ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 71484.3— 2024 (ИСО/МЭК 5259-3:2024)

Искусственный интеллект

КАЧЕСТВО ДАННЫХ ДЛЯ АНАЛИТИКИ И МАШИННОГО ОБУЧЕНИЯ

Часть 3

Требования и рекомендации по управлению качеством данных

(ISO/IEC 5259-3:2024, MOD)

Издание официальное

Москва
Российский институт стандартизации
2024

gost.e-ecolog.ru

Предисловие

1 ПОДГОТОВЛЕН Научно-образовательным центром компетенций в области цифровой экономики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова» (МГУ имени М.В. Ломоносова) и Обществом с ограниченной ответственностью «Институт развития информационного общества» (ИРИО) на основе собственного перевода на русский язык англоязычной версии стандарта, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 28 октября 2024 г. № 1538-ст

4 Настоящий стандарт является модифицированным по отношению к международному стандарту ИСО/МЭК 5259-3:2024 «Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 3. Требования и рекомендации по управлению качеством данных» (ISO/IEC 5259-3:2024 «Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines», MOD) путем изменения отдельных фраз (слов, значений, показателей, ссылок), которые выделены в тексте курсивом.

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте, приведены в дополнительном приложении ДА

5 ВВЕДЕН ВПЕРВЫЕ

*Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (*gost.e-ecolog.ru

© ISO, 2024
© IEC, 2024
© Оформление. ФГБУ «Институт стандартизации», 2024

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

Содержание

1 Область применения 1
2 Нормативные ссылки 1
3 Термины и определения 2
4 Сокращения 2
5 Способы применения 2
6 Общее управление качеством данных 2
6.1 Цель 2
6.2 Общие положения 2
6.3 Требования и рекомендации 3
6.4 Результаты 5
7 Управление качеством данных с учетом жизненного цикла 5
7.1 Цель 5
7.2 Общие положения 5
7.3 Требования и рекомендации 8
7.4 Результаты 14
8 Горизонтальные процессы 15
8.1 Цель 15
8.2 Общие положения 15
8.3 Требования и рекомендации 15
8.4 Результаты 17
9 Управление качеством данных в цепочках поставок 17
9.1 Цель 17
9.2 Требования и рекомендации 17
9.3 Результаты 18
10 Управление инструментами для обработки данных 18
10.1 Цель 18
10.2 Требования и рекомендации 18
10.3 Результаты 18
11 Управление зависимостями, связанными с качеством данных 18
11.1 Цель 18
11.2 Требования и рекомендации 18
11.3 Результаты 18
12 Управление качеством данных в проекте 19
12.1 Цель 19
12.2 Требования и рекомендации 19
12.3 Спецификация и управление требованиями к качеству данных 19
12.4 Роли и ответственность в управлении качеством данных 20
12.5 Адаптация мероприятий по обеспечению качества данных 20
12.6 Планирование и координация деятельности по обеспечению качества данных 20
12.7 Продвижение по жизненному циклу качества данных 21
12.8 Обоснование качества данных 21
12.9 Вывод из эксплуатации 21
12.10 Результаты 21
Приложение ДА (справочное) Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте 22
Библиография 23

Введение

Качество продуктов и услуг на основе аналитики и машинного обучения зависит от качества данных, используемых для обучения моделей машинного обучения. Следовательно, управление качеством данных имеет важное значение, поскольку оно часто помогает обеспечить успех аналитики и использования технологий машинного обучения.

Внедрение системы управления качеством данных облегчает управление качеством продуктов и услуг, в которых используются технологии аналитики и машинного обучения. Настоящий стандарт определяет терминологию, требования и рекомендации по обмену информацией, а также процедур по согласованию и по управлению качеством данных. Система управления качеством данных обеспечивает прозрачность и возможность проверки посредством самооценки или оценки третьей стороной. Это способствует удовлетворению интересов заинтересованных сторон, а также позволяет управлять требованиями к качеству, производительности и представлению данных. В частности, настоящий стандарт определяет требования к системе управления качеством данных со ссылками на показатели качества данных, которые применимы к наиболее часто используемым технологиям аналитики и машинного обучения.

Поскольку требования к качеству данных различаются в зависимости от контекста и сферы применения, в настоящем стандарте представлен типовой набор требований и рекомендаций, относящийся к общим стадиям жизненного цикла данных. Жизненный цикл данных, как правило, тесно интегрирован с сопутствующим жизненным циклом системы искусственного интеллекта и, следовательно, имеет несколько взаимозависимостей. Настоящий стандарт не предписывает, какой жизненный цикл для системы искусственного интеллекта следует использовать. Вместо этого он предоставляет общие рекомендации, которые позволяют гибко сочетать несколько моделей жизненного цикла при условии, что процессы жизненного цикла могут быть сопоставлены.

Настоящий стандарт является частью серии стандартов ИСО/МЭК 5259. Другие части данной серии включают следующие стандарты:

- ИСО/МЭК 5259-1 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры;
- ISO/IEC FDIS 5259-2 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 2. Показатели качества данных;
- ИСО/МЭК 259-4 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 4. Инструментарий для мониторинга качества данных;
- ISO/IEC FDIS 5259-5 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 5. Управление качеством данных *[1]*;
- ISO/IEC CD TR 5259-6 Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 6. Структура визуализации качества данных *[2]*.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Искусственный интеллект

КАЧЕСТВО ДАННЫХ ДЛЯ АНАЛИТИКИ И МАШИННОГО ОБУЧЕНИЯ

Часть 3

Требования и рекомендации по управлению качеством данных

Artificial intelligence. Data quality for analytics and machine learning. Part 3. Data quality management requirements and guidelines

Дата введения — 2025—01—01

1 Область применения

Настоящий стандарт устанавливает требования и дает рекомендации по созданию, внедрению, поддержанию и постоянному улучшению качества данных, используемых для аналитики и машинного обучения.

Настоящий стандарт не описывает детально процессы, методы или показатели, но определяет требования и рекомендации для процесса управления качеством, а также перечень эталонных процессов и методов, которые могут быть адаптированы для соответствия требованиям данного стандарта.

Требования и рекомендации, изложенные в настоящем стандарте, являются типовыми и применимы к любой организации, независимо от размера, типа и рода деятельности.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ Р 71476 (ИСО/МЭК 22989:2022) Искусственный интеллект. Концепции и терминология искусственного интеллекта

ГОСТ Р 71484.1 (ИСО/МЭК 5259-1:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, термины и примеры

ГОСТ Р 71484.4 (ИСО/МЭК 5259-4:2024) Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 4. Структура процесса управления качеством данных

ГОСТ Р ИСО 9001 Системы менеджмента качества. Требования

ГОСТ Р ИСО/МЭК 42001 Искусственный интеллект. Система управления

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная

Издание официальное

ссылка, внесено изменение, затрагивающая положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями:

3.1 **заявление о качестве данных** (data quality claim): Утверждение о том, в какой степени данные удовлетворяют требованиям к качеству.

3.2 **план управления качеством данных** (data quality plan): Описание методов, процессов и способов распределения ресурсов для достижения целей в области качества данных как результата планирования качества данных.

3.3 **планирование качества данных** (data quality planning): Сформулированное в результате планирования описание методов, процессов и способов распределения ресурсов для достижения целей качества данных.

3.4

соглашение о взаимодействии при разработке (development interface agreement, DIA): Соглашение между заказчиком и поставщиком, в котором указывается ответственность за действия, доказательства или результаты работы, подлежащие обмену между сторонами и связанные с разработкой изделий или элементов.

Примечание — Соглашение о взаимодействии при разработке относится к стадии разработки, тогда как договор на поставку относится к стадии производства.

[*ГОСТ Р ИСО 26262-1—2020, раздел 3.32*]

4 Сокращения

В настоящем стандарте применены следующие сокращения:

ИИ — искусственный интеллект;

МО — машинное обучение.

5 Способы применения

Настоящий стандарт может применяться в одном или нескольких случаях, например:

- организацией для создания и адаптации процесса управления качеством данных при использовании данных в аналитике и машинном обучении, а также для постоянного улучшения процессов;
- в проекте машинного обучения для определения, отслеживания и оценки требований к качеству данных;
- пользователем данных и обладателем данных для совместного определения характеристик качества данных и обеспечения соблюдения согласованных требований, что облегчает заключение соглашения о передаче данных.

