ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ
РОССИЙСКОЙ
ФЕДЕРАЦИИ

ГОСТ (Р)
_____–_____.

Информационные технологии

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Качество данных для аналитики и машинного обучения (ML). Часть 3. Требования и рекомендации по управлению качеством данных

(ISO/IECWD 5259-3 Artificial intelligence — Data quality for analytics and ML — Part 3: Data quality management requirements and guidelines)

Настоящий проект стандарта не подлежит применению до его принятия

Москва
Стандартинформ
20XX

Предисловие

1 ПОДГОТОВЛЕН 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__ _____ 202_ г. № ____.

4 Настоящий стандарт идентичен разрабатываемому международному стандарту ISO/IECWD 5259-3 Artificial intelligence — Data quality for analytics and ML — Part 3: Data quality management requirements and guidelines , учитывающих специфику национальной стандартизации в области искусственного интеллекта и больших данных.

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 (www.rst.gov.ru)

© ISO, 20XX — Все права сохраняются
© Стандартинформ, оформление, 20XX

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

Оглавление

Предисловие II
Введение IX
1 Область применения 11
2 Нормативные ссылки 11
3 Термины и определения 12
3.1 Ввод данных 12
3.2 Элемент данных 12
3.3 Соответствие данных требуемой характеристике качества 13
3.4 проект машинного обучения 13
3.5 элемент метаданных 13
4 Обозначения и сокращения 13
5.1 Применение 13
5.2 Степени соответствия 14
6 Общее управление качеством данных 14
6.1 Общие положения 14
6.3 Требования и рекомендации 14
6.3.1 Общие положения 15
6.3.2 Культура качества 15
6.3.3 Управление аномалиями качества данных 15
6.3.4 Управление компетенциями 16
6.3.5 Управление ресурсами 16
6.3.6 Интеграция системы управления 16
6.3.7 Документация 17
6.3.8 Аудит и оценка качества данных 17
7. Рекомендации и требования, относящиеся к жизненному циклу управления качеством данных 19
7.2.2.4 Комплектование данных 22
7.2.2.5 Предварительная обработка данных 22
7.2.2.6 Расширение данных 22
7.2.2.7 Предоставление данных 22
7.2.2.8 Вывод данных из эксплуатации 22

7.2.3 Независимая от проекта адаптация жизненного цикла управления качеством данных 23
7.2.4 Горизонтальные аспекты жизненного цикла управления качеством данных 23
7.2.4.1 Валидация и верификация 23
7.2.4.2 Управление изменениями....24
7.2.4.3 Управление конфигурацией....24
7.3.1 Мотивация данных и концептуализация....24
7.3.1.1 Общие положения 24
7.3.1.2 Анализ заинтересованных сторон....24
7.3.1.3 Анализ сложности и реализуемости 25
7.3.2 Спецификация данных....25
7.3.2.1 Общие положения 25
7.3.2.2 Формат данных 26
7.3.2.3 Статистические свойства и делимость....26
7.3.2.4 Совместимость 27
7.3.2.5 Вспомогательные ресурсы и инструменты....27
7.3.3 Планирование данных 27
7.3.3.1 Общие положения 27
7.3.3.2 Специальный план жизненного цикла данных 28
7.3.4 Комплектование данных 28
7.3.4.1 Общие положения 28
7.3.4.2 Источник данных 29
7.3.4.3 Сбор данных 29
7.3.4.4 Обработка данных....30
7.3.5 Предварительная обработка данных....30
7.3.5.1 Общие положения 30
7.3.5.2 Очистка данных 31
7.3.6 Расширение данных....31
7.3.6.1 Общие положения 31
7.3.6.2 Маркировка данных....31
7.3.6.3 Оценка результатов расширения....31
7.3.7 Предоставление данных....32
7.3.7.1 Общие положения 32

7.3.7.2 Элементы предоставления данных 32
7.3.7.3 Документация 33
7.3.7.4 Отслеживание и улучшение 33
7.3.7.5 Оптимизация данных 34
7.3.7.6 Обязательство поддержки 34
7.3.8 Вывод данных из эксплуатации 34
7.3.8.1 Общие положения 34
7.3.8.2 Передача данных 35
7.3.8.3 Удаление данных 35
7.3.8.4 Частичное удаление данных по запросу 36
7.4 Артефакты 36
7.4.1 Артефакты этапа мотивации данных и концептуализации 36
7.4.2 Артефакты этапа спецификации данных 37
7.4.3 Артефакты этапа планирования данных 37
7.4.4 Артефакты этапа комплектования данных 37
7.4.5 Артефакты этапа предварительной обработки данных 37
7.4.6 Артефакты этапа расширения данных 38
7.4.7 Артефакты этапа предоставления данных 38
7.4.8 Артефакты этапа вывода данных из эксплуатации 38
8 Горизонтальные процессы 38
8.1 Цель 38
8.2 Общие положения 38
8.3 Требования и рекомендации 39
8.3.1 Верификация и валидация 39
8.3.1.1 Общие положения 39
8.3.1.2 Контрольные точки жизненного цикла качества данных 39
8.3.1.3 Улучшение 39
8.3.2 Управление конфигурацией 39
8.3.3 Управление изменениями 40
8.3.3.1 Общие положения 40
8.3.3.2 Планирование управления изменениями 40
8.3.3.3 Спецификация запроса на изменение 40
8.3.3.4 Анализ запроса на изменение 41
8.3.3.5 Оценка запроса на изменение 41

8.3.3.6 Внедрение и документирование изменений.....41
8.4 Артефакты 41
8.4.1 Артефакты верификации и валидации 42
8.4.2 Артефакты управления конфигурацией 42
8.4.3 Артефакты управления изменениями.....42
9 Управление качеством данных в цепочках поставок.....42
9.1 Цель.....42
9.2 Требования и рекомендации 43
9.3 Артефакты 43
10 Управление средствами обработки данных.....43
10.1 Цель.....44
10.2 Требования и рекомендации 44
10.3 Артефакты 44
11 Управление зависимостями качества данных 44
11.1 Цель.....44
11.2 Требования и рекомендации 45
11.3 Артефакты 45
12 Управление качеством данных по конкретному проекту.....45
12.1 Цель.....45
12.2 Требования и рекомендации 45
12.2.1 Контекст и предполагаемое использование.....46
12.2.1.1 Цель.....46
12.2.1.2 Требования и рекомендации 46
12.3 Спецификация и управление требованиями к качеству данных 47
12.3.1 Цель.....47
12.3.2 Требования и рекомендации 47
12.4 Управление рисками 47
12.4.1 Цель.....47
12.4.2 Требования и рекомендации 48
12.4.3 Артефакты.....48
12.5 Роли и обязанности в управлении качеством данных.....48
12.5.1 Цель.....48
12.5.2 Требования и рекомендации 49
12.5.3 Артефакты.....49

