных и системами для работы с ними. Наряду с проблемами описания системно-технологического окружения при использовании больших данных не менее существенными являются вопросы упорядочения взаимоотношений между заказчиками и их исполнителями (подрядчиками, поставщиками) при оперировании большими данными. Для заказчика основным становится удовлетворение его потребностей, которые должны быть сформулированы в виде однозначных, полных требований к результатам действий с большими данными и быть понятны исполнителю. Как правило, такие требования фиксируются в документе, представляющем собой техническое задание, регламентирующее взаимоотношения заказчика и исполнителя, а также видение заказчиком ожидаемого результата.

Разработанный в 2020–2021 годах проект национального стандарта «Информационные технологии. Большие данные. Техническое задание. Требования к содержанию и оформлению» [40] направлен на типизацию и стандартизацию требований, включаемых в технические задания для следующих видов деятельности, связанной с большими данными:

- поставка массивов (наборов) больших данных;
- оперирование большими данными;
- разработка и/или поставка технологий и/или средств (решений) для оперирования большими данными;
- внедрение технологий и/или средств (решений) для оперирования большими данными;
- предоставление в пользование (аренда) технологий и/или средств (решений) оперирования большими данными;
- техническая эксплуатация средств (решений) для оперирования большими данными.

При разработке национального стандарта в значительной мере использованы термины и определения, установленные международными стандартами и другими национальными стандартами. В то же время возникла необходимость в уточнении или введении новых понятий для типизации субъектов процессов оперирования большими данными или для описания взаимоотношений между ними. Например, под терминами «заказчик», «исполнитель (подрядчик)» и «поставщик» в стандарте понимаются лица, которые могут заказать, дать задание или поручение, тем самым определяя свои потребности, и, соответственно, лица, которые могут эти заказы, задания или поручения выполнять для удовлетворения потребностей заказчика. Эти лица могут выступать как в роли «внешних», так и «внутренних» заказчиков, исполнителей (подрядчиков) или поставщиков.

Общие положения стандарта предусматривают:

- определение техническим заданием предмета выполняемых работ (оказываемых услуг) и/или поставок, а также основные требования к ним и их результатам;
- использование в качестве основы для технического задания существующей и прогнозируемой потребности заказчика, результатов исследований проблем, связанных с оперированием большими данными, национальных, международных и отраслевых стандартов и иных нормативных технических документов, требований, установленных нормативными правовыми актами, опыта предыдущих аналогичных выполненных работ (оказанных услуг), выполненных поставок;
- исключение возможности различных толкований как содержания технического задания в целом, так и устанавливаемых им требований в частности;
- обеспечение логической связанности содержания технического задания, достаточной для понимания целей, задач и требований к выполнению работ (оказанию услуг), осуществлению поставок и ожидаемым результатам;
- наличие положений о проверке соответствия выполнения работ (оказания услуг), осуществления поставок и их результатов установленным техническим заданием целям, задачам и требованиям.Еще одной особенностью национального стандарта является разделение видов деятельности, направленных на непосредственное использование больших данных по назначению и на обеспечение такого использования. Ко второй категории относятся поставка массивов (наборов) больших данных, разработка и/или поставка технологий и/или средств (решений) для оперирования большими данными, внедрение технологий и/или средств (решений) для оперирования большими данными, предоставление в пользование (аренда) технологий и/или средств (решений) оперирования большими данными, техническая эксплуатация средств (решений) для оперирования большими данными. Подобный подход позволяет устанавливать и детализировать специфические требования, относящиеся к каждому из перечисленных видов деятельности.

Разработанный стандарт не ограничивает возможности включения в задание произвольной комбинации различных видов деятельности, в том числе оперирования данными и деятельности, направленной на обеспечение такого использования. Предусмотрена возможность формирования технического задания для оперирования большими данными в рамках всего «жизненного цикла данных».

В отдельном разделе стандарта определен порядок формирования технических заданий по осуществлению деятельности, связанной с использованием больших данных для государственных и муниципальных нужд.

9 Качество данных

Обеспечение единообразных требований к качеству данных имеет первостепенное значение для эффективной работы с большими данными и достижения полезных результатов в современной динамике глобальной цифровой экономики. Первыми результатами работ по стандартизации требований к качеству данных стали международные и национальные стандарты качества данных [61, 62] и серия стандартов ISO 9000 [63], а также разрабатываемые в настоящее время проекты международных стандартов ISO/IEC 5259-X «Искусственный интеллект. Качество данных для аналитики и машинного обучения» [18–21].

Понятие качества данных вобрало в себя смыслы двух понятий: качество и данные. Апогей стандартизованного формирования понятия «качество» можно отнести к построению системы менеджмента качества в условиях рыночной экономики. Феномен качества наиболее рельефно выражен в серии стандартов ISO 9000. Качество – это определённое явление рыночной экономики, сочетающее в себе два процесса: рост удовлетворённости потребителей продукции при постоянном снижении её себестоимости (издержек). Содержательный интерес может представлять понимание термина «качество», как синтеза двух явлений: аналитики и романтики [64]. Первое явление предполагает возможность своей формализованной репрезентации, в том числе, с применением различных методов математики, физики, инструментов визуализации и искусственного интеллекта. Второе явление отрицает возможность его непосредственной формализации, оно больше ассоциируется с чисто субъективными особенностями человека, такими как: эмоции, мысли, чувства, интуиция, свобода воли и др.

С развитием тренда больших данных обострился вопрос качества данных. Это связано с необходимостью высококачественного поиска информации в больших массивах данных, и, что более важно, использования данных для обучения нейронных сетей, тестирования моделей искусственного интеллекта. Обеспечение качества данных становится всё более сложным процессом, поскольку в данных содержится много нерелевантных данных, шума, мусора. Неслучайно появляются результаты исследований, название которых красноречиво подтверждают актуальность темы качества данных, а именно, «тёмные данные» [65], «грязные данные» [66]]. Например, к грязным данным относят данные, содержащие ошибочную информацию. При этом полное удаление грязных данных из источника нецелесообразно или практически невозможно. В этой связи всё более актуализируется вопрос очистки данных, см., например, работу [67]. В этих работах акцентируется внимание на таких аспектах, как:

- известные данные в массиве отсутствуют, в данных есть пробелы;
- пользователь может даже не знать, что ему не хватает данных;
- неправильный выбор критериев для включения в выборку;
- самовыбор, когда люди сами решают, стоит ли данные включать в базу данных;
- неправильная оценка критических аспектов системы для обработки данных;
- возможность предположения иных данных, если бы были предприняты другие действия;
- данные могут быть неверными, неточными, дублировать друг друга и вводить в заблуждение;
- данные могут быть неинтегрируемыми и не допускать обобщённого форматирования;
- изменение данных и их значений во времени и др.