Примечание — Организация может запросить гарантии конфиденциальности и доказательства, подтверждающие надлежащее использование.

6 Общее управление качеством данных

6.1 Цель

Целью процесса управления качеством данных является реализация приемлемых (т. е. повторяемых и проверяемых) процессов для обеспечения качества данных и надежного удовлетворения заданному набору требований, установленных организацией.

6.2 Общие положения

Качество данных влияет на результаты аналитики и использования алгоритмов машинного обучения. Качество данных зависит от внутренне присущих характеристик и системно-зависимых

характеристик. Данные могут подходить для одного приложения, но не подходить для другого. Настоящий стандарт помогает установить и поддерживать качество данных для каждого приложения аналитики и машинного обучения.

6.3 Требования и рекомендации

6.3.1 Общие положения

Следующие требования и рекомендации применимы ко всей организации.

6.3.2 Культура качества данных

Организация должна поддерживать культуру качества данных. Организация должна:

a) иметь правила и процессы, способствующие достижению качества (согласно настоящему стандарту), с учетом модели качества данных, применяемых к соответствующим продуктам и услугам;

b) определять и внедрять процессы управления качеством данных и выполнять соответствующие мероприятия по обеспечению качества данных;

c) интегрировать процессы и действия по управлению качеством данных, насколько это возможно, в другие процессы и мероприятия по управлению, такие как общее управление качеством и управление рисками;

d) документировать выполненные действия;

e) предоставлять ресурсы, достаточные для управления качеством данных;

f) контролировать и, по мере необходимости, анализировать и совершенствовать процессы управления качеством данных;

g) предоставлять требуемые полномочия персоналу, участвующему в процессе обеспечения качества;

h) доводить до сведения сотрудников политику обеспечения качества данных внутри организации.

6.3.3 Решение проблемных вопросов, связанных с качеством данных

Организация должна обеспечивать соответствие требованиям, связанным с качеством данных, посредством:

a) реализации процессов информирования, анализа, оценки, решения и закрытия вопросов;

b) документирования вопросов, связанных с качеством данных;

c) делегирования вопросов, которые не удается решить, или их эскалации на более высокий уровень управления для урегулирования.

Примечания

1 Одним из способов разрешения и закрытия проблемных вопросов, связанных с качеством данных, может быть ограничение или уточнение сферы охвата проекта МО.

2 Проблемный вопрос о качестве данных может быть закрыт путем реализации соответствующих мер или принятия решения о закрытии вопроса на основе заданных критериев.

6.3.4 Управление компетенциями

Организация должна управлять компетенциями сотрудников посредством:

a) документирования необходимых навыков и инструментов для обработки данных;

b) обеспечения наличия у привлеченного персонала достаточной квалификации для осуществления своей деятельности и выполнения своих обязанностей;

c) ведения учета сотрудников с обозначением их уровня владения необходимыми навыками и инструментами;

d) ведения соответствующих записей об обучении и опыте, подтверждающих утверждения о наличии соответствующих навыков. Организация может использовать внешние источники получения требуемых компетенций.

6.3.5 Управление ресурсами

Организация должна предоставить ресурсы, необходимые для управления качеством данных, включая:

a) программное обеспечение, обучение и поддержку, необходимые для управления качеством данных;

b) ИТ-инфраструктуру или сервисы, необходимые для управления качеством данных (например, вычислительные ресурсы, системы хранения, сеть);

c) персонал, обладающий необходимыми навыками для управления качеством данных.

6.3.6 Интеграция системы управления

Организация должна интегрировать свою деятельность по управлению качеством данных в существующую систему управления, включая системы управления качеством продукции или услуг, а также в разработку и использование систем ИИ.

Последствия, связанные с двойной ролью заинтересованных сторон, должны регулироваться системой менеджмента качества, в том числе для смягчения любых конфликтов интересов.

Примечания

1 Руководство заинтересованной стороны может учитывать возможность выполнения отдельным сотрудником нескольких ролей. Пользователь продуктов или услуг на основе аналитики и машинного обучения также может быть обладателем или создателем данных.

2 Организации могут использовать *ГОСТ Р ИСО/МЭК 42001* для определения системы менеджмента для разработки или использования систем ИИ.

3 Организации могут использовать *ГОСТ Р ИСО 9001* или другие отраслевые системы менеджмента качества для определения своей системы менеджмента качества.

6.3.7 Документация

Документация должна быть понятной заинтересованным сторонам проекта с соответствующими ролями. Документы на языке, который не понимает та или иная заинтересованная сторона, должны сопровождаться кратким изложением на понятном для нее языке.

При необходимости документация должна быть доступна заинтересованным сторонам, имеющим соответствующее разрешение. Затраты на доступ должны быть сведены к минимуму.

Документация должна содержать всю требуемую информацию или ссылки, необходимые для того, чтобы сделать ее понятной будущим заинтересованным сторонам, которые не участвуют в текущем проекте. Это позволит им оценить набор данных с точки зрения возможности потенциального повторного использования, частичного или полного.

6.3.8 Аудит и оценка качества данных

При необходимости внедренные процессы должны подвергаться аудиту, который должен основываться на оценке:

a) соответствия плана управления качеством данных правилам и процессам организации;

b) аргументов и обоснований, подробно описывающих, как выполнялись требования модели качества данных;

c) аргументов, подробно описывающих, как были достигнуты цели плана управления качеством данных;

d) полноты, последовательности и правильности плана управления качеством данных и всех результатов в соответствии с настоящим стандартом;

e) рекомендаций по улучшению качества данных.

Должна проводиться оценка качества данных, основанная на определении того, достигаются ли цели, определенные в настоящем стандарте, при использовании современных технологий и прикладных знаний в области инженерии.

План оценки качества данных должен быть сформирован на стадии спецификации. Оценка качества данных должна выполняться перед предоставлением данных (см. рисунок 1, стадия 7: предоставление данных) или через соответствующий интервал при использовании непрерывного обучения или при использовании потоковых данных.

Оценка качества данных может выполняться на подмножестве данных, если можно показать, что качество подмножества является репрезентативным для обеспечения качества полного набора данных.

6.3.9 Подтверждение и показатели качества данных

Качество данных должно быть подтверждено с использованием соответствующих показателей качества данных — см. *[3]*. Проверка качества данных должна как минимум содержать:

a) обоснованное подтверждение основных результатов, которое должно:
   1) быть завершено до предоставления данных;
   2) основываться на том, достигнуты ли цели настоящего стандарта;

b) аудиты качества реализуемых процессов;

c) оценку качества данных.

Все результаты должны быть проверены.

Персонал, выполняющий такие проверки, должен иметь доступ к сотрудникам, выполняющим соответствующие процессы, требуемую информацию и необходимые ресурсы.

Примечание — Проверку достоверности основных результатов можно делегировать, но ответственность остается за назначенным лицом.

6.3.10 Управление качеством данных в проекте

Организация должна управлять качеством данными конкретного проекта посредством:

a) создания соответствующего процесса управления качеством данных, который отвечает всем специфическим требованиям проекта;

b) ведения перечня соответствующих требований к качеству данных. Там, где это применимо, должны быть задокументированы количественные и качественные бенчмарки для показателей качества данных;

c) внедрения соответствующих процессов для определения и управления всеми показателями качества данных, имеющими отношение к проекту.

Процесс управления качеством данных конкретного проекта должен соответствовать требованиям раздела 12.

6.4 Результаты

Результаты процесса управления качеством данных должны включать:

a) принятые для конкретной организации правила и процессы обеспечения качества данных (например, согласно *ГОСТ Р 71484.4*);

b) доказательство наличия управления компетенциями;

c) подтверждение наличия системы управления качеством данных;

d) перечень примененных показателей качества данных;

e) документирование примененных бенчмарков показателей качества данных;

f) отчеты о выявленных отклонениях в качестве.

7 Управление качеством данных с учетом жизненного цикла

7.1 Цель

Целью жизненного цикла управления качеством данных является установление и поддержание качества данных на протяжении всего жизненного цикла данных. Пример модели жизненного цикла данных описан в *ГОСТ Р 71484.1* (рисунок 3).