12.6 Адаптация мероприятий по обеспечению качества данных 49
12.7 Планирование и координация деятельности по обеспечению качества данных 49
12.7.1 Общие положения 49
12.7.2 План качества данных.....50
12.7.3 Планирование процессов 50
12.8 Развитие жизненного цикла качества данных.....50
12.9 Обоснование качества данных.....51
12.10 Выпуск 51
12.11 Артефакты.....51
Библиография 52
Приложение А (справочное) Сведения о соответствии ссылочных международных стандартов национальным стандартам 53

Введение

Многие современные средства аналитики и машинного обучения (МО) представляют собой «черный ящик», поэтому сложно оценить их производительность и качество только на основе результатов (или обученной модели). Качество аналитики, продуктов и услуг на основе машинного обучения зависит от качества данных, используемых для обучения моделей. Следовательно, управление качеством данных имеет важное значение, поскольку оно часто помогает предсказать результаты аналитики на основе технологии машинного обучения.

Внедрение системы управления качеством данных облегчает управление качеством продуктов и услуг, использующих аналитику и технологии машинного обучения. В настоящем документе определены словарь, требования и рекомендации для обмена информацией и согласования при управлении качеством данных. Система управления качеством данных обеспечивает прозрачность и возможность аудита посредством самооценки или оценки третьей стороной. Это способствует достижению удовлетворенности заинтересованных сторон и упрощает управление требованиями к качеству, производительности. В частности, определяются требования к системе управления качеством данных, предложены меры качества данных, для наиболее часто используемых технологий аналитики и машинного обучения.

Поскольку требования к качеству данных различаются в зависимости от контекста и области применения, в документе представлен общий набор требований и рекомендаций, касающихся общих этапов жизненного цикла данных. Жизненный цикл данных обычно тесно связан с жизненным циклом системы ИИ, которая их использует, и, следовательно, зависит от него. Данный документ не предписывает, какой жизненный цикл системы ИИ следует использовать. Вместо этого он предоставляет общие принципы, которые позволяют пользователям, заявляющим о соответствии этому документу, гибко использовать несколько моделей жизненного цикла, если процессы жизненного цикла могут быть совместно использованы.

Этот документ является частью серии стандартов ISO/IEC 5259. Другие части серии ISO/IEC 5259 включают:

- ISO/IEC 5259-1:—1, Искусственный интеллект. Качество данных для аналитики и машинного обучения (ML). Часть 1. Обзор, терминология и примеры;
- ISO/IEC 5259-2:–2, Искусственный интеллект. Качество данных для аналитики и машинного обучения (ML). Часть 2. Показатели качества данных;

ГОСТ (Р) _____—_____.
(проект)

• ISO/IEC 5259-3:—3, Искусственный интеллект. Качество данных для аналитики и машинного обучения (ML). Часть 3. Требования и рекомендации по управлению качеством данных;

• ISO/IEC 5259-5:—4, Искусственный интеллект. Качество данных для аналитики и машинного обучения (ML). Часть 5. Управление качеством данных.

ГОСТ (Р) _____—_____.
(проект)

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Информационные технологии

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Структура жизненного цикла данных

Artificial intelligence — Data quality for analytics and ML — Part 3: Data quality management requirements and guidelines

---

Дата введения – 2022-__-__

1 Область применения

Этот документ определяет требования и рекомендации по созданию, внедрению, поддержке и постоянному повышению качества данных, используемых в областях аналитики и машинного обучения. В документе не определяются детально сам процесс, методы или показатели. Он содержит требования и рекомендации для организации процесса управления качеством, а также эталонный процесс и методы, которые могут быть адаптированы для выполнения данных требований. Требования и рекомендации, изложенные в этом документе, являются общими и предназначены для применения ко всем организациям, независимо от их типа, размера или характера.

2 Нормативные ссылки

В настоящем стандарте использованы следующие нормативные ссылки. Для датированных ссылок применяют только указанное издание ссылочного стандарта, для недатированных — последнее издание (включая все изменения).

ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (Информационные технологии. Искусственный интеллект. Концепции искусственного интеллекта и терминология).

ISO/IEC 5259-1:202X: Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples (Качество данных для аналитики и машинного обучения. Часть 1. Обзор, терминология и примеры).

ISO/IEC 5259-2:202X — Information technology — Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures (Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 2. Показатели качества данных).

ISO/IEC 5259-4: 202X — Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework (Искусственный интеллект. Качество данных для аналитики и машинного обучения. Часть 4. Структура процесса обеспечения качества данных

.

3 Термины и определения

ИСО и МЭК поддерживают терминологические базы данных для использования в стандартизации, расположенные по следующим адресам:

Платформа ИСО для онлайн-просмотра материалов по стандартам (Online Browsing Platform, OBP) доступна по адресу https://www.iso.org/obp/ui

База данных МЭК «Электропедия» (IEC Electropedia) доступна по адресу http://www.electropedia.org/

Для целей настоящего документа применяются термины и определения, приведенные в ISO/IECWD 5259-3 Artificial intelligence — Data quality for analytics and ML — Part 3: Data quality management requirements and guidelines (ISO / IECWD 5259-3 Искусственный интеллект – Качество данных для аналитики и машинного обучения – Часть 3: Требования и руководящие принципы управления качеством данных).

3.1 Ввод данных

единичный фрагмент информации (single piece of information)

3.2 Элемент данных

часть входных данных (part of a data entry)

3.3 Соответствие данных требуемой характеристике качества

утверждение о том, в какой степени данные удовлетворяют характеристике качества данных (statement to what degree data satisfy a data quality characteristic).