С ростом объёма данных всё более актуальным становится вопрос их семантической интерпретации. Если раньше истинность, и, соответственно, качество данных обычно оценивалась через их отображение на объекты или артефакты реальной действительности, то развитие технологий искусственного интеллекта потребовало всё больше внимания уделять неформализуемым семантическим интерпретациям моделей, так называемым, когнитивным семантикам [68]. Они отражают мыслительные, эмоциональные и трансцендентальные аспекты индивидуального и коллективного сознания, феномены сознательного и бессознательного. По-видимому, охват именно таких семантик будет лежать в основе перспективного развития методов искусственного интеллекта.

Интерес в контексте качества данных может представлять процесс потери сведений и информации при преобразовании и фильтрации исходных аналоговых данных в цифровую форму [69]. Например, это может происходить при обработке сигналов, получаемых из космоса или с ускорителей квантовых частиц.

Таким образом, современные стандарты, направленные на обеспечение высокого качества данных, можно оценить на соответствие следующим критериям:

- соответствие стандартам системы менеджмента качества;
- полнота охвата феномена «данные» в контексте понятий «информация» и «сведения»;
- учёт формализуемых (денотативных) и неформализуемых (когнитивных) семантик;
- взаимосвязь данных с инструментами их обработки и анализа, например, глубокого обучения;
- минимизация искажений при трансформации исходных аналоговых данных в цифровую форму;
- учёт нарастающего объёма «тёмных» данных;

- учёт построения систем очистки и сохранения данных.

Указанные международные и национальные стандарты формируют систему понятий относительно феномена качества данных, определяют характеристики качества данных, устанавливают к ним требования и указывают способы повышения качества информации.

Понятие «качество» в рассматриваемых стандартах задаётся согласно стандартам системы менеджмента качества, а именно, это степень соответствия совокупности присущих характеристик объекта требованиям. Однако в сочетании с термином «данные» соответствие новой коллокации остальным критериям вызывает сомнение.

Так, определение термина «данные (data)» в серии стандартов ISO 9000 механически заимствовано из иного стандарта (ISO/IEC 2382:2015 [70]). Оно формулируется как «интерпретируемое представление информации в соответствующей форме, удобной для передачи, интерпретации и обработки». Такое заимствование этого термина из другого стандарта в контексте темы качества данных далеко не в полной мере отражает аспект возможной семантической интерпретации данных несмотря на то, что термин «интерпретация» в этом коротком определении встречается дважды. То есть в определении термина «качество данных» и самом стандарте умаляется учёт формализуемых и неформализуемых особенностей семантической интерпретации данных, которая может быть как денотативной, так и когнитивной. Рассматриваемые стандарты далеко не в полной мере отвечают и другим перечисленным выше критериям соответствия.

Определение термина «качество данных» в проектах серии стандартов ISO 5259-X осуществляется через задание соответствующих характеристик на основе международного стандарта ISO/IEC 25012:2008 [62] и демонстрацией взаимосвязи между моделью качества данных, требованиями к качеству данных для аналитики и машинного обучения, а также характеристиками качества данных. Рассматривается полнота и применимость для аналитики и машинного обучения следующих пятнадцати характеристик качества данных: аккуратность, полнота, согласованность, надёжность, правильность, доступность, соответствие, конфиденциальность, эффективность, точность, прослеживаемость, понятность, доступность, переносимость и возможность восстановления.

Таким образом, анализ разрабатываемых и разработанных базовых стандартов искусственного интеллекта в части адекватности определения термина «качества данных» требованиям реальной теории и практики позволяет сделать следующие выводы относительно этого определения:

- проводится в основном механическое сложение определений двух терминов, приведённых в других стандартах, что не позволяет обеспечить должную синергию термина и его понятия;
- не учитываются различные виды семантической интерпретации данных, что затрудняет перспективное развитие систем обработки данных, в том числе с применением искусственного интеллекта;
- в уточнении и определении нуждаются учёт в стандарте нарастающего объёма ошибок данных и процедуры трансформации исходных аналоговых данных в цифровую форму.

Заключение

Обзор процесса и содержания стандартизации работы с большими данными демонстрирует его актуальность и значимость для социально-экономического развития. В Российской Федерации за последние годы резко сократилось отставание от международных процессов стандартизации работы с большими данными. Стандарты работы с большими данными в первую очередь будут востребованы органами власти и коммерческими компаниями, которые решают управленческие задачи или ведут бизнес, принимая решения на основе данных. Следует отметить, что введение в действие серии национальных стандартов, аналогичных международным стандартам серии 20547-X, имеет важное значение в реализации задач Национальной программы «Цифровая экономика Российской Федерации», в том числе в части внедрения сквозной цифровой технологии «Большие данные».

Очевидно, что одним из наиболее перспективных и направлений дальнейшей стандартизации больших данных станет разработка стандартов, фиксирующих требования к качеству массивов данных для разнообразных вариантов использования и различных технологий,

начиная от традиционной аналитики больших данных и заканчивая технологиями машинного/глубокого обучения. Данная тенденция является логическим развитием процесса стандартизации, первый этап которого завершился утверждением международных общеотраслевых стандартов эталонной архитектуры больших данных.

Вторым важным направлением станет разработка «отраслевых» стандартов эталонных архитектур систем для работы с большими данными применительно к конкретным сферам деятельности, таким как телекоммуникации, здравоохранение, образование или транспорт.

Благодарности

В работе использованы результаты проекта «Мониторинг и стандартизация развития и использования технологий хранения и анализа больших данных в цифровой экономике Российской Федерации», выполняемого в рамках реализации программы Центра компетенций Национальной технологической инициативы «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по договору МГУ имени М.В.Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 15.08.2019 № 7/1251/2019.

Работа выполнена при частичной поддержке РФФИ, грант 18-29-03086.