7.2 Общие положения

7.2.1 Жизненный цикл управления качеством данных

Управление качеством данных должно осуществляться на всех стадиях жизненного цикла данных. Модель жизненного цикла управления качеством данных, показанная на рисунке 1, содержит рекомендации по соблюдению требований к качеству данных для использования в аналитике и машинном обучении. На рисунке выделены отдельные стадии, имеющие отношение к управлению качеством данных, что упрощает группировку и упорядочивание требований и рекомендаций, которые следует учитывать при управлении качеством данных. Модель жизненного цикла данных не предписывает временной порядок стадий. Стадии жизненного цикла управления качеством данных описаны в 7.2.2.

Проблемы с качеством данных могут возникнуть на любой стадии жизненного цикла данных. Для управления качеством данных необходимо формировать и поддерживать процессы управления качеством данных с начала жизненного цикла данных. Если организация делегирует ответственность за процесс, это делегирование должно быть задокументировано и отслеживаться.

Примечание — Обычно труднее обнаружить и устранить проблемы с качеством данных постфактум, нежели управлять рисками, связанными с качеством данных, когда они возникают впервые. Например, ошибок, возникших при сборе данных, легче избежать путем надлежащего управления качеством, чем пытаться обнаружить и исправить ошибки на более поздней стадии жизненного цикла данных.

7.2.2 Стадии жизненного цикла управления качеством данных

7.2.2.1 Потребность в качественных данных и концептуализация

Управление качеством данных начинается со стадии потребности в качественных данных и концептуализации. Потенциальные проблемы с качеством данных следует выявлять и устранять, когда становится очевидной первая потребность в данных для аналитики и машинного обучения. В частности, выполняется валидация и верификация потребности в качественных данных и предполагаемого использования данных для управления такими характеристиками качества, как согласованность и релевантность.

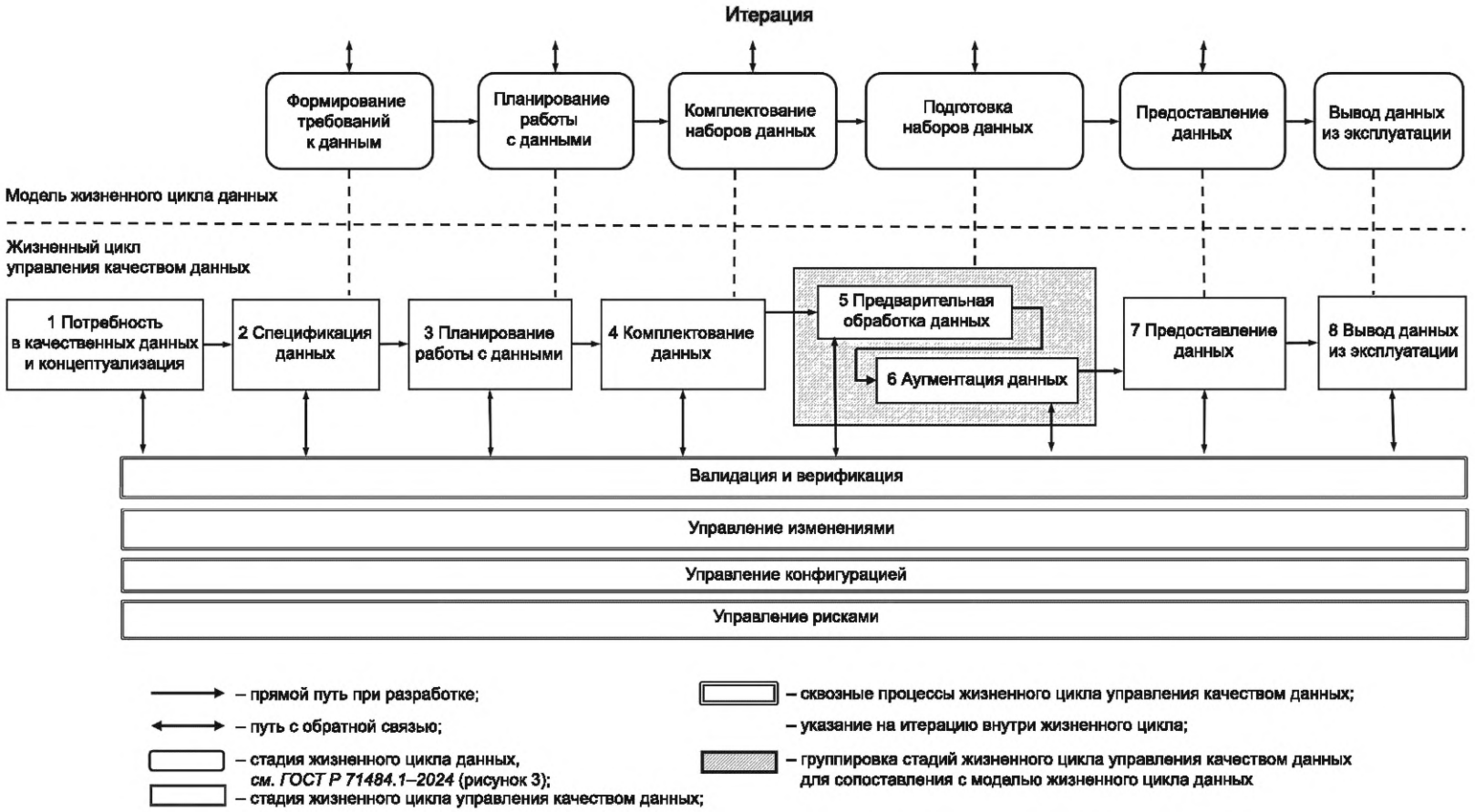

Рисунок 1 — Жизненный цикл управления качеством данных

7.2.2.2 Спецификация данных

На стадии спецификации данных формируются требования к данным, в том числе требования к форматам данных, статистическим свойствам и разделимости. Рекомендации по управлению качеством данных облегчают выявление ошибочных, неполных или противоречивых требований и планов. Например, с учетом контекста аналитики и машинного обучения процесс управления качеством подтверждает, что данные соответствуют требованиям этого контекста.

7.2.2.3 Планирование работы с данными

На стадии планирования работы с данными разрабатывается план, соответствующий спецификации данных. Он включает планирование конкретных мероприятий и ресурсов для сбора и обработки данных на протяжении всего жизненного цикла данных, а также методов оценки и критериев приемлемости.

7.2.2.4 Комплектование данных

С точки зрения управления качеством данных к стадии комплектования данных может относиться, в том числе, генерация данных или получение и объединение существующих наборов данных. Выбор метода сбора данных влияет на управление качеством данных. Независимо от метода получения данных результат этой стадии считается необработанным входом для последующих стадий жизненного цикла данных, даже если импортируется уже обрабатывавшийся ранее набор данных.

7.2.2.5 Предварительная обработка данных

На стадии предварительной обработки данных формулируются все требования и рекомендации по управлению качеством данных, связанные с обработкой исходных данных, полученных на стадии комплектования данных. Основные действия на этой стадии, как правило, связаны с удалением элементов данных, например, путем фильтрации и очистки этих данных.

7.2.2.6 Аугментация данных

Аугментация данных использует входные данные, полученные на стадии предварительной обработки данных, и упорядочивает их в соответствии со всеми требованиями и рекомендациями управления качеством, связанными с добавлениями в набор данных. Это включает сопоставление метаданных, добавление меток данных или другие преобразования.

7.2.2.7 Предоставление данных

Целью предоставления данных является предоставление набора данных туда, где это необходимо. Требования к управлению качеством данных на этой стадии включают, в том числе, поддержание целостности наборов данных во время доставки, контроль доступа, документирование и проверку того, что контекст использования соответствует предполагаемому использованию набора данных.

7.2.2.8 Вывод данных из эксплуатации

Управление качеством данных завершается удалением или передачей данных другой стороне на стадии вывода данных из эксплуатации. В каждом из этих взаимоисключающих случаев ответственность за данные снимается. Надлежащее обращение с выводимыми из эксплуатации данными является основным фактором доверия к организации и может способствовать предотвращению их неправомерного использования.