3.4 проект машинного обучения

проект, на основе системы искусственного интеллекта, использующей машинное обучение и связанные с ним данные на протяжении всего жизненного цикла данных (project that is responsible for an AI system using ML and its associated data throughout the data's entire life cycle)

3.5 элемент метаданных

часть метаданных

4 Обозначения и сокращения

ИИ	—	искусственный интеллект (artificial intelligence)
МО	—	машинное обучение (machine learning)
DQMLC	—	жизненный цикл управления качеством данных (data quality management life cycle)
DLC		жизненный цикл данных (data life cycle)

5.1 Применение

Данный документ может применяться:

- организацией для создания и адаптации системы управления качеством данных в аналитике и машинном обучении, постоянного улучшения процессов, для определения, отслеживания и оценки требований к данным;
- пользователем данных и оператором данных для согласования характеристик качества данных и обеспечения выполнения согласованных требований к качеству данных, что позволяет упростить обмен данными.

5.2 Степени соответствия

Степени соответствия положениям данного документа:

а) полное соответствие при этом:

1) предоставляется Заявление о полном соответствии, в котором указывается набор процессов, которые соответствуют документу.

2) все процессы, перечисленные в заявлении соответствуют требованиям.

3) предоставляются доказательства, подтверждающие полное соответствие.

б) частичное соответствие при этом:

1) предоставляется Заявление о частичном соответствии, в котором указывается набор процессов, которые соответствуют документу.

2) соблюдены требования раздела 6 и соответствующие требования других разделов настоящего документа.

3) предоставляются доказательства, подтверждающие частичное соответствие.

6 Общее управление качеством данных

6.1 Общие положения

Качество данных оказывает большое влияние на результаты аналитики и алгоритмов машинного обучения. Качество данных имеет внутреннюю составляющую и составляющую, зависящую только от системы. Поэтому данные могут подходить для использования в одном приложении, но не подходить для другого. Следовательно, качество данных устанавливается и поддерживается для каждого приложения аналитики и машинного обучения.

ISO/IEC 5259-1: - описывает терминологию и понятия качества данных, используемые в этом документе. ISO/IEC 5259-2: - описывает модель качества данных и меры качества данных, используемые в этом документе. ISO/IEC 5259-4: - описывает структуру процесса качества данных, используемую в этом документе.

6.3 Требования и рекомендации

6.3.1 Общие положения

Следующие требования и рекомендации относятся к организации в целом.

6.3.2 Культура качества

Организация должна поддерживать культуру качества путем:

1) наличия правил и процессов для достижения качества (согласно этому документу) с учетом модели качества данных применительно к продуктам и услугам;
2) определения и внедрения процессов управления качеством данных и выполнения соответствующих действий по обеспечению качества данных;
3) интеграции процессов и действий по управлению качеством данных в другие процессы и действия по управлению, такие как общее управление качеством и управление рисками;
4) сбора документации о проведенных мероприятиях;
5) предоставления ресурсов, достаточных для выполнения действий по управлению качеством данных;
6) мониторинга, анализа и улучшения процессов управления качеством данных;
7) предоставления необходимых полномочий заинтересованным лицам.

6.3.3 Управление аномалиями качества данных

Организация должна обеспечить соответствие требованиям к качеству данных посредством:

а) наличия процессов для сообщения, анализа, оценки, разрешения и закрытия (осуществления решения или объяснения причины отсутствия необоснованного риска) проблем качества данных;

б) документирования закрытых вопросов;

в) передачей на уровень выше вопросов, которые не могут быть закрыты на текущем уровне.

6.3.4 Управление компетенциями

Организация должна управлять компетенциями посредством:

a) документирования необходимых технологий и инструментов для обработки данных;

б) обеспечением того, чтобы лица, участвующие в обработке данных, обладали достаточными навыками для выполнения своей деятельности и обязанностей;

c) подтверждением наличия достаточных навыков работы с системой качества данных;

d) ведением соответствующих записей об обучении и опыте, подтверждающих заявления о соответствующих навыках.

ПРИМЕЧАНИЕ. Организация может использовать внешние источники компетенций.

6.3.5 Управление ресурсами

Организация должна предоставлять ресурсы, необходимые для управления качеством данных, в том числе:

a) соответствующие лицензии на программное обеспечение и техническую поддержку необходимых инструментов;

b) необходимую ИТ-инфраструктуру или сервисы для удовлетворения требований к хранению данных и их обработке;

в) наличие квалифицированных кадров.

6.3.6 Интеграция системы управления

Организации следует интегрировать свою деятельность по управлению качеством данных в существующую систему управления, включая системы управления качеством продукции или услуг, а также для разработки и использования систем ИИ. Последствия наличия нескольких ролей и заинтересованных сторон должны регулироваться системой менеджмента качества, включая решение любых конфликтов интересов.

ПРИМЕЧАНИЕ 1. Управление заинтересованными сторонами рассматривает наличие нескольких ролей у отдельного лица. Пользователь продуктов или услуг, основанных на аналитике и машинном обучении, также может быть оператором или поставщиком данных.

ПРИМЕЧАНИЕ 2. Организации могут использовать ИСО/МЭК 42001:—1) [3] для определения системы управления для разработки или использования систем ИИ.

ПРИМЕЧАНИЕ 3. Организации могут использовать ISO 9001:2015 [4] для определения своей системы менеджмента качества.

6.3.7 Документация

Документация должна быть понятна всем заинтересованным сторонам проекта. Источники на языке, который не понятен соответствующей заинтересованной стороне, должны сопровождаться кратким изложением на языке, понятном этой стороне.

Документация должна быть доступна для всех заинтересованных сторон, когда это необходимо. Накладные расходы на доступ должны быть сведены к минимуму.

Документация должна включать контекст или ссылки, необходимые для того, чтобы сделать ее понятной для будущих заинтересованных сторон, не участвующих в текущем проекте. Эта практика может позволить им оценить набор данных для потенциального повторного использования, частично или полностью.