Литература

1. ISO/IEC 20546:2019 Information technology – Big data – Overview and vocabulary) // International Organization for Standardization. URL: https://www.iso.org/standard/68305.html (дата обращения: 01.10.2021).
2. ГОСТ Р ИСО/МЭК 20546–2021 «Информационные технологии. Большие данные. Обзор и словарь» // Российский институт стандартизации. URL: https://www.gostinfo.ru/catalog/Details/?id=6859575 (дата обращения: 01.10.2021).
3. ISO/IEC DIS 22989 Information technology – Artificial intelligence – Artificial intelligence concepts and terminology // International Organization for Standardization. URL: https://www.iso.org/standard/74296.html (дата обращения: 01.10.2021).
4. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh Ch., Byers A.H. Big data: The next frontier for innovation, competition, and productivity // McKinsey Global Institute. P. 12. URL: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation# (дата обращения: 01.10.2021).
5. Big Data interoperability Framework. V1.0 Final Version // National Institute of Standards and Technology. URL: https://bigdatawg.nist.gov/V1_output_docs.php (дата обращения: 01.10.2021).
6. Big Data interoperability Framework. V2.0 Final Version // National Institute of Standards and Technology. URL: https://bigdatawg.nist.gov/V2_output_docs.php (дата обращения: 01.10.2021).
7. Big Data interoperability Framework. V3.0 Final Version // National Institute of Standards and Technology. URL: https://bigdatawg.nist.gov/V3_output_docs.php (дата обращения: 01.10.2021).
8. Big Data: Big today, normal tomorrow. ITU T Technology Watch Report. November 2013. // International Telecommunication Union. URL: https://www.itu.int/en/ITU T/techwatch/Pages/big data standards.aspx (дата обращения: 01.10.2021).
9. Big data: Preliminary Report 2014. ISO/IEC JTC1, 2015. // International Organization for Standardization. URL: https://www.iso.org/files/live/sites/isoorg/files/developing_standards/docs/en/big_data_report-jtc1.pdf (дата обращения: 01.10.2021).
10. ITU-T Y.3600 (11/2015) Big data – Cloud computing-based requirements and capabilities. // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=12584&lang=en (дата обращения: 01.10.2021).
11. ITU-T Y Suppl. 40 (07/2016) Big data standardization roadmap // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=13022&lang=en (дата обращения: 01.10.2021).

12. ISO/IEC TR 20547-2:2018 Information technology – Big data reference architecture – Part 2: Use cases and derived requirements // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:20547:-2:ed-1:v1:en (дата обращения: 01.10.2021).
13. ISO/IEC TR 20547-5:2018 Information technology – Big data reference architecture – Part 5: Standards roadmap // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:20547:-5:ed-1:v1:en (дата обращения: 01.10.2021).
14. ISO/IEC TR 20547-1:2020 Information technology – Big data reference architecture – Part 1: Framework and application process // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:20547:-1:ed-1:v1:en (дата обращения: 01.10.2021).
15. ISO/IEC 20547-3:2020 Information technology – Big data reference architecture – Part 3: Reference architecture // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:20547:-3:ed-1:v1:en (дата обращения: 01.10.2021).
16. ISO/IEC 20547-4:2020 Information technology – Big data reference architecture – Part 4: Security and privacy // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:20547:-4:ed-1:v1:en (дата обращения: 01.10.2021).
17. Walshe R. The Road to Big Data Standardisation // The Elements of Big Data Value / E. Curry et al. (eds.). Springer, Cham, 2021. https://doi.org/10.1007/978-3-030-68176-0_14.
18. ISO/IEC WD 5259-1 Data quality for analytics and ML – Part 1: Overview, terminology, and examples // International Organization for Standardization. URL: https://www.iso.org/standard/81088.html (дата обращения: 01.10.2021).
19. ISO/IEC AWI 5259-2 Data quality for analytics and ML – Part 2: Data quality measures // International Organization for Standardization. URL: https://www.iso.org/standard/81860.html (дата обращения: 01.10.2021).
20. ISO/IEC WD 5259-3 Data quality for analytics and ML – Part 3: Data quality management requirements and guidelines // International Organization for Standardization. URL: https://www.iso.org/standard/81092.html (дата обращения: 01.10.2021).
21. ISO/IEC WD 5259-4 Data quality for analytics and ML – Part 4: Data quality process framework // International Organization for Standardization. URL: https://www.iso.org/standard/81093.html (дата обращения: 01.10.2021).
22. ITU-T Y.3603 Big data – Requirements and conceptual model of metadata for data catalogue» // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=14137&lang=en (дата обращения: 01.10.2021).
23. ITU-T Y.3604 Big data – Overview and requirements for data preservation» // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=14138&lang=en (дата обращения: 01.10.2021).
24. Resolution 166. Registration of a PWI entitled «Information technology – Artificial intelligence – Data life cycle framework». ISO/IEC JTC 1/SC 42 "Artificial intelligence". Resolutions taken during the Closing Plenary - JTC 1/SC 42 - 7 May 2021.
25. ITU-T Y.3651 Big-data-driven networking – mobile network traffic management and planning // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=13818&lang=en (дата обращения: 01.10.2021).
26. ITU-T Y.3653 Big data driven networking – functional architecture // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=14615&lang=en (дата обращения: 01.10.2021).
27. Приказ Федерального агентства по техническому регулированию и метрологии (Росстандарт) от 25.07.2019 № 1732 (ред. от 20.01.2021) «О создании технического комитета по стандартизации "Искусственный интеллект» // Росстандарт. URL: https://www.rst.gov.ru/portal/gost/home/activity/documents/orders#/order/104460 (дата обращения: 01.10.2021).
28. Приказ Федерального агентства по техническому регулированию и метрологии (Росстандарт) от 20.08.2020 № 1415 «О внесении изменений в приказ Федерального агентства по техническому регулированию и метрологии от 25 июля 2019 г. № 1732 «О создании технического комитета по стандартизации "Искусственный интеллект» // Росстандарт. URL:

https://www.gost.ru/portal/gost/home/activity/documents/orders#/order/179415 (дата обращения: 01.10.2021).
29. Распоряжение Правительства РФ от 23.03.2018 N 482-р (ред. от 28.05.2020) «Об утверждении плана мероприятий («дорожной карты») по совершенствованию законодательства и устранению административных барьеров в целях обеспечения реализации Национальной технологической инициативы по направлению «Технет» (передовые производственные технологии)» // СЗ РФ. 2018. № 15 (ч. V). Ст. 2173.
30. Российская венчурная компания. URL: https://www.rvc.ru/upload/iblock/ac9/Long-term_plan_of_standardization_Technet.pdf (дата обращения: 01.10.2021).
31. Приказ Росстандарта от 05.02.2019 г. № 166 «О внесении изменений в Программу национальной стандартизации на 2019 год, утверждённую приказом Федерального агентства по техническому регулированию и метрологии от 1 ноября 2018 г. № 2285» // Росстандарт. URL: https://www.gost.ru/portal/gost/home/activity/standardization (дата обращения: 01.10.2021).
32. Протокол результатов общественного обсуждения эффективности применения принятых во исполнение планов мероприятий «дорожных карт» по совершенствованию законодательства и устранению административных барьеров в целях обеспечения реализации Национальной технологической инициативы нормативных правовых актов и документов по стандартизации, достигнутых целей их применения не менее чем в течение одного года с даты начала применения (реализации) акта // Национальная технологическая инициатива. URL: https://nti2035.ru/upload/351708.2.protocol.pdf дата обращения: 01.10.2021).
33. Дорожная карта развития «сквозной» цифровой технологии «Нейротехнологии и искусственный интеллект» // Министерство цифрового развития, связи и массовых коммуникаций Российской Федерации. URL: https://digital.gov.ru/ru/documents/6658/ (дата обращения: 01.10.2021).
34. Паспорт Федерального проекта «Нормативное регулирование цифровой среды» // Министерство экономического развития Российской Федерации. URL: https://www.economy.gov.ru/material/directions/gosudarstvennoe_upravlenie/normativnoe_regulirovanie_cifrovoy_sredy/ (дата обращения: 01.10.2021). .
35. Национальная стратегия развития искусственного интеллекта на период до 2030 года, утверждённая Указом Президента РФ от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» // СЗ РФ. 2019. № 41. Ст. 5700.
36. Федеральный проект «Искусственный интеллект» // Министерство экономического развития Российской Федерации. URL: https://www.economy.gov.ru/material/directions/tehnologicheskoe_razvitie/federalnyy_proekt_iskusstvennyy_intellekt/ (дата обращения: 01.10.2021).
37. Перспективная программа стандартизации по приоритетному направлению «искусственный интеллект» на период 2021-2024 годы, утверждённая Министерством экономического развития Российской Федерации и Федеральным агентством по техническому регулированию и метрологии 22.12.2020 г. // URL: https://www.economy.gov.ru/material/news/v_rossii_poyavyatsya_standarty_v_oblasti_iskusstvennogo_intellekta.html (дата обращения: 01.10.2021).
38. Окончательная редакция проекта ГОСТ Р «Информационные технологии. Эталонная архитектура больших данных. Часть 1. Структура и процесс применения» // Технический комитет 164 «Искусственный интеллект». URL: https://www.tc164.ru/окончательные-редакции (дата обращения: 01.10.2021).
39. Окончательная редакция проекта ГОСТ-Р «Информационные технологии. Эталонная архитектура больших данных. Часть 2. Варианты использования и производные требования» // Технический комитет 164 «Искусственный интеллект». URL: https://www.tc164.ru/окончательные-редакции (дата обращения: 01.10.2021).
40. Проект ПНСТ «Информационные технологии. Большие данные. Типовая архитектура» // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/554715621 (дата обращения: 01.10.2021).
41. Первая редакция проекта ГОСТ Р «Информационные технологии. Эталонная архитектура больших данных. Часть 5. Направления стандартизации» // Технические комитет 164

«Искусственный интеллект». URL: https://www.tc164.ru/первые-редакции (дата обращения: 01.10.2021).
42. Окончательная редакция проекта ГОСТ Р «Информационные технологии. Большие данные. Техническое задание. Требования к содержанию и оформлению» // Технический комитет 164 «Искусственный интеллект». URL: https://www.tc164.ru/окончательные-редакции (дата обращения: 01.10.2021).
43. Первая редакция проекта ГОСТ Р «Информационные технологии – Искусственный интеллект – Структура управления процессами аналитики больших данных» // Технический комитет 164 «Искусственный интеллект». URL: https://www.tc164.ru/первые-редакции (дата обращения: 01.10.2021).
44. ISO/IEC 24668 Information technology — Artificial intelligence — Process management framework for big data analytics // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:24668:dis:ed-1:v1:en (дата обращения: 01.10.2021).
45. Raban, D.R., Gordon, A. The evolution of data science and big data research: A bibliometric analysis // Scientometrics. 2020. Vol 122. P. 1563–1581. https://doi.org/10.1007/s11192-020-03371-2.
46. Т.В. Ершова, Ю.Е. Хохлов, С.Б. Шапошник. Методология мониторинга развития и использования технологий работы с большими данными // Информационное общество. 2021. № 4–5. С. 2-32. https://doi.org/10.52605/16059921_2021_04_02
47. Kalantari, A., Kamsin, A., Kamaruddin, H.S. et al. A bibliometric approach to tracking big data research trends // J Big Data. 2017. Vol 4, № 30. https://doi.org/10.1186/s40537-017-0088-1.
48. de Meer, J. A Theory on Big Data. // INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge) / Draude, C., Lange, M. & Sick, B. (Hrsg.). Bonn: Gesellschaft für Informatik e.V. 2019. S. 261-269. https://doi.org/10.18420/inf2019_ws30.
49. Walshe R., Casey K., Kernan J., Fitzpatrick D. AI and Big Data Standardization: Contributing to United Nations Sustainable Development Goals // Journal of ICT Standardization: 2020: Vol 8 Iss 2. URL: https://journals.riverpublishers.com/index.php/JICTS/article/view/2649 (дата обращения: 01.10.2021).
50. Caballero I., Parody L., Bermejo I., Gómez López M.T., Gasca R.M., Piattini M. Service level agreement for data quality governed by Iso 8000-1X0 // Proceedings of the 19th International Conference on Information Quality, ICIQ. 2014. P. 114–127.
51. Xiao Y., Lu L.Y.Y., Liu J.S., Zhou Z. Knowledge diffusion path analysis of data quality literature: A main path analysis // Journal of Informetrics. 2014. Vol 8 Iss 3. P. 594–605. https://doi.org/10.1016/j.joi.2014.05.001.
52. Perin U. Reference Architectures and Standards for the Internet of Things and Big Data in Smart Manufacturing // International Journal of Recent Technology and Engineering. 2019. Vol 8 Iss 1C2. P. 884–887. https://doi.org/10.1109/FiCloud.2019.00041.
53. Heale B.S.E., Overby C.L., Del Fiol G. et al. Integrating genomic resources with electronic health records using the HL7 infobutton standard // Applied Clinical Informatics. 2016. Vol 7 Iss 3. P. 817–831. https://doi.org/10.4338/ACI-2016-04-RA-0058.
54. Raghunath N. A Standard for Benchmarking Big Data Systems // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014. Vol 8585. P. 193–201. https://doi.org/10.1007/978-3-319-10596-3_15.
55. Laney, D. 3D Data Management: Controlling Data Volume, Velocity, and Variety // META Group Research Note, 6, February 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (дата обращения: 01.10.2021).
56. Cloutier R., Muller G., Verma D., Nilchiani R., Hole E., Bone M. The Concept of Reference Architecture // Systems Engineering. 2009. Vol 13, Iss 1. P 14-27. https://doi.org/10.1002/sys.20129.
57. Big data: The next frontier for innovation, competition, and productivity. Report // McKinsey Global Institute. 2011. May. URL: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (дата обращения: 01.10.2021).
58. Mohd Rehan Ghazi Durgaprasad Gangodkar Hadoop, MapReduce and HDFS: A Developers Perspective // Procedia Computer Science. 2015. Vol 48. P. 45–50.