7.2.3 Адаптация жизненного цикла управления качеством данных вне зависимости от типа проекта

Организации следует адаптировать жизненный цикл, если она:

a) объединяет или разделяет стадии;

b) выполняет требуемые процессы на различных или дополнительных стадиях (см. *ГОСТ Р 71484.4*);

c) повторяет стадии;

d) выполняет не зависящие друг от друга действия одновременно;

e) обоснованно исключает неприменимые стадии;

f) применяет рекомендации по управлению качеством данных для валидации и верификации, управления изменениями и управления конфигурацией, как показано на рисунке 1.

7.2.4 Сквозные аспекты жизненного цикла управления качеством данных

7.2.4.1 Общие положения

На рисунке 1 показаны четыре процесса, которые охватывают все стадии жизненного цикла управления качеством данных, включая:

a) валидацию и верификацию;

b) управление изменениями;

c) управление конфигурацией;

d) управление рисками.

Эти процессы должны выполняться вне зависимости от типа проекта.

7.2.4.2 Валидация и верификация

Валидация и верификация качества данных должны осуществляться на всех релевантных стадиях жизненного цикла управления качеством данных.

7.2.4.3 Управление изменениями

Управление изменениями гарантирует, что любое изменение данных или процессов не окажут негативного влияния на качество данных.

7.2.4.4 Управление конфигурацией

Управление конфигурацией помогает гарантировать, что на данные и процессы можно однозначно ссылаться и что никакие изменения в управлении конфигурацией не окажут негативного влияния на качество данных.

7.2.4.5 Управление рисками

Управление рисками должно осуществляться на всех релевантных стадиях жизненного цикла управления качеством данных.

7.3 Требования и рекомендации

7.3.1 Потребность в качественных данных и концептуализация

7.3.1.1 Общие положения

Организация должна:

a) документировать мотивацию и истоки потребности в данных;

b) указывать состав, назначение и предполагаемое использование данных;

c) документировать требования к качеству данных.

Примечание — Дополнительную информацию об этических и общественных аспектах см. в *[3]*.

7.3.1.2 Анализ заинтересованных сторон

Организация должна:

a) провести анализ для выявления всех заинтересованных сторон, имеющих отношение к данным, и последствий для управления качеством данных;

b) выявить потенциально противоречивые требования к качеству данных.

Примечание — Описание потенциальных ролей и подролей заинтересованных сторон в области ИИ представлено в *ГОСТ Р 71476*.

7.3.1.3 Технико-экономическое обоснование

Организация должна проводить технико-экономическое обоснование, основанное на имеющихся у нее опыте и ресурсах и в котором оценивается способность организации достичь своих целей в области качества данных. Анализ реализуемости такого типа должен обновляться на протяжении всего жизненного цикла данных.

Примечание — Предыдущие проекты МО и общедоступные наборы данных, которые достигли целевых показателей качества, являются примерами доказательств, подтверждающих реализуемость.

7.3.2 Спецификация данных

7.3.2.1 Общие положения

Организация должна указать требования к данным в спецификации данных и подтвердить, что эти требования являются согласованными и полными для предполагаемого использования.

Спецификация данных должна включать информацию о том, какие аспекты являются минимально необходимыми, а какие необязательными в зависимости от предполагаемого использования системой ИИ.

Спецификация должна содержать:

a) описание природы данных;

b) предполагаемую цель проекта аналитики и машинного обучения, использующего данные;

c) требования законодательства;

d) требования безопасности и защищенности;

e) потенциальную возможность нежелательной предвзятости;

f) требования к конфиденциальности;

g) требования к данным, относящимся к предметной области и конкретному проекту;

h) модель качества данных и требуемый уровень качества для достижения цели использования данных.

Примечания

1 Спецификация данных может быть основана на описании обработки нежелательной систематической ошибки на протяжении жизненного цикла системы ИИ — см. *[5]*.

2 Спецификация данных включает в себя различные взаимосвязи с проектом МО (например, обучающие данные, тестовые данные или текущее управление данными).

3 Некоторые алгоритмы МО требуют определенных статистических свойств обучающих данных.

4 Соображения конфиденциальности могут включать принцип минимизации данных (также известный как избегание излишних данных). Минимизация данных означает, что собирается или используется только минимум данных, необходимый для функционирования системы.

7.3.2.2 Формат данных

Организация должна определить, какая информация должна быть включена в сведения о формате данных. Такая информация может включать, помимо прочего:

a) кодировку;

b) частоту (время);

c) разрешение (пространство);

d) синтаксис;

e) семантику;

f) структуру связанных метаданных;

g) диапазоны ожидаемых значений;

h) обязательные и необязательные элементы;

i) опционально допустимые форматы свойств и допустимые альтернативы (например, размеры изображений);

j) ссылки и перекрестные ссылки.

Примечание — Семантика относится к интерпретации и использованию данных, включая допустимые операции с данными.

7.3.2.3 Статистические свойства и разделимость

Спецификация данных должна включать в себя описание допустимых разделений на подмножества (например, обучающие, валидационные и тестовые данные). Подмножества не должны пересекаться, за исключением случаев, когда они используются в сочетании с научно обоснованными методами, такими как бэггинг (bagging) — технология классификации, использующая композиции алгоритмов), кросс-валидация и обобщенная кросс-валидация (bootstrapping).

Организация должна определять и поддерживать для каждого из этих подмножеств:

a) соответствующие статистические свойства;

b) репрезентативность;

c) соответствующий размер.

7.3.2.4 Вспомогательные ресурсы и инструменты

Спецификация данных должна включать все соответствующие требования по обработке данных и, как минимум, охватывать:

a) спецификации вспомогательных инструментов;

b) минимальные технические требования к аппаратным средствам для обработки данных и работы с инструментами;

c) требуемое разрешение при визуализации;

d) требования к хранилищу;

e) требования к сети;

f) требования к доступности.

7.3.3 Планирование работы с данными

7.3.3.1 Общие положения

Организация должна разработать план управления качеством данных, охватывающий процессы, действия и ресурсы для соответствия спецификации данных на протяжении всего жизненного цикла данных.

План управления качеством данных должен включать:

a) цели в области управления качеством данных в соответствии со спецификацией данных;

b) гарантированное выделение ресурсов;

c) положения о соответствии требованиям законодательства;

d) процессы контроля и обновления плана управления качеством данных (при необходимости);

e) положения, гарантирующие прослеживаемость и воспроизводимость процессов.

7.3.3.2 План управления качеством данных, соответствующий стадиям жизненного цикла данных

План управления качеством данных должен включать спецификации процессов, включая их описания, связанные с ними действия, предполагаемые результаты с назначением ответственного лица за каждый процесс, и охватывать следующие этапы жизненного цикла данных:

a) план комплектования данных;

b) план предварительной обработки данных;

c) план аугментации данных;

d) план предоставления данных;

e) план вывода данных из эксплуатации.

План управления качеством данных, соответствующий стадиям жизненного цикла данных, должен соответствовать требованиям, указанным в 7.3.4—7.3.8.

7.3.4 Комплектование данных

7.3.4.1 Общие положения

Организация должна определить процессы сбора данных в соответствии со спецификацией и включить их в план управления качеством данных.

Соответствующие требования спецификации данных (см. 7.3.2) должны быть выполнены до предварительной обработки данных. Риски, возникающие из-за неполноты данных, должны быть документированы и смягчены. Организация должна контролировать процессы комплектования данных и документировать отклонения от ожидаемых результатов или запланированных процессов.

Организация, получающая данные, должна взять на себя ответственность за данные, включая управление их качеством.

Примечание — Процессы комплектования данных могут применяться независимо от источника данных.

7.3.4.2 Источник данных

При выборе метода комплектования данных организация должна учитывать, что:

a) необходимые данные уже существуют и напрямую доступны для повторного использования;

b) существующие данные могут быть преобразованы в соответствии с требованиями;

c) данные можно приобрести или лицензировать;

d) необходимо собрать новые данные;

e) данные могут быть сгенерированы (например, с помощью имитации или других вычислительных инструментов).

Организация может делегировать комплектование данных, но при этом несет ответственность за соблюдение требований к качеству данных. Должны быть приняты во внимание задокументированные условия повторного использования или лицензирования набора данных.