6.3.8 Аудит и оценка качества данных

Процессы, которые внедрены подлежат проверке с использованием аудита качества данных, который должен быть основан на оценке:

а) плана обеспечения качества данных в соответствии с организационными правилами и процессами;

б) результатов применения требований, описываемых характеристиками модели качества данных;

с) аргументов, обосновывающих, почему цели этого документа были достигнуты;

д) являются ли план качества данных и все рабочие продукты полными, непротиворечивыми и правильными в соответствии с этим документом;

е) рекомендаций по улучшению.

Качество данных должно оцениваться на основе того, обеспечивают ли эти данные достижение целей настоящего документа, текущего уровня техники и прикладных знаний в инженерной области.

Оценка качества данных должна планироваться в начале этапа спецификации и выполняться до предоставления данных (см. рисунок 1, этап 7: предоставление данных).

Рисунок 1 — Рисунок 1 – Общий DQPF и примеры результатов

6.3.9 Проверка соответствия и меры качества данных

Качество данных должно быть подтверждено соответствующими мерами качества данных в соответствии с ИСО/МЭК 5259-2:—. Проверка качества данных включает, но не ограничивается:

а) подтверждающими обзорами ключевых продуктов работы. Каждый такой обзор:

1) должен быть завершен до предоставления данных;

2) может основываться на том, были ли достигнуты цели настоящего документа;

3) могут выполняться другими организациями (ответственность остается за данной организацией).

б) содержит аудит качества реализованных процессов;

c) содержит оценку качества данных.

Все рабочие продукты должны пройти подтверждающие проверки.

Лица, проводящие эти проверки, должны иметь доступ к заинтересованным лицам, соответствующей информации и необходимым ресурсам.

6.3.10 Особенности управление качеством в проекте

Организация должна управлять специфическими для проекта данными посредством:

a) создания подходящей системы управления качеством данных для конкретного проекта, которая отвечает всем требованиям конкретного проекта МО;

b) ведения списка соответствующих заявлений о качестве данных. Там, где это применимо, должны быть задокументированы количественные контрольные показатели для показателей качества данных, в противном случае должны быть задокументированы качественные контрольные показатели;

c) применением соответствующих процессов для определения и управления всеми показателями качества данных, относящимися к проекту.

Управление качеством данных для конкретного проекта должно соответствовать требованиям раздела 12.

П р и м е ч а н и е — Меры качества данных описаны в ИСО/МЭК 5259-2:—.

6.4 Артефакты

Артефакты системы управления качеством данных включают:

a) специфичные для организации правила и процессы качества данных;

б) свидетельство управления компетенцией;

c) свидетельство системы управления качеством данных;

d) идентификацию используемых показателей качества данных;

e) документирование применяемых контрольных показателей качества данных;

f) выявленные отчеты об аномалиях качества.

7. Рекомендации и требования, относящиеся к жизненному циклу управления качеством данных

7.1 Цель

Целью жизненного цикла управления качеством данных (DQMLC) является установление и поддержание качества данных на протяжении всего жизненного цикла данных. Модель жизненного цикла данных описана в ИСО/МЭК 5259-1: пункт 5.2.2, рисунок 2

Рисунок 2 — Связь между моделью DLC и DQPF

7.2 Общие положения

7.2.1 Жизненный цикл управления качеством данных

Качеством данных необходимо управлять на всех этапах жизненного цикла данных. Организация может использовать жизненный цикл данных, описанный в ИСО/МЭК 8183, или организация может использовать другую подходящую модель жизненного цикла для управления качеством данных. Модель жизненного цикла управления качеством данных (DQMLC), показанная на рис. 1, дает рекомендации по выполнению требований к качеству данных для использования в контексте аналитики и машинного обучения. Он выделяет отдельные этапы, которые имеют отношение к управлению качеством данных, и облегчает группировку и организацию требований и руководств, которые следует учитывать при управлении качеством

данных. Модель DQMLC не предписывает временной порядок этапов. Этапы DQMLC описаны в пункте 7.2.2.

Недостатки в качестве данных могут возникать на любом этапе жизненного цикла данных. Управление качеством данных должно устанавливать и поддерживать качество данных как можно раньше в жизненном цикле данных.

ПРИМЕЧАНИЕ. Как правило, более сложно обнаружить и исправить недостатки качества данных постфактум, чем управлять рисками качества данных при их появлении. Например, ошибок, возникающих при сборе данных, легче избежать с помощью соответствующего управления качеством, чем пытаться обнаруживать и исправлять ошибки на более поздних этапах жизненного цикла данных.

7.2.2 Этапы жизненного цикла управления качеством данных

7.2.2.1 Мотивация данных и концептуализация

Управление качеством данных начинается со стадии мотивации и концептуализации. Потенциальные упущения в качестве данных необходимо выявлять и устранять, когда становится очевидной первая потребность в данных для аналитики и контекста машинного обучения. Среди прочего, оценивается мотивация и каким образом предполагается использовать данные для управления характеристиками качества, такими как соответствие требованиям, актуальность и возможные этические проблемы.

7.2.2.2 Спецификация данных

На этапе спецификации данных управление качеством данных предъявляет требования к созданию спецификации данных. Рекомендации по управлению качеством данных облегчают выявление ошибочных, неполных или противоречивых требований и планов. Например, учитывая содержание (цель) аналитики и машинного обучения, управление качеством подтверждает, что данные соответствуют требованиям данной цели.

7.2.2.3 Планирование данных

На этапе планирования данных составляется план, обеспечивающей соответствие спецификации данных. Он включает в себя планирование конкретных задач и ресурсов для сбора и обработки данных на протяжении всего жизненного цикла данных, а также методы оценки и критерии соответствия.

7.2.2.4 Комплектование данных

С точки зрения управления качеством данных комплектование данных может относиться к сбору данных, созданию данных или сбору и объединению существующих наборов данных. Выбор способа комплектования данных влияет на управление качеством данных. Независимо от способа получения данных результат этого этапа считается необработанным входом для последующих этапов жизненного цикла данных, даже если импортируется уже структурированный набор данных.

7.2.2.5 Предварительная обработка данных

На этапе предварительной обработки данных формируются все требования и рекомендации по управлению качеством данных, связанные с обработкой необработанных входных данных, полученных на этапе комплектования данных. Действия на этом этапе позволяют сократить объем данных и повысить их информативность, за счет фильтрации и очищения необработанных данных.