https://doi.org/10.1016/j.procs.2015.04.108. URL: https://www.sciencedirect.com/science/article/pii/S1877050915006171 (дата обращения: 01.10.2021).
59. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters // OSDI'04: Sixth Symposium on Operating System Design and Implementation. San Francisco, CA, 2004. P. 137-150. URL: https://static.googleusercontent.com/media/research.google.com/ru//archive/mapreduce-osdi04.pdf (дата обращения: 01.10.2021).
60. ГОСТ Р ИСО/МЭК 27002–2021 Информационные технологии. Методы и средства обеспечения безопасности. Свод норм и правил применения мер обеспечения информационной безопасности // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200179669 (дата обращения: 01.10.2021).
61. ГОСТ Р 56214–2014/ISO/TS 8000–1:2011 Качество данных. Часть 1. Обзор // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200114769?section=text (дата обращения: 01.10.2021).
62. ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:25012:ed-1:v1:en (дата обращения: 01.10.2021).
63. ISO 9000:2015 Quality management systems — Fundamentals and vocabulary // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:ru (дата обращения: 01.10.2021).
64. Пирсиг Р. Дзэн и искусство ухода за мотоциклом/ пер. с англ. М.Горшкова. – СПб.: «Симпозиум», 2002.
65. David J. Hand. Dark Data. Princeton University Press. Princeton and Oxford., 2020.
66. Dirty Data // Techopedia. URL: https://www.techopedia.com/definition/1194/dirty-data (дата обращения: 01.10.2021).
67. Raikov A.N., Avdeeva Z., and Ermakov A. Big Data Refining on the Base of Cognitive Modeling // Proceedings of the 1st IFAC Conference on Cyber-Physical&Human-Systems, Florianopolis, Brazil. 2016. P. 147-152. https://doi.org/10.1016/j.ifacol.2016.12.205.
68. Raikov A. Cognitive Semantics of Artificial Intelligence: A New Perspective // Springer Singapore, Topics: Computational Intelligence XVII, 2021. https://doi.org/10.1007/978-981-33-6750-0.
69. Райков А.Н., Котельников В.А. Предвосхищение будущего цифровизации // Информатизация и связь. 2019. № 1. – С. 7–11. https://doi.org/10.34219/2078-8320-2019-10-1-7-11.
70. ISO/IEC 2382:2015 Information technology — Vocabulary // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en (дата обращения: 01.10.2021).

BIG DATA STANDARDIZATION: INTERNATIONAL AND NATIONAL STANDARDS

Averkin, Alexei Nikolaevich

Candidate of physical and mathematical sciences, associate professor
Plekhanov Russian University of Economics, Educational and scientific laboratory for artificial intelligence, neurotechnology and business analytics, leading researcher
Moscow, Russia
averkin2003@inbox.ru

Afanasev, Sergei Dmitrievich

Candidate of law sciences
Lomonosov Moscow State University, National center for digital economy, lead specialist
Moscow Region State University, Institute of economics, management and law, Faculty of law, Department of constitutional and municipal law, associate professor
Moscow, Russian Federation
sergei.afanasev@digital.msu.ru

Mikryukov, Andrey Alexandrovich

Candidate of engineering sciences, associate professor
Plekhanov Russian University of Economics, Institute of mathematics, information systems and digital economy, Department of applied informatics and information security, associate professor
Moscow, Russian Federation
mikrukov.aa@rea.ru

Padzhev, Valentin Valentinovich

Institute of the Information Society, head of Directorate of law
Moscow, Russian Federation
vpadzhev@iis.ru

Raikov, Alexander Nikolaevich

Doctor of engineering sciences, professor
Lomonosov Moscow State University, National center for digital economy, head of Department for intellectual technologies
V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, senior researcher
Moscow, Russia Federation
Alexander.N.Raikov@gmail.com

Hohlov, Yuri Evgenyevich

Candidate of physical and mathematical sciences, associate professor
Institute of the Information Society, chairman of the Board of directors
Plekhanov Russian University of Economics, IIS-based Digital economy department, scientific advisor
Moscow, Russian Federation
yuri.hohlov@iis.ru

Khramtsovskaya, Natalya Alexandrovna

Candidate of historical sciences
Electronic Office Systems LLC, leading expert in documentation management
Moscow, Russian Federation
sspchram@tochka.ru

Abstract

The standards that establish requirements for big data technologies are designed to increase the efficiency of using these technologies in various sectors of the economy. The article examines the features of international and Russian national industry-wide data standards based on ISO/IEC 20546, 20547-X, 9000 series and 5259-X series projects in terms of basic concepts, reference architecture and requirements for big data and their quality, as well as approaches

to standardizing customer requirements for actions related to the use of big data. It is concluded that the development of documents on big data standardization in Russia is of great importance and entails the need to increase the pace of standardization.