Примечание — Доступные открытые данные являются хорошим вариантом для растущего числа приложений и способствуют повторному использованию данных. Существующие наборы данных, уже принадлежащие организации, также могут быть рассмотрены для повторного использования.

7.3.4.3 Сбор данных

В случае если организации необходимо собрать новые данные, она должна предоставить:

a) причины сбора новых данных;

b) метод сбора данных (например, с помощью датчиков, ручного ввода, преобразования значений, имитационного моделирования, синтетических данных);

c) спецификацию сбора данных, включая:

1) соответствующие настройки и параметры методов сбора данных;

2) условия эксплуатации;

3) обнаружение ошибок и их устранение;

4) необходимые навыки и ресурсы;

5) если применимо, технические характеристики и местоположение установки датчиков;

d) необходимые ресурсы и перечень навыков для сбора и обработки данных;

e) методы обезличивания или псевдонимизации, если применимо;

f) документацию о шкалах измерения данных (например, номинальных, порядковых, интервальных, отношений) и единицах измерения;

g) ожидаемые отклонения от целевого качества данных.

Примечание — Метод сбора данных может привести к систематическим ошибкам в собранных данных, особенно если порядок действий для сбора данных отличается от предполагаемого. Одним из возможных способов смягчения последствий является использование разных методов сбора данных для обеспечения избыточности.

7.3.4.4 Обработка данных

Организация должна разработать соответствующие процессы и инструменты для управления полученными данными, их визуализации и анализа. Полученные данные должны соответствовать спецификации данных.

Организация должна определить процессы управления полученными данными, включая:

a) проверку качества;

b) контроль доступа;

c) отслеживание происхождения данных и их модификаций;

d) контроль версий набора данных;

e) хранение, размещение и резервное копирование данных;

f) операции с наборами данных, такие как обновление, объединение, сортировка и разделение на фрагменты;

g) процессы обнаружения, минимизации и смягчения последствий повреждения данных;

h) назначение ответственного за данные.

7.3.5 Предварительная обработка данных

7.3.5.1 Общие положения

Организация должна определить процессы предварительной обработки данных, соответствующие спецификации данных и включить их в план управления качеством данных.

7.3.5.2 Очистка данных

Организация должна, как минимум, указать, какие показатели качества данных были применены для:

a) обнаружения и решения проблемы недостающих данных;

b) обнаружения и решения проблемы дублирования данных;

c) обнаружения и устранения выбросов и других проблем;

d) обнаружения и решения проблем смещенности, дрейфа и масштабирования;

e) обнаружения и обработки нестандартных значений;

f) обнаружения и удаления ненужных данных;

g) преобразования данных, включая нормализацию данных;

h) методов проверки результатов очистки данных;

i) обезличивания данных (если применимо).

Примечание — Некоторые преобразования данных необратимы.

7.3.6 Аугментация данных

7.3.6.1 Общие положения

Организация должна определить процессы аугментации данных, чтобы они соответствовали спецификации данных, и включить их в план управления качеством данных.

7.3.6.2 Разметка и аннотирование данных

Организация должна определить порядок разметки данных, включая:

a) спецификацию разметки данных;

b) необходимые навыки и ресурсы;

c) процедуру отбора данных для разметки;

d) мониторинг и управление качеством процессов разметки;

e) потенциальное физическое и психологическое воздействие на разметчика данных, включая стратегии смягчения последствий.

7.3.6.3 Компьютерная аугментация

Если используется компьютерная аугментация, организация должна указать:

a) инструменты и методы, используемые для компьютерной аугментации;

b) выбранные признаки;

c) сформированные журналы и метаданные.

7.3.7 Предоставление данных

7.3.7.1 Общие положения

Организация должна:

a) определить процессы предоставления данных в соответствии со спецификацией данных и включить их в план управления качеством данных;

b) предъявить проверяемые доказательства того, что предоставленные данные и метаданные соответствуют всем указанным требованиям;

c) внедрить соответствующие процессы для обеспечения предоставления элементов и версий, описанных в 7.3.7.2;

d) убедиться, что процесс предоставления не приводит к изменению данных;

e) внедрить соответствующие процессы обезличивания и доступа к данным;

f) осуществить контроль версий всех предоставленных элементов;

g) упаковать и доставить данные, включая документацию и другие сопутствующие элементы, чтобы пользователь данных мог:

1) определить, соответствуют ли данные требованиям к использованию;

2) использовать данные в соответствии с предполагаемым назначением.

Примечание — Например, представление данных с большим количеством десятичных знаков может вводить в заблуждение, предполагая более высокую, чем достигнутая, точность и, следовательно, может привести к несоответствующему использованию или допущению об использовании.

7.3.7.2 Элементы предоставления данных

В описание предоставления данных должны быть включены следующие пункты:

a) описание данных, включая любые их разделения на части, если это необходимо для обучения, тестирования и проверки;

b) документация (см. 7.3.7.3);

c) образцы данных, представляющие их синтаксис и формат;

d) документация спецификации данных, других связанных элементов (см. 7.3.7.1 ж), модели качества данных, показателей качества данных и результатов оценки качества данных.

7.3.7.3 Документация

Вместе с данными должна быть представлена документация, содержащая следующую информацию:

a) область использования;

b) инструкция по применению;

c) статистические характеристики данных;

d) описание всех элементов данных и метаданных;

e) перечень заинтересованных сторон;

f) роли и обязанности получателя, включая законодательные требования и договорные обязательства;

g) требования к условиям использования; обычно это охватывает среды тестирования, разработки и эксплуатации системы [см. стадию 8 в *ГОСТ Р 71484.1—2024*, (рисунок 3)];

h) факторы, которые могут негативно повлиять на показатели качества;

i) доступные методы приемочных испытаний и их ожидаемые результаты;

j) стратегия взаимодействия с пользователем данных, включая то, как происходят обновления и как уведомляются пользователи.

7.3.7.4 Отслеживание и улучшение

Организация должна оценить, существует ли требование отслеживать процесс и результаты использования данных. В случае, если такие требования существуют или организация решает внедрить отслеживание, организация должна внедрить соответствующие процессы для удовлетворения требований к отслеживанию и улучшению.

В случае внедрения отслеживания использования данных организация должна оценить необходимость:

a) обновления документации, касающейся области использования и предполагаемого использования данных;

b) улучшения данных;

c) обновления системы управления качеством данных.

Организация должна создать канал для сбора добровольных отзывов о данных от соответствующих заинтересованных сторон и разработать процессы оценки обратной связи для постоянного улучшения данных и системы управления качеством данных.

7.3.7.5 Оптимизация данных

Организация должна документировать применяемую оптимизацию данных:

a) оптимизация места в хранилище;

b) оптимизация доступа;

c) оптимизация требований к сетевым ресурсам;

d) оптимизация обработки данных (например, буферизация, кэширование).

7.3.7.6 Обязательства по поддержке

Уровень технической поддержки, предоставляемой соответствующим заинтересованным сторонам, должен быть достаточным для выполнения требований к качеству данных в контексте предполагаемого использования данных. Должны быть задокументированы следующие пункты:

a) необходимые навыки или обучение использованию данных;

b) документирование предоставленной поддержки, включая объем, период и доступность;

c) план обслуживания и стремление к постоянному улучшению данных;

d) наличие канала связи с соответствующими заинтересованными сторонами.

7.3.8 Вывод данных из эксплуатации

7.3.8.1 Общие положения

Организация должна выбрать подходящий метод вывода данных из эксплуатации либо путем удаления данных, либо путем передачи данных другой стороне.

Организация должна определить процессы вывода данных из эксплуатации, отвечающие всем необходимым требованиям (см. 7.3.8.2—7.3.8.4), и включить их в план управления качеством данных.

7.3.8.2 Передача данных

В случае передачи данных организация должна определить получателя данных, который имеет право принимать данные, и выполнить все связанные с этим законодательные требования.

Организация должна документально подтвердить, что передача данных не нарушает никаких обязательств.

Организация должна документально подтвердить явно выраженное согласие получателя принять данные и выполнить все связанные с этим обязательства.

Организация должна сформировать и обеспечить независимое подтверждение отчета о передаче данных.