7.2.2.6 Расширение данных

Расширение данных начинается на этапе их предварительной обработки и формирует все требования и рекомендации по управлению качеством, связанные с добавлением дополнительных данных. Это включает связывание метаданных, добавление меток данных или другие преобразования, которые увеличивают размер набора данных.

7.2.2.7 Предоставление данных

Цель предоставления данных — обеспечить наличие данных там, где они необходимы. Требования к управлению качеством данных на этом этапе включают, среди прочего, поддержание целостности данных во время передачи, контроль доступа, документирование и проверку того, что контекст использования соответствует предполагаемому использованию данных.

7.2.2.8 Вывод данных из эксплуатации

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

7.2.3 Независимая от проекта адаптация жизненного цикла управления качеством данных

Организация может использовать данный жизненный цикл только в том случае, если она:

а) объединяет или разделяет этапы

b) выполняет требуемые процессы на разных или дополнительных стадиях (см. ISO/IEC 5259-4);

в) повторяет этапы;

г) выполняет действия одновременно, если они не зависят друг от друга;

е) обоснованно пропускает этапы, которые не применяются

f) использует средства управления качеством данных для верификации и валидации, управления изменениями и управления конфигурацией, как показано на рисунке 1.

7.2.4 Горизонтальные аспекты жизненного цикла управления качеством данных

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

а) валидацию и верификацию;

б) управление изменениями.

в) управление конфигурацией.

Эти процессы должны выполняться вне независимости от проекта.

7.2.4.1 Валидация и верификация

Валидацию и верификацию качества данных следует проводить на всех соответствующих этапах жизненного цикла управления качеством данных.

7.2.4.2 Управление изменениями

Управление изменениями гарантирует, что любое изменение данных или процессов не повлияет отрицательно на качество данных.

7.2.4.3 Управление конфигурацией

Управление конфигурацией гарантирует, что на данные и процессы можно ссылаться однозначно, и что никакие изменения управления конфигурацией не повлияют на качество данных.

Управление конфигурацией гарантирует, что на данные и процессы можно ссылаться однозначно, и что никакие изменения управления конфигурацией не повлияют на качество данных.

7.3.1 Мотивация данных и концептуализация

7.3.1.1 Общие положения

Организация должна:

а) задокументировать необходимость и причину потребности в данных;

б) указать объем, цель и предполагаемое использование данных;

с) выявить потенциальные социальные и этические проблемы в отношении предполагаемого использования данных;

г) показать соответствие ценностям и культуре организации;

е) оценить экономическую ценность, затраты и потенциальное значение данных для бизнеса;

f) документировать требования к качеству данных.

П р и м е ч а н и е — Для получения дополнительной информации об общественных и этических проблемах см. ISO/IEC TR 24368:—1).

7.3.1.2 Анализ заинтересованных сторон

Организация должна:

определить и выполнить анализ всех заинтересованных сторон в отношении данных и последствий для управления качеством данных;

выявить потенциально противоречивые требования к качеству данных;

установить соответствие ценностям и культуре организации.

ПРИМЕЧАНИЕ. Описание потенциальных ролей и подролей заинтересованных сторон ИИ представлено в ISO/IEC DIS 22989:—.

7.3.1.3 Анализ сложности и реализуемости

Организация должна поддерживать анализ сложности и реализуемости для установления доступности данных и достижимости целей организации в области качества на основе имеющихся навыков и ресурсов. Анализ сложности и реализуемости должен обновляться на протяжении всего жизненного цикла данных.

П р и м е ч а н и е — Предыдущие проекты данных, открытые наборы данных или научная литература, которые достигли целевых показателей качества, являются примерами доказательств, подтверждающих осуществимость.

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 непротиворечивы и охватывают все требования для предполагаемого использования.

Спецификация данных должна включать минимальный перечень характеристик, а также тех, которые являются необязательными в зависимости от предполагаемого использования системой ИИ.

Спецификация должна по крайней мере охватывать:

а) описание характера данных;

б) предполагаемую цель проекта аналитики и машинного обучения, которой должны соответствовать данные;

с) юридические требования и договорные обязательства;

г) обеспечение безопасности и защищенности;

е) социальные и этические проблемы;

f) возможность предвзятости и необъективности;

g) соображения конфиденциальности;

h) требования к данным предметной области и проекта;

i) модель качества данных и требуемый уровень качества для обеспечения предполагаемого использования данных.

ПРИМЕЧАНИЕ 1 Спецификация данных может быть основана на описании ISO/IEC TR 24027 [1] обработки нежелательной предвзятости и необъективности на протяжении всего жизненного цикла системы ИИ.

ПРИМЕЧАНИЕ 2. Спецификация данных включает различные интерфейсы с проектом МО (например, данные для обучения, данные для тестирования или текущее управление данными).

ПРИМЕЧАНИЕ 3. Некоторые алгоритмы МО требуют определенных статистических свойств обучающих данных.

ПРИМЕЧАНИЕ 4. Соображения конфиденциальности могут включать принцип минимизации данных (также известный как обезличивание). Минимизация данных означает, что собирается или используется только минимальный объем данных, необходимых для работы системы.

7.3.2.2 Формат данных

Спецификация данных должна включать следующую информацию о формате данных:

а) кодирование;

б) частота (время);

в) разрешение (интервал);

г) синтаксис;

д) семантика;

f) структура ассоциированной метаинформации;

g) диапазоны ожидаемых значений;

h) необязательные и обязательные элементы;

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

j) ссылки и перекрестные ссылки.

П р и м е ч а н и е — Семантика относится к интерпретации и использованию данных, включая допустимые операции над данными.

7.3.2.3 Статистические свойства и делимость

Спецификация данных должна включать спецификацию делимости на подмножества (например, данные обучения, валидации и тестирования). Подмножества не должны пересекаться.

Каждое подмножество должно:

а) иметь соответствующие статистические характеристики;

б) быть репрезентативными;

в) иметь соответствующий размер.