Keywords

big data; data; standardization; national standard; international standard; ISO; reference architecture; data quality; artificial intelligence

References

1. ISO/IEC 20546:2019 Information technology – Big data – Overview and vocabulary) // International Organization for Standardization. URL: https://www.iso.org/standard/68305.html (accessed: 01.10.2021).
2. GOST R ISO/MEK 20546–2021 «Informacionnye tekhnologii. Bol'shie dannye. Obzor i slovar'» // Rossijskij institut standartizacii. URL: https://www.gostinfo.ru/catalog/Details/?id=6859575 (accessed: 01.10.2021).
3. ISO/IEC DIS 22989 Information technology – Artificial intelligence – Artificial intelligence concepts and terminology // International Organization for Standardization. URL: https://www.iso.org/standard/74296.html (accessed: 01.10.2021).
4. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh Ch., Byers A.H. Big data: The next frontier for innovation, competition, and productivity // McKinsey Global Institute. P. 12. URL: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation# (accessed: 01.10.2021).
5. Big Data interoperability Framework. V1.0 Final Version // National Institute of Standards and Technology. URL: https://bigdatawg.nist.gov/V1_output_docs.php (accessed: 01.10.2021).
6. Big Data interoperability Framework. V2.0 Final Version // National Institute of Standards and Technology. URL: https://bigdatawg.nist.gov/V2_output_docs.php (accessed: 01.10.2021).
7. Big Data interoperability Framework. V3.0 Final Version // National Institute of Standards and Technology. URL: https://bigdatawg.nist.gov/V3_output_docs.php (accessed: 01.10.2021).
8. Big Data: Big today, normal tomorrow. ITU T Technology Watch Report. November 2013. // International Telecommunication Union. URL: https://www.itu.int/en/ITU T/techwatch/Pages/big data standards.aspx (accessed: 01.10.2021).
9. Big data: Preliminary Report 2014. ISO/IEC JTC1, 2015. // International Organization for Standardization. URL: https://www.iso.org/files/live/sites/isoorg/files/developing_standards/docs/en/big_data_report-jtc1.pdf (accessed: 01.10.2021).
10. ITU-T Y.3600 (11/2015) Big data – Cloud computing based requirements and capabilities. // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=12584&lang=en (accessed: 01.10.2021).
11. ITU-T Y Suppl. 40 (07/2016) Big data standardization roadmap // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=13022&lang=en (accessed: 01.10.2021).
12. ISO/IEC TR 20547-2:2018 Information technology – Big data reference architecture – Part 2: Use cases and derived requirements // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:20547:-2:ed-1:v1:en (accessed: 01.10.2021).
13. ISO/IEC TR 20547-5:2018 Information technology – Big data reference architecture – Part 5: Standards roadmap // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:20547:-5:ed-1:v1:en (accessed: 01.10.2021).
14. ISO/IEC TR 20547-1:2020 Information technology – Big data reference architecture – Part 1: Framework and application process // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:20547:-1:ed-1:v1:en (accessed: 01.10.2021).
15. ISO/IEC 20547-3:2020 Information technology – Big data reference architecture – Part 3: Reference architecture // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:20547:-3:ed-1:v1:en (accessed: 01.10.2021).
16. ISO/IEC 20547-4:2020 Information technology – Big data reference architecture – Part 4: Security and privacy // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:20547:-4:ed-1:v1:en (accessed: 01.10.2021).

17. Walshe R. The Road to Big Data Standardisation // The Elements of Big Data Value / E. Curry et al. (eds.). Springer, Cham, 2021. https://doi.org/10.1007/978-3-030-68176-0_14.
18. ISO/IEC WD 5259-1 Data quality for analytics and ML – Part 1: Overview, terminology, and examples // International Organization for Standardization. URL: https://www.iso.org/standard/81088.html (accessed: 01.10.2021).
19. ISO/IEC AWI 5259-2 Data quality for analytics and ML – Part 2: Data quality measures // International Organization for Standardization. URL: https://www.iso.org/standard/81860.html (accessed: 01.10.2021).
20. ISO/IEC WD 5259-3 Data quality for analytics and ML – Part 3: Data quality management requirements and guidelines // International Organization for Standardization. URL: https://www.iso.org/standard/81092.html (accessed: 01.10.2021).
21. ISO/IEC WD 5259-4 Data quality for analytics and ML – Part 4: Data quality process framework // International Organization for Standardization. URL: https://www.iso.org/standard/81093.html (accessed: 01.10.2021).
22. ITU-T Y.3603 Big data – Requirements and conceptual model of metadata for data catalogue» // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=14137&lang=en (accessed: 01.10.2021).
23. ITU-T Y.3604 Big data – Overview and requirements for data preservation» // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=14138&lang=en (accessed: 01.10.2021).
24. Resolution 166. Registration of a PWI entitled «Information technology – Artificial intelligence – Data life cycle framework». ISO/IEC JTC 1/SC 42 "Artificial intelligence". Resolutions taken during the Closing Plenary - JTC 1/SC 42 - 7 May 2021.
25. ITU-T Y.3651 Big-data-driven networking – mobile network traffic management and planning // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=13818&lang=en (accessed: 01.10.2021).
26. ITU-T Y.3653 Big data driven networking – functional architecture // International Telecommunication Union. URL: https://www.itu.int/ITU-T/recommendations/rec.aspx?id=14615&lang=en (accessed: 01.10.2021).
27. Prikaz Federal'nogo agentstva po tekhnicheskomu regulirovaniyu i metrologii (Rosstandart) ot 25.07.2019 № 1732 (red. ot 20.01.2021) «O sozdanii tekhnicheskogo komiteta po standartizacii "Iskusstvennyj intellekt"» // Rosstandart. https://www.rst.gov.ru/portal/gost/home/activity/documents/orders#/order/104460 (accessed: 01.10.2021).
28. Prikaz Federal'nogo agentstva po tekhnicheskomu regulirovaniyu i metrologii (Rosstandart) ot 20.08.2020 № 1415 «O vnesenii izmenenij v prikaz Federal'nogo agentstva po tekhnicheskomu regulirovaniyu i metrologii ot 25 iyulya 2019 g. № 1732 «O sozdanii tekhnicheskogo komiteta po standartizacii "Iskusstvennyj intellect"» // Rosstandart. https://www.gost.ru/portal/gost/home/activity/documents/orders#/order/179415 (accessed: 01.10.2021).
29. Rasporyazhenie Pravitel'stva RF ot 23.03.2018 N 482-r (red. ot 28.05.2020) «Ob utverzhdenii plana meropriyatij («dorozhnoj karty») po sovershenstvovaniyu zakonodatel'stva i ustraneniyu administrativnyh bar'erov v celyah obespecheniya realizacii Nacional'noj tekhnologicheskoj iniciativy po napravleniyu «Tekhnet» (peredovye proizvodstvennye tekhnologii)» // SZ RF. 2018. № 15 (ch. V). St. 2173.
30. Rossijskaya venchurnaya kompaniya. https://www.rvc.ru/upload/iblock/ac9/Long-term_plan_of_standardization_Technet.pdf (accessed: 01.10.2021).
31. Prikaz Rosstandarta ot 05.02.2019 g. № 166 «O vnesenii izmenenij v Programmu nacional'noj standartizacii na 2019 god, utverzhdyonnuyu prikazom Federal'nogo agentstva po tekhnicheskomu regulirovaniyu i metrologii ot 1 noyabrya 2018 g. № 2285» // Rosstandart. URL: https://www.gost.ru/portal/gost/home/activity/standardization (accessed: 01.10.2021).
32. Protokol rezul'tatov obshchestvennogo obsuzhdeniya effektivnosti primeneniya prinyatyh vo ispolnenie planov meropriyatij «dorozhnyh kart» po sovershenstvovaniyu zakonodatel'stva i ustraneniyu administrativnyh bar'erov v celyah obespecheniya realizacii Nacional'noj tekhnologicheskoj iniciativy normativnyh pravovyh aktov i dokumentov po standartizacii, dostignutyh celej ih primeneniya ne menee chem v techenie odnogo goda s daty nachala