7.3.8.3 Удаление данных

Организация должна документально подтвердить, что удаление данных не нарушает обязательств, включая, но не ограничиваясь этим, законодательные требования и обязательства по хранению данных.

Организация должна гарантировать, что ни один пользователь данных не будет нуждаться в доступе к данным. Пользователями данных могут быть системы, которым требуется доступ к данным для обслуживания и повторного обучения.

В случае удаления организация должна удалить данные из всех мест хранения и задокументировать методы безопасного удаления данных (такие как перезапись диска случайными данными).

Организация должна проверить, существует ли возможность частичного или полного восстановления данных из моделей МО, которые были обучены с использованием данных. Организация должна учитывать такие случаи в своих обязательствах и в плане вывода данных из эксплуатации.

Организация должна сформировать и обеспечить независимое подтверждение отчета об удалении данных.

Организация должна проверить, имеют ли данные культурное, историческое или социальное значение. Если данные имеют важное значение, организация должна документально подтвердить действия по передаче данных получателю, который может выполнить все обязательства по предоставлению таких данных.

Организация должна рассмотреть возможность и целесообразность передачи данных в общественное пользование.

Примечания

1 Места хранения данных включают резервные копии, которые могут автоматически создаваться ИТ-инфраструктурой.

2 Организации рекомендуется предоставить аргументы, почему передача данных вредит ее интересам, чтобы обосновать решение организации удалить данные.

3 Передача данных в общественное пользование потенциально может пойти на благо исследованиям и образованию, а также косвенно самой организации.

7.3.8.4 Частичное удаление данных по запросу

Организация должна владеть соответствующими процессами обработки запросов на частичное удаление данных, если это применимо.

В случае удаления части данных организация должна оценить, соответствуют ли оставшиеся данные своим спецификациям. В случае, если данные не соответствуют спецификации, организация

должна уведомить соответствующие заинтересованные стороны о рисках и проблемах с качеством этих данных и, если возможно, обеспечить соответствующие меры по их смягчению. В случае, если соответствующие меры по смягчению рисков не могут быть реализованы, организация должна вывести из эксплуатации все данные, что может включать передачу оставшихся данных.

В случае частичного удаления данных организация должна удалить эту часть данных из всех мест хранения. Организация должна сформировать и обеспечить независимое подтверждение отчета о частичном удалении данных.

Примечание — Нормативные правовые акты в отношении данных, такие как законы о персональных данных, могут обязать организацию удалить некоторые части набора данных.

7.4 Результаты

7.4.1 Результаты стадии определения потребности в качественных данных и концептуализации

Результатами реализации этой стадии жизненного цикла управления качеством данных являются:

a) документация о предполагаемом использовании данных;

b) технико-экономическое обоснование;

c) анализ потребностей заинтересованных сторон;

d) свидетельство того, что верификация и валидация результатов, полученных на стадии «Потребность в качественных данных и концептуализация», были успешно завершены.

7.4.2 Результаты стадии спецификации данных

Результатами реализации этой стадии жизненного цикла управления качеством данных являются:

a) формат данных и спецификацию их использования;

b) спецификация вспомогательных инструментов и связанных с ними требований;

c) свидетельство того, что верификация и валидация результатов, полученных на стадии спецификации данных, были успешно завершены.

7.4.3 Результаты стадии планирования работы с данными

Результатами реализации этой стадии жизненного цикла управления качеством данных являются:

a) план управления качеством данных;

b) свидетельство того, что верификация и валидация результатов, полученных на стадии планирования работы с данными, были успешно завершены.

7.4.4 Результаты стадии комплектования данных

Результатами реализации этой стадии жизненного цикла управления качеством данных являются:

a) собранные данные и связанная с ними документация;

b) спецификация методов комплектования данных и соответствующих конфигураций;

c) инфраструктура для управления данными;

d) анализ качества данных;

e) свидетельство того, что верификация и валидация результатов, полученных на стадии комплектования данных, были успешно завершены.

7.4.5 Результаты стадии предварительной обработки данных

Результатами реализации этой стадии жизненного цикла управления качеством данных являются:

a) очищенные данные;

b) анализ качества данных;

c) статистические характеристики, описывающие свойства данных;

d) свидетельство того, что верификация и валидация результатов, полученных на стадии предварительной обработки данных, были успешно завершены.

7.4.6 Результаты стадии аугментации данных

Результатами реализации этой стадии жизненного цикла управления качеством данных являются:

a) аугментированные данные, включая метаданные и метки;

b) описания характеристик;

c) спецификации применяемых инструментов и методов аугментации;

d) свидетельство того, что верификация и валидация результатов, полученных на стадии аугментации данных, были успешно завершены.

7.4.7 Результаты стадии предоставления данных

Результатами реализации этой стадии жизненного цикла управления качеством данных являются:

a) предоставленные элементы, в том числе данные, образцы данных и документация;

b) свидетельство того, что верификация и валидация результатов, полученных на стадии предоставления данных, были успешно завершены.

7.4.8 Результаты стадии вывода данных из эксплуатации

Результатами реализации этой стадии жизненного цикла управления качеством данных являются:

a) отчет о выводе данных из эксплуатации;

b) свидетельство того, что верификация и валидация результатов, полученных на стадии вывода данных из эксплуатации, были успешно завершены.

8 Горизонтальные процессы

8.1 Цель

Целью горизонтальных процессов управления качеством данных является консолидация действий, применимых на каждой стадии жизненного цикла управления качеством данных.

8.2 Общие положения

Организация должна осуществлять и документировать горизонтальные процессы с входными и выходными данными, относящимися к конкретному процессу.

8.3 Требования и рекомендации

8.3.1 Верификация и валидация

8.3.1.1 Общие положения

Процессы верификации и валидации служат контролю и оценке качества данных на всех стадиях жизненного цикла управления качеством данных. Верификация позволяет получить объективные свидетельства того, что требования к качеству данных были выполнены. Валидация позволяет установить, что данные соответствуют поставленным целям.

Особенности и ограничения процессов верификации и валидации должны быть задокументированы, а потенциальное влияние на качество данных должно быть оценено.

8.3.1.2 Разброс параметров качества жизненного цикла данных

На каждой стадии жизненного цикла качества данных должны быть определены объективно оцениваемые целевые показатели качества полученных результатов. Верификация и валидация должны подтвердить, что все требования каждой стадии жизненного цикла качества данных были выполнены. Переход к стадии жизненного цикла данных должен осуществляться только при наличии верифицированных и валидированных входных данных. Результаты верификации и валидации должны быть задокументированы.

8.3.1.3 Улучшение

После успешного прохождения верификации и валидации результатов имеющиеся отзывы о них в ходе предполагаемого использования должны быть задокументированы и оценены. Несоответствие полученного результата требованиям о предполагаемом использовании может повлечь за собой обновление процессов верификации и валидации.

Результаты верификации и валидации должны использоваться для регулярного обновления и улучшения процессов управления качеством данных.

8.3.2 Управление конфигурацией

Управление конфигурацией должно планироваться и поддерживаться на протяжении всего жизненного цикла управления качеством данных.

Результаты, определенные планом управления качеством данных, должны основываться на стратегии управления конфигурацией, которая определяет требования и цели для однозначно идентифицируемых и воспроизводимых элементов.

Управление конфигурацией должно установить требования, применимые к качеству данных для каждой допустимой конфигурации.

В процессе управления конфигурацией должно быть установлено, что все допустимые конфигурации соответствуют всем применимым требованиям к качеству данных.

8.3.3 Управление изменениями

8.3.3.1 Общие положения

Ожидается, что изменения требований к качеству данных и соответствующих процессов будут часто происходить во время разработки систем ИИ и поддерживаться в жизненном цикле управления

качеством данных итеративно. Данные и связанные с ними метаданные часто подвергаются изменениям. В основном это происходит на стадии спецификации и внедрения и в меньшей степени — на стадиях валидации и верификации. Изменения в данных также могут произойти при повторении процесса комплектования данных в результате действий по предоставлению данных.