7.3.2.4 Совместимость

Спецификация данных должна включать описание совместимости данных. Описание функциональной совместимости должно учитывать, с какими другими наборами данных и алгоритмами данные могут взаимодействовать, как новые версии могут быть созданы путем агрегирования или слияния и как данные поддерживаются в согласованном состоянии при обновлении ссылок.

Спецификация данных должна включать требования к версии и обновлению для обеспечения функциональной совместимости.

7.3.2.5 Вспомогательные ресурсы и инструменты

Спецификация данных должна включать все соответствующие требования по обработке данных, в том числе:

a) спецификации вспомогательных инструментов;

b) минимальную спецификацию аппаратных средств для обработки данных и управления инструментами;

c) размер экрана и разрешение для визуализации данных;

г) требования к хранению;

e) требования к сети;

f) требования к доступности.

7.3.3 Планирование данных

7.3.3.1 Общие положения

Организация должна разработать план данных, охватывающий процессы, действия и ресурсы, чтобы соответствовать спецификации данных на протяжении всего жизненного цикла данных.

План данных должен включать:

а) цели качества данных в соответствии со спецификацией данных

б) распределение выделенных ресурсов;

с) положения о выполнении юридических и договорных обязательств;

d) процессы для управления и обновления плана данных, когда это необходимо;

е) положения, обеспечивающие прослеживаемость и воспроизводимость процессов.

ПРИМЕЧАНИЕ. Можно делегировать ответственность за процесс, если ответственное лицо всегда может быть однозначно идентифицировано.

7.3.3.2 Специальный план жизненного цикла данных

План данных должен включать спецификации процесса, включая описание, соответствующие действия, предполагаемые результаты, включая назначение ответственного лица за каждый процесс, охватывая следующие этапы жизненного цикла данных:

а) план комплектования данных;

б) план предварительной обработки данных;

с) план расширения данных;

г) план предоставления данных;

е) план вывода данных из эксплуатации.

План данных, соответствующий этапам жизненного цикла данных, должен соответствовать требованиям, изложенным в пунктах 7.3.4–7.3.8.

7.3.4 Комплектование данных

7.3.4.1 Общие положения

Организация должна определить процессы комплектования данных в соответствии со спецификацией данных и включить их в план данных.

Полученные данные должны соответствовать спецификации данных либо непосредственно во время получения, либо у организации должны быть планы по приведению в соответствие спецификации после выполнения этапов предварительной обработки данных и расширения данных на этапах жизненного цикла данных.

Организация должна отслеживать процессы комплектования данных и документировать отклонения от ожидаемых результатов или отклонения от запланированных процессов.

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

ПРИМЕЧАНИЕ. Процессы комплектования данных применимы независимо от источника данных.

7.3.4.2 Источник данных

Организация должна установить режим комплектования данных, если:

- требуемые данные уже существуют и непосредственно доступны для повторного использования;
- существующие данные могут быть преобразованы в соответствии с требованиями;
- данные можно купить;
- необходимо собрать новые данные;
- данные могут быть сгенерированы (например, с помощью моделирования или других вычислительных средств).

Организация может делегировать комплектование данных, но организация по-прежнему несет ответственность за соблюдение требований организации к качеству данных. Набор данных может иметь задокументированные условия повторного использования или лицензионные требования, которые должны быть учтены.

ПРИМЕЧАНИЕ. Доступные наборы данных с открытым исходным кодом являются жизнеспособным 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, организация должна предоставить:

a) обоснование для сбора новых данных

b) описание режима сбора данных (например, с помощью датчиков, ручной ввод данных, значение производной, моделирование, синтетические данные);

c) спецификацию сбора данных, включая:

1) соответствующие настройки конфигурации и параметров;

2) условия эксплуатации;

3) обнаружение и устранение ошибок;

4) необходимые навыки и ресурсы;

5) если применимо, то спецификацию и место установки датчиков;

г) необходимые ресурсы и навыки для сбора и обработки данных;

е) методы анонимизации или псевдонимизации, если применимо;

f) документацию по шкалам измерения данных (например, номинальная, порядковая, интервальная и относительная) и единицам измерения;

g) ожидаемые отклонения от целевого качества данных.

ПРИМЕЧАНИЕ Метод сбора данных может внести систематические ошибки в собранные данные, особенно если установка сбора данных отличается от указанной в спецификации установки. Одним из возможных способов является использование различных методов сбора для обеспечения избыточности.

7.3.4.4 Обработка данных

Организация должна установить соответствующие процессы и инструменты для управления, визуализации и анализа полученных данных. Полученные данные должны соответствовать спецификации данных.

Организация должна верифицировать полученные данные.

Организация должна определить процессы для управления полученными данными, включая:

а) обзор качества;

б) контроль доступа;

с) прослеживаемость происхождения и модификации данных;

г) контроль версий;

е) хранение, размещение и резервное копирование данных;

f) операции с наборами данных, такие как обновление, слияние, сортировка и срез;

g) процессы для обнаружения, минимизации и устранения 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 в соответствии со спецификацией данных и включить их в план данных.

7.3.5.2 Очистка данных

Организация должна определить, каким образом выполняется:

a) обнаружение и устранение недостающих данных;

б) обнаружение и устранение дублирования данных;

c) обнаружение и устранение выбросов и других аномалий;

d) обнаружение и устранение предвзятости и небъективности, отклонений и масштабирования;

e) обнаружение и обработка нестандартных значений;

f) обнаружение и удаление ненужных данных;

g) преобразование данных, включая нормализацию данных;

h) методы проверки результатов очистки данных;

i) если применимо, анонимизация данных.

ПРИМЕЧАНИЕ. Некоторые преобразования данных необратимы.

7.3.6 Расширение данных

7.3.6.1 Общие положения

Организация должна определить процессы расширения данных для соответствия спецификации данных и включить их в план данных.

7.3.6.2 Маркировка данных

Если применимо, организация должна указать маркировку данных, включая:

a) спецификация маркировки данных;

б) необходимые навыки и ресурсы;

в) выбор данных для маркировки;

г) мониторинг и управление качеством процессов маркировки;

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

7.3.6.3 Оценка результатов расширения

Организация должна указать:

а) инструменты и методы, используемые для компьютерного расширения;

б) выбранные характеристики;

с) сформированные журналы и метаданные.