primeneniya (realizacii) akta // https://nti2035.ru/upload/351708.2.protocol.pdf (accessed: 01.10.2021).
33. Dorozhnaya karta razvitiya «skvoznoj» cifrovoj tekhnologii «Nejrotekhnologii i iskusstvennyj intellekt» // Ministerstvo cifrovogo razvitiya, svyazi i massovyh kommunikacij Rossijskoj Federacii. URL: https://digital.gov.ru/ru/documents/6658/ (accessed: 01.10.2021).
34. Pasport Federal'nogo proekta «Normativnoe regulirovanie cifrovoj sredy» // Ministerstvo ekonomicheskogo razvitiya Rossijskoj Federacii. URL: https://www.economy.gov.ru/material/directions/gosudarstvennoe_upravlenie/normativnoe_regulirovanie_cifrovoy_sredy/ (accessed: 01.10.2021).
35. Nacional'naya strategiya razvitiya iskusstvennogo intellekta na period do 2030 goda, utverzhdyonnaya Ukazom Prezidenta RF ot 10.10.2019 № 490 «O razvitii iskusstvennogo intellekta v Rossijskoj Federacii» // SZ RF. 2019. № 41. St. 5700.
36. Federal'nyj proekt «Iskusstvennyj intellekt» // Ministerstvo ekonomicheskogo razvitiya Rossijskoj Federacii. URL: https://www.economy.gov.ru/material/directions/tehnologicheskoe_razvitie/federalnyy_proekt_iskusstvennyy_intellekt/ (accessed: 01.10.2021).
37. Perspektivnaya programma standartizacii po prioritetnomu napravleniyu «iskusstvennyj intellekt» na period 2021-2024 gody, utverzhdyonnaya Ministerstvom ekonomicheskogo razvitiya Rossijskoj Federacii i Federal'nym agentstvom po tekhnicheskomu regulirovaniyu i metrologii 22.12.2020 g. // URL: https://www.economy.gov.ru/material/news/v_rossii_poyavyatsya_standarty_v_oblasti_iskusstvennogo_intellekta.html (accessed: 01.10.2021).
38. Okonchatel'naya redakciya proekta GOST R «Informacionnye tekhnologii. Etalonnaya arhitektura bol'shih dannyh. CHast' 1. Struktura i process primeneniya» // Tekhnicheskij komitet 164 «Iskusstvennyj intellekt». URL: https://www.tc164.ru/окончательные-редакции. (accessed: 01.10.2021).
39. Okonchatel'naya redakciya proekta GOST-R «Informacionnye tekhnologii. Etalonnaya arhitektura bol'shih dannyh. CHast' 2. Varianty ispol'zovaniya i proizvodnye trebovaniya» // Tekhnicheskij komitet 164 «Iskusstvennyj intellekt». URL: https://www.tc164.ru/окончательные-редакции (accessed: 01.10.2021).
40. Proekt PNST «Informacionnye tekhnologii. Bol'shie dannye. Tipovaya arhitektura» // Elektronnyj fond pravovyh i normativno-tekhnicheskih dokumentov. URL: https://docs.cntd.ru/document/554715621 (accessed: 01.10.2021).
41. Pervaya redakciya proekta GOST R «Informacionnye tekhnologii. Etalonnaya arhitektura bol'shih dannyh. CHast' 5. Napravleniya standartizacii» // Tekhnicheskie komitet 164 «Iskusstvennyj intellekt». URL: https://www.tc164.ru/первые-редакции (accessed: 01.10.2021).
42. Okonchatel'naya redakciya proekta GOST R «Informacionnye tekhnologii. Bol'shie dannye. Tekhnicheskoe zadanie. Trebovaniya k soderzhaniyu i oformleniyu» // Tekhnicheskij komitet 164 «Iskusstvennyj intellekt». URL: https://www.tc164.ru/окончательные-редакции (accessed: 01.10.2021).
43. Pervaya redakciya proekta GOST R «Informacionnye tekhnologii – Iskusstvennyj intellekt – Struktura upravleniya processami analitiki bol'shih dannyh» // Tekhnicheskij komitet 164 «Iskusstvennyj intellekt». URL: https://www.tc164.ru/первые-редакции (accessed: 01.10.2021).
44. ISO/IEC 24668 Information technology — Artificial intelligence — Process management framework for big data analytics // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:24668:dis:ed-1:v1:en (accessed: 01.10.2021).
45. Raban, D.R., Gordon, A. The evolution of data science and big data research: A bibliometric analysis // Scientometrics. 2020. Vol 122. P. 1563–1581. https://doi.org/10.1007/s11192-020-03371-2.
46. T.V. Ershova, Yu.E. Hohlov, S.B. Shaposhnik. Metodologiya monitoringa razvitiya i ispol'zovaniya tekhnologij raboty s bol'shimi dannymi // Informacionnoe obshchestvo. 2021. № 4–5. S. 2–32. https://doi.org/10.52605/16059921_2021_04_02
47. Kalantari, A., Kamsin, A., Kamaruddin, H.S. et al. A bibliometric approach to tracking big data research trends // J Big Data. 2017. Vol 4, № 30. https://doi.org/10.1186/s40537-017-0088-1.
48. de Meer, J. A Theory on Big Data. // INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge) / Draude, C., Lange, M. & Sick, B. (Hrsg.). Bonn: Gesellschaft für Informatik e.V. 2019. S. 261-269. https://doi.org/10.18420/inf2019_ws30.