8.3.3.2 Планирование управления изменениями

Управление изменениями должно:

a) планироваться и предшествовать любым изменениям;

b) включать план, в котором указаны элементы, подлежащие изменению, и график изменений;

c) определять процесс, который включает в себя:

1) спецификацию запроса на изменение (см. 8.3.3.3);

2) анализ запроса на изменение (см. 8.3.3.4);

3) оценку запроса на изменение (см. 8.3.3.5);

4) документацию (см. 8.3.3.6).

8.3.3.3 Спецификация запроса на изменение

Запросы на изменение должны содержать:

a) уникальный идентификатор;

b) дату;

c) причину, описание и конфигурацию запрошенного изменения;

d) решение и обоснование запрошенного изменения.

8.3.3.4 Анализ запроса на изменение

Воздействие запроса на изменение должно быть проанализировано с учетом:

a) типа (ошибка, адаптация, дополнение к данным или меткам и т. д.);

b) затрагиваемых элементов (данные, метки, требования и т. д.);

c) затрагиваемых сторон, включая их ответственность;

d) влияния на качество;

e) влияния на ранее существовавшие данные или метки;

f) графика.

8.3.3.5 Оценка запроса на изменение

Запрос на изменение и анализ его воздействия должны быть оценены для принятия решений уполномоченным лицом:

a) о статусе (принято, отклонено и т. п.);

b) персонале, который будет осуществлять изменения;

c) о порядке действий.

8.3.3.6 Внедрение и документирование изменений

Изменения должны быть выполнены и проверены в соответствии с планом.

Изменение данных, их качества, способа использования или верификации приводят к обновлению связанных с ними оценок. Документация, касающаяся изменений, должна включать:

a) список измененных элементов, включая конфигурацию и версии;

b) подробности выполненных изменений;

c) дату вступления изменений в силу.

8.3.4 Управление рисками

8.3.4.1 Общие положения

Целью управления рисками является обеспечение того, чтобы все риски для качества данных были учтены, оценены и при необходимости смягчены либо устранены.

8.3.4.2 Требования и рекомендации

Организация должна проводить оценку рисков и вести документацию в отношении рисков, связанных с качеством данных. При оценке рисков следует учитывать разумно предполагаемое использование данных не по назначению и связанные с этим последствия. Организация должна управлять и поддерживать соответствующие процессы для минимизации выявленных рисков.

Риски предполагаемого использования данных должны быть определены, обоснованы и задокументированы как часть действий, предпринимаемых организацией для их оценки и устранения связанных с ними рисков. Состав соответствующих показателей качества данных должен обновляться и управляться системой управления качеством данных с учетом связанных с этим рисков. Организация должна сообщать о рисках, связанных с данными, соответствующим заинтересованным сторонам.

Примечание — Для определения процессов управления рисками при использовании ИИ организации могут применять *[6]*.

8.4 Результаты

8.4.1 Результаты верификации и валидации

Результаты верификации и валидации включают:

a) документацию о результатах верификации и валидации на всех стадиях жизненного цикла;
b) документацию об ограничениях и недочетах верификации и валидации;
c) предложения по улучшению.

8.4.2 Результаты управления конфигурацией

Результаты управления конфигурацией включают:

a) документированный план управления конфигурацией;
b) документированную стратегию конфигурации;
c) документацию рабочих конфигураций и связанных с ними требований к качеству.

8.4.3 Результаты управления изменениями

Результаты управления изменениями включают:

a) план управления изменениями;
b) запросы на изменения;
c) анализ воздействия изменений;
d) отчеты об изменениях.

8.4.4 Результаты управления рисками

Результаты оценки риска включают:

a) документацию о рисках, связанных с качеством данных;
b) план обработки рисков.

9 Управление качеством данных в цепочках поставок

9.1 Цель

Целью требований к управлению цепочкой поставок является обеспечение того, чтобы все выбранные поставщики предоставляли данные, соответствующие требованиям организации.

9.2 Требования и рекомендации

При выборе поставщика следует учитывать возможности поставщика предоставлять качественные данные в соответствии с разделами 6—8 настоящего стандарта.

Требования к запросу ценового предложения включают:

a) запрос ценового предложения, включающий:
   1) требования к соблюдению настоящего стандарта;
   2) соответствующие спецификации данных;
   3) соответствующие требования к качеству.
b) соглашение о взаимодействии при разработке между заказчиком и поставщиком с указанием:
   1) поставщиков, заказчиков и менеджеров по качеству;
   2) всех действий в жизненном цикле управления качеством данных, которые должны выполняться каждой стороной;
   3) совместно используемой информации и результатов;
   4) обязанностей, возложенных на каждую сторону по каждому виду деятельности;
   5) требований к качественным и количественным бенчмаркам (см. 6.3.10, перечисление b);
   6) процессов, методов и инструментов, связанных с соглашением;
   7) мероприятий по оценке качества;
   8) планируемого отчета поставщика об оценке качества;
   9) положения, которое позволяет назначенному заказчиком аудитору получить доступ ко всем необходимым ресурсам для проведения аудита качества;
   10) обязанностей и действий в ходе эксплуатации и при выводе из эксплуатации [см. стадии 8—10 в *ГОСТ Р 71484.1—2024*, (рисунок 3)];
   11) требований об информировании по проблемам, которые влияют на качество данных или создают риски невыполнения соглашения;
   12) требований к продолжительности периода эксплуатации данных для обеспечения их доступности и использования.

9.3 Результаты

Результаты управления качеством данных в цепочках поставок включают:

a) отчет о выборе поставщика;

b) соглашение о взаимодействии при разработке;

c) отчет об оценке качества.

10 Управление инструментами для обработки данных

10.1 Цель

Целью управления инструментами для обработки данных является обеспечение соответствия данных требованиям организаций при каждом использовании того или иного инструмента для обработки данных или его применении к данным на протяжении всего жизненного цикла данных.

10.2 Требования и рекомендации

Организация должна определить влияние на качество данных как минимум для следующих категорий инструментов:

a) для комплектования данных (например, веб-скраперы, утилиты для захвата экрана, анкетные формы, датчики);

b) для предварительной обработки данных (например, инструменты для очистки или преобразования данных);

c) для оценки качества данных (например, инструменты измерения, инструменты анализа измерений);

d) для разметки данных;

e) для хранения данных (например, инструменты для создания, чтения, обновления и удаления);

f) для передачи данных (например, сети, шины данных);

g) для защиты данных (например, межсетевые экраны, инструменты для шифрования).

10.3 Результаты

К результатам управления инструментами для обработки данных относятся:

a) документация обо всех использованных инструментах для обработки данных;

b) документация обо всех воздействиях инструментов для обработки данных на качество данных.

11 Управление зависимостями, связанными с качеством данных

11.1 Цель

Целью управления зависимостями, связанными с качеством данных, является обеспечение того, чтобы они не нарушали требований организации к данным. Зависимости могут быть внутренними (например, обработка данных, выполняемая под непосредственным контролем организации) или внешними (например, облачные сервисы, разметка данных по договору).

11.2 Требования и рекомендации

Организация должна определить как минимум следующее:

a) внутренние зависимости, которые могут повлиять на качество данных;

b) внешние зависимости, которые могут повлиять на качество данных;

c) меры по смягчению последствий, которые организация может предпринять для устранения влияния любой зависимости на качество данных.

11.3 Результаты

Результаты управления зависимостями, связанными с качеством данных, включают:

a) документацию всех внутренних зависимостей и их потенциального влияния на качество данных;

b) документацию всех внешних зависимостей и их потенциального влияния на качество данных;

c) документирование мер по смягчению рисков, предпринятых для обеспечения соответствия данных требованиям организации.

12 Управление качеством данных в проекте

12.1 Цель

Целью управления качеством данных конкретного проекта для организации является обеспечение того, чтобы требования к качеству данных конкретного проекта принимались во внимание достаточным и надлежащим образом.

12.2 Требования и рекомендации

12.2.1 Контекст и предполагаемое использование

Контекст требований и рекомендаций по управлению качеством данных — это проект организации в области аналитики и машинного обучения для конкретной системы ИИ. Проект в области аналитики и машинного обучения может быть связан с несколькими наборами данных, алгоритмами или моделями машинного обучения. Требования и рекомендации по управлению качеством данных для конкретного проекта относятся к компонентам системы ИИ, зависящим от качества данных.