7.3.7 Предоставление данных

7.3.7.1 Общие положения

Организация должна:

а) определить процессы предоставления данных соответствующие спецификации данных и включить их в план данных;

b) обеспечить проверяемые доказательства того, что предоставленные данные и метаданные соответствуют всем установленным требованиям;

с) документировать и устранять разумно ожидаемое неправомерное использование данных;

г) предоставить документацию о разумно ожидаемых непреднамеренных последствиях использования данных;

е) внедрить соответствующие процессы для обеспечения предоставления только элементов и версий в соответствии с пунктом 7.3.7.2;

ж) гарантировать, что предоставление не изменяет данные;

з) внедрить соответствующие процессы аутентификации и доступа к данным;

и) контроль версий всех подготовленных элементов;

к) скомплектовать и предоставить данные, включая документацию и другие связанные с ней компоненты, чтобы пользователь данных мог:

1) определить, соответствуют ли данные требованиям использования;

2) использовать данные по назначению.

П р и м е ч а н и е — Например, представление данных, в котором хранится большое количество десятичных знаков, может ввести в заблуждение, предполагая более высокую точность, чем достигнутая, и, следовательно, может привести к непреднамеренному использованию или необоснованным допущениям при использовании.

7.3.7.2 Элементы предоставления данных

В описание этапа предоставления данных должны быть включены следующие пункты:

а) данные, включая любые разбиения, если это необходимо, для формирования данных обучения, тестирования и валидации;

б) документация (см. пункт 7.3.7.3);

c) образец данных, имеющий соответствующий синтаксис и формат;

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

7.3.7.3 Документация

Вместе с данными должна быть представлена следующая документация:

a) область использования;

б) инструкции по применению;

c) статистические свойства данных;

d) все схемы элементов данных и элементов метаданных;

e) соответствующие заинтересованные стороны данных;

f) роли и обязанности получателя, включая юридические и договорные обязательства;

g) требования к среде использования; обычно охватывает среды тестирования, разработки и эксплуатации системы (см. этап 8 в ISO/IEC 5259-1:—, рисунок 2);

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

i) доступные методы приемочных испытаний и их ожидаемые результаты;

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

7.3.7.4 Отслеживание и улучшение

Организация должна оценить, есть ли требование отслеживать использование и результаты использования данных. Если такие требования существуют или организация решает добровольно внедрить отслеживание, организация должна внедрить соответствующие процессы для выполнения требований по отслеживанию и улучшению.

В случае реализации отслеживания использования данных организация должна оценить необходимость:

a) обновить задокументированный объем и предполагаемое использование данных;

б) улучшить данные;

c) обновить систему управления качеством данных.

Организация должна создать канал для сбора добровольных отзывов заинтересованных сторон о данных и установить процессы для оценки отзывов для постоянного улучшения данных и системы управления качеством данных.

7.3.7.5 Оптимизация данных

Организация должна документировать способы, применяемые для оптимизации данных:

a) оптимизация пространства для хранения;

б) оптимизация доступа;

c) оптимизация требований к сетевым ресурсам;

d) оптимизация обработки данных (например, буферизация);

e) совместимость с интерфейсами передачи данных.

7.3.7.6 Обязательство поддержки

Уровень предоставляемой поддержки должен соответствовать требованиям к качеству данных и предполагаемому использованию данных. Должны быть задокументированы следующие пункты:

a) необходимые навыки или обучение для использования данных;

b) документация о предоставленной поддержке, включая объем, период и доступность;

c) план технического обслуживания и обязательства по постоянному улучшению данных;

d) доступный канал связи с заинтересованными сторонами данных;

e) план распределения ресурсов и управления.

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 посредством передачи данных.

Организация должна определить процессы вывода данных из эксплуатации для выполнения всех соответствующих требований (см. пункты 7.3.8.2–7.3.8.4) и включить их в план данных.

7.3.8.2 Передача данных

В случае передачи данных организация должна определить получателя данных, который имеет право принять данные и все связанные с этим юридические и договорные обязательства.

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

Организация должна документировать выраженное получателем согласие принять данные и все связанные с ними обязательства.

Организация должна создать и независимо утвердить отчет о передаче данных.

7.3.8.3 Удаление данных

Организация должна документально подтвердить, что удаление данных не нарушает никаких юридических, договорных обязательств и обязательств по хранению данных.

Организация должна гарантировать, что нет ни одного пользователя данных, которому по-прежнему требуется доступ к данным. В качестве Пользователей данных могут выступать системы, которым требуется доступ к данным для обслуживания и переобучения.

В случае закрытия организация должна удалить данные со всех мест хранения.

Организация должна предусмотреть возможность частичного или полного восстановления данных из моделей машинного обучения, обученных на ее данных. Организация должна учитывать такие случаи в своих обязательствах и плане удаления данных.

Организация должна составить и независимо подтвердить отчет об удалении данных.

Организации следует проверить, имеют ли данные культурное, историческое или общественное значение. В случае, если данные имеют значение, организация должна документировать усилия по передаче данных получателю, который может выполнить все обязательства по данным и понимая их значимость.

Организации следует рассмотреть возможность и целесообразность передачи данных в общественное достояние.

ПРИМЕЧАНИЕ 1. Места хранения данных включают резервные копии, которые могут быть автоматически созданы ИТ-инфраструктурой.

ПРИМЕЧАНИЕ 2. Для организации является хорошей практикой предоставление аргументов, почему передача данных наносит ущерб ее интересам, чтобы обосновать выбор организации в отношении удаления данных.

ПРИМЕЧАНИЕ 3. Предоставление данных в общественное достояние потенциально приносит пользу исследованиям и образованию и, косвенно, самой организации.

7.3.8.4 Частичное удаление данных по запросу

В организации должны быть предусмотрены соответствующие процессы для обработки запросов на частичное удаление данных.

В случае удаления части данных организация должна оценить, соответствуют ли оставшиеся данные ее спецификации. В случае, если данные не соответствуют спецификации, организация должна уведомить все заинтересованные стороны о несоответствии качества и, если возможно, принять соответствующие меры. В случае невозможности реализации соответствующего смягчения, организация должна вывести из эксплуатации все данные, что может включать передачу оставшихся данных.

В случае частичного удаления данных организация должна удалить частичные данные со всех мест хранения.

Организация должна создать и независимо утвердить отчет о частичном удалении данных.

ПРИМЕЧАНИЕ Правила, касающиеся данных, такие как законы о защите конфиденциальности, могут вынудить организацию удалить части набора данных.

7.4 Артефакты

7.4.1 Артефакты этапа мотивации данных и концептуализации

Артефакты этапа мотивации и концептуализации данных включают:

а) намерение документирования данных;

б) общественно-этическая экспертиза;

в) технико-экономическое обоснование;

г) анализ заинтересованных сторон;

е) свидетельство того, что верификация и валидация на этапе мотивации данных и концептуализации были успешно завершены.

7.4.2 Артефакты этапа спецификации данных

Артефакты этапа спецификации данных включают:

a) формат данных и спецификацию использования;

b) спецификация вспомогательных инструментов и связанных с ними требований;

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

7.4.3 Артефакты этапа планирования данных

Артефакты этапа планирования данных включают:

a) план данных;

b) свидетельство того, что верификация и валидация на этапе планирования данных были успешно завершены.

7.4.4 Артефакты этапа комплектования данных

Артефакты этапа комплектования данных включают:

a) собранные данные и сопровождающая их документация;

b) спецификация режима сбора данных и соответствующая ему конфигурация;

c) инфраструктура для управления данными;

г) проверка качества данных;

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

7.4.5 Артефакты этапа предварительной обработки данных

Артефакты этапа предварительной обработки данных включают:

a) очищенные данные;

б) проверка качества данных;

c) описательная статистика, описывающая свойства данных;

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

7.4.6 Артефакты этапа расширения данных

Артефакты этапа расширения данных включают:

a) расширенные данные, включая метаданные и метки;

б) описание характеристик;

c) спецификация применяемых инструментов и методов обогащения;

d) свидетельство того, что верификация и валидация на этапе достоверности данных завершены успешно.

7.4.7 Артефакты этапа предоставления данных

Артефакты этапа предоставления данных включают:

a) предоставленные элементы, включая данные, выборку данных и документацию;

b) подтверждение того, что верификация и валидация на этапе предоставления данных были успешно завершены.

7.4.8 Артефакты этапа вывода данных из эксплуатации

Артефакты этапа вывода данных из эксплуатации включают:

a) отчет о выводе данных из эксплуатации;

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

8 Горизонтальные процессы

8.1 Цель

Целью горизонтальных процессов управления качеством данных является консолидация действий, применимых к каждому этапу жизненного цикла управления качеством данных.

8.2 Общие положения

Горизонтальные процессы должны применяться на каждом этапе жизненного цикла с входными и выходными данными для конкретных этапов жизненного цикла.

8.3 Требования и рекомендации

8.3.1 Верификация и валидация

8.3.1.1 Общие положения

Процессы верификации и валидации отслеживают и оценивают качество данных на всех этапах жизненного цикла управления качеством данных. Верификация устанавливает объективные доказательства того, что требования к качеству данных были соблюдены. Валидация устанавливает, что данные соответствуют намеченным целям.

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

8.3.1.2 Контрольные точки жизненного цикла качества данных

На каждом этапе жизненного цикла качества данных должны быть указаны объективно поддающиеся оценке целевые показатели качества артефактов. Верификация и валидация должны подтвердить, что все требования на каждом этапе жизненного цикла качества данных выполнены. Стадия жизненного цикла данных должна вводиться только с входными данными для которых выполнена верификация и валидация. Результаты верификации и валидации должны быть задокументированы.

8.3.1.3 Улучшение

После успешного прохождения верификации и валидации артефакта результаты, полученные при их регламентированном использовании должны быть задокументированы и оценены. Несоответствие артефакта спецификации предполагаемому использованию должно привести к обновлению 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.

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

Менеджмент конфигурации должен установить требования, применяемые к качеству данных для каждой допустимой конфигурации.

Управление конфигурацией должно установить, что все допустимые конфигурации соответствуют всем применимым требованиям к качеству данных.

8.3.3 Управление изменениями

8.3.3.1 Общие положения

Очевидно, что изменения процессов и требований к качеству данных будут часто происходить во время разработки систем ИИ и поддерживаться в жизненном цикле управления качеством данных итерациями. Данные и связанные с ними метаданные часто подвергаются изменениям, в основном на этапах спецификации и реализации и в меньшей степени в результате процессов проверки и проверки.

8.3.3.2 Планирование управления изменениями

Управление изменениями должно:

a) планироваться и начинаться до любых изменений;

b) включать план, который идентифицирует элементы, подлежащие изменению, и определяет график изменений;

c) определить процесс, который включает:

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

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

3) оценка запроса на изменение (см. пункт 8.3.3.5);

4) документацию (см. 8.3.3.6).

8.3.3.3 Спецификация запроса на изменение

Запросы на изменение должны иметь:

а) уникальный идентификатор;

б) дату;

с) причину, описание и конфигурацию запрашиваемого изменения;

d) решение и обоснование запрошенного изменения.

8.3.3.4 Анализ запроса на изменение

Влияние запроса на изменение должно быть проанализировано с учетом:

а) его типа (ошибка, расширения, добавление данных или меток и т. д.);

b) затрагиваемые элементы (данные, метки, требования и т. д.);

с) затрагиваемые стороны, включая их обязанности;

г) влияние на качество;

е) влияние на ранее существовавшие данные или метки;

е) график.

8.3.3.5 Оценка запроса на изменение

Запрос на изменение и анализ его воздействия должны быть оценены для принятия решения:

а) о его статусе (принят, отклонен и т.п.) уполномоченным лицом;

б) лице, которое будет осуществлять изменения;

в) временных рамках.

8.3.3.6 Внедрение и документирование изменений

Изменения должны быть реализованы и проверены в соответствии с планом.

Изменения данных, их качества, использование или проверки должны инициировать обновление выполненных ранее оценок и обзоров.

Документация по изменениям должна включать:

а) список измененных элементов, включая конфигурацию и версии;

б) детали выполненных изменений;

в) дату вступления изменений в силу.

8.4 Артефакты

8.4.1 Артефакты верификации и валидации