49. Walshe R., Casey K., Kernan J., Fitzpatrick D. AI and Big Data Standardization: Contributing to United Nations Sustainable Development Goals // Journal of ICT Standardization: 2020: Vol 8 Iss 2. URL: https://journals.riverpublishers.com/index.php/JICTS/article/view/2649 (accessed: 01.10.2021).
50. Caballero I., Parody L., Bermejo I., Gómez López M.T., Gasca R.M., Piattini M. Service level agreement for data quality governed by Iso 8000-1X0 // Proceedings of the 19th International Conference on Information Quality, ICIQ. 2014. P. 114–127.
51. Xiao Y., Lu L.Y.Y., Liu J.S., Zhou Z. Knowledge diffusion path analysis of data quality literature: A main path analysis // Journal of Informetrics. 2014. Vol 8 Iss 3. P. 594–605. https://doi.org/10.1016/j.joi.2014.05.001.
52. Perin U. Reference Architectures and Standards for the Internet of Things and Big Data in Smart Manufacturing // International Journal of Recent Technology and Engineering. 2019. Vol 8 Iss 1C2. P. 884–887. https://doi.org/10.1109/FiCloud.2019.00041.
53. Heale B.S.E., Overby C.L., Del Fiol G. et al. Integrating genomic resources with electronic health records using the HL7 infobutton standard // Applied Clinical Informatics. 2016. Vol 7 Iss 3. P. 817–831. https://doi.org/10.4338/ACI-2016-04-RA-0058.
54. Raghunath N. A Standard for Benchmarking Big Data Systems // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014. Vol 8585. P. 193–201. https://doi.org/10.1007/978-3-319-10596-3_15.
55. Laney, D. 3D Data Management: Controlling Data Volume, Velocity, and Variety // META Group Research Note, 6, February 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed: 01.10.2021).
56. Cloutier R., Muller G., Verma D., Nilchiani R., Hole E., Bone M. The Concept of Reference Architecture // Systems Engineering. 2009. Vol 13, Iss 1. P 14-27. https://doi.org/10.1002/sys.20129.
57. Big data: The next frontier for innovation, competition, and productivity. Report // McKinsey Global Institute. 2011. May. URL: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (accessed: 01.10.2021).
58. Mohd Rehan Ghazi Durgaprasad Gangodkar Hadoop, MapReduce and HDFS: A Developers Perspective // Procedia Computer Science. 2015. Vol 48. P. 45–50. https://doi.org/10.1016/j.procs.2015.04.108. URL: https://www.sciencedirect.com/science/article/pii/S1877050915006171 (accessed: 01.10.2021).
59. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters // OSDI'04: Sixth Symposium on Operating System Design and Implementation. San Francisco, CA, 2004. P. 137-150. URL: https://static.googleusercontent.com/media/research.google.com/ru//archive/mapreduce-osdi04.pdf (accessed: 01.10.2021).
60. GOST R ISO/MEK 27002–2021 Informacionnye tekhnologii. Metody i sredstva obespecheniya bezopasnosti. Svod norm i pravil primeneniya mer obespecheniya informacionnoj bezopasnosti // Elektronnyj fond pravovyh i normativno-tekhnicheskih dokumentov. URL: https://docs.cntd.ru/document/1200179669 (accessed: 01.10.2021).
61. GOST R 56214–2014/ISO/TS 8000-1:2011 Kachestvo dannyh. CHast' 1. Obzor // Elektronnyj fond pravovyh i normativno-tekhnicheskih dokumentov. URL: https://docs.cntd.ru/document/1200114769?section=text (accessed: 01.10.2021).
62. ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:25012:ed-1:v1:en (accessed: 01.10.2021).
63. ISO 9000:2015 Quality management systems — Fundamentals and vocabulary // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:ru (accessed: 01.10.2021).
64. Pirsig R. Dzen i iskusstvo uhoda za motociklom/ per. s angl. M.Gorshkova. – SPb.: «Simpozium», 2002.
65. David J. Hand. Dark Data. Princeton University Press. Princeton and Oxford., 2020.
66. Dirty Data // Techopedia. URL: https://www.techopedia.com/definition/1194/dirty-data (accessed: 01.10.2021).

67. Raikov A.N., Avdeeva Z., and Ermakov A. Big Data Refining on the Base of Cognitive Modeling // Proceedings of the 1st IFAC Conference on Cyber-Physical&Human-Systems, Florianopolis, Brazil. 2016. P. 147-152. https://doi.org/10.1016/j.ifacol.2016.12.205.
68. Raikov A. Cognitive Semantics of Artificial Intelligence: A New Perspective // Springer Singapore, Topics: Computational Intelligence XVII, 2021. https://doi.org/10.1007/978-981-33-6750-0.
69. Rajkov A.N., Kotel'nikov V.A. Predvoskhishchenie budushchego cifrovizacii // Informatizaciya i svyaz'. 2019. № 1. – С. 7 – 11. https://doi.org/10.34219/2078-8320-2019-10-1-7-11.
70. ISO/IEC 2382:2015 Information technology — Vocabulary // International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en (accessed: 01.10.2021).