12.2.2 Цель

Цель состоит в том, чтобы гарантировать, что данные, используемые в конкретном проекте аналитики и машинного обучения, соответствовали требованиям и рекомендациям.

12.2.3 Требования и рекомендации

Организация должна:

a) обозначить сферу действия проекта;

b) документировать потребность в использовании данных;

c) документировать условия для использования данных;

d) документировать предполагаемое использование данных;

e) документировать известные ограничения из-за выявленных несовместимостей или конфликтов, связанных с:

1) периодом технического обслуживания и продолжительностью использования;
2) техническими ограничениями (например, алгоритмами, библиотеками);
3) областью применения;
4) ошибочным использованием данных.

Список известных ограничений должен обновляться и представляться, когда станет известно о новых ограничениях.

Модель МО обычно взаимодействует с несколькими компонентами, например, пользовательскими интерфейсами, базами данных и др. Если они влияют на качество данных, их следует учитывать в системе управления качеством данных.

Там, где это возможно, организация должна документировать приложения, в которых данные были успешно использованы.

12.3 Спецификация и управление требованиями к качеству данных

12.3.1 Цель

Целью спецификации и управления требованиями к качеству данных является обеспечение соблюдения таких необходимых критериев, как прослеживаемость и воспроизводимость.

12.3.2 Требования и рекомендации

Организация должна предоставить документацию, относящуюся к требованиям и рекомендациям, связанным с качеством данных. Организация должна предоставить рекомендации по формулированию требований к качеству данных, например, на естественном языке, с использованием онтологии или математических формул. Такие требования должны быть сформулированы недвусмысленно и в строгой форме, чтобы обеспечить их проверяемость.

Требования к качеству данных должны:

a) быть идентифицируемы как таковые;

b) применяться к соответствующим данным;

c) являться частью управления конфигурацией;

d) быть проверяемыми.

Требования к качеству данных должны иметь уникальный идентификатор, статус и прослеживаться относительно данных. Проверка может осуществляться различными способами и как минимум охватывать:

- проверку экспертами в предметной области;
- проверку привлеченными разработчиками систем ИИ;
- формальную проверку (если применимо).

12.4 Роли и ответственность в управлении качеством данных

12.4.1 Цель

Цель распределения ролей и ответственности при управлении качеством данных состоит в том, чтобы обеспечить выполнение соответствующих функций по обеспечению качества данных и назначение ответственных за каждую роль. Персонал, назначенный на определенные роли, должен иметь профессиональную квалификацию и располагать необходимыми ресурсами для выполнения своих обязанностей.

12.4.2 Требования и рекомендации

Организация должна назначить:

a) менеджера проекта в начале проекта, обладающего необходимыми полномочиями для обеспечения:

- выполнения мероприятий по обеспечению качества;
- достижения соответствия настоящему стандарту;
- назначения менеджера по качеству;

b) менеджера по качеству.

12.4.3 Результаты

К результатам распределения ролей и ответственности при управлении качеством данных относятся:

a) документ, определяющий роли в управлении качеством данных (например, менеджер проекта, менеджер по качеству);

b) документ, определяющий обязанности для каждой роли в управлении качеством данных.

12.5 Адаптация мероприятий по обеспечению качества данных

Мероприятия по обеспечению качества, касающиеся жизненного цикла данных и используемые в настоящем стандарте, могут быть адаптированы при условии, что они определены в плане управления качеством данных и для этого приведено обоснование.

12.6 Планирование и координация деятельности по обеспечению качества данных

12.6.1 Общие положения

Мероприятия по управлению качеством данных должны включать:

a) планирование и координацию мероприятий по обеспечению качества;

b) формирование плана управления качеством данных;

c) мониторинг мероприятий по обеспечению качества в соответствии с планом управления качеством данных;

d) распределение и информирование об ответственности за мероприятия по обеспечению качества.

12.6.2 План управления качеством данных

Все участвующие организации обязаны поддерживать план управления качеством данных, который должен:

a) быть явным образом включен в план проекта (или на него должны быть приведены ссылки) с выделенными мероприятиями по управлению качеством данных;

b) содержать мероприятия по достижению качества данных;

c) регулярно пересматриваться и обновляться, включая все соответствующие результаты.

12.6.3 Планирование процессов

Планирование мероприятий по управлению качеством данных должно включать:

a) показатели качества данных;

b) зависимости от других процессов или информации;

c) указание на лицо, ответственное за выполнение процесса;

d) перечень необходимых ресурсов для выполнения процесса;
e) время начала, окончания или указание на продолжительность процесса;
f) определение соответствующего результата.

12.7 Продвижение по жизненному циклу качества данных

Мероприятия по управлению качеством данных должны соответствовать жизненному циклу качества данных, применяемому организацией:

a) стадии должны начинаться только в том случае, если недостающая информация об обнаруженных рисках на предыдущих стадиях обрабатывается в рамках деятельности организации по управлению рисками;

b) результаты плана по управлению качеством данных должны быть частью процесса управления конфигурацией и изменениями.

12.8 Обоснование качества данных

При необходимости организация должна предоставить понятное обоснование заявленного качества данных. Обоснование качества данных должно:

a) предоставить аргументацию для достижения соответствующего качества данных;
b) содержать все результаты, созданные в течение жизненного цикла качества данных.

12.9 Вывод из эксплуатации

Перед выводом из эксплуатации продукции или услуги, для которых осуществляется управление качеством данных, организация должна гарантировать, что:

a) обоснование качества данных (если применимо) завершается до стадии развертывания системы [см. *ГОСТ Р 71484.1—2024*, (рисунок 3)];

b) элементы предоставления данных из 7.3.7.2 поставляются (см. стадию 7 на рисунке 1) при условии, что имеются достаточные доказательства их качества;

c) требования 7.3.8 выполнены.

12.10 Результаты

Результаты управления качеством данных в конкретном проекте должны включать:

a) требования к качеству данных;
b) план управления качеством данных;
c) обоснование качества (если применимо);
d) отчеты о показателях, подтверждающих качество;
e) отчет о выводе данных из эксплуатации.

Приложение ДА
(справочное)

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном стандарте

Таблица ДА.1

Обозначение ссылочного национального стандарта	Степень соответствия	Обозначение и наименование ссылочного международного стандарта
ГОСТ Р 71476—2024 (ИСО/МЭК 22989:2022)	MOD	ISO/IEC 22989:2022 «Искусственный интеллект. Концепции и терминология искусственного интеллекта»
ГОСТ Р 71484.1—2024 (ИСО/МЭК 5259-1:2024)	MOD	ISO/IEC 5259-1:2024 «Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры»
Примечание — В настоящей таблице использовано следующее условное обозначение степени соответствия стандартов: - MOD — модифицированные стандарты.		

Библиография

[1] ISO/IEC FDIS 5259-5, Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance framework

[2] ISO/IEC CD TR 5259-6, Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 6: Visualization framework for data quality

[3] ISO/IEC FDIS 5259-2, Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures

[4] *ПНСТ 840—2023 (ISO/IEC TR 24368:2022) Искусственный интеллект. Обзор этических и общественных аспектов*

[5] *ПНСТ 839—2023 (ISO/IEC TR 24027:2021) Искусственный интеллект. Смещенность в системах искусственного интеллекта и при принятии решений с помощью искусственного интеллекта*

[6] *ПНСТ 776—2022 (ISO/IEC FDIS 23894:2023) Информационные технологии. Интеллект искусственный. Управление рисками*

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: искусственный интеллект, качество данных, характеристики качества данных, набор данных, подготовка наборов данных, вывод данных из эксплуатации, управление качеством данных, интеграция системы управления качеством данных, жизненный цикл управления качеством данных, свойства данных, комплектование данных, верификация и валидация данных, управление средствами обработки данных, роли и обязанности в управлении качеством данных, план качества данных

Технический редактор *И.Е. Черепкова*
Корректор *Р.А. Ментова*
Компьютерная верстка *Л.А. Круговой*

Сдано в набор 31.10.2024. Подписано в печать 12.11.2024. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 3,26. Уч.-изд. л. 2,80.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.

gost.e-ecolog.ru gost.e-ecolog.ru

gost.e-ecolog.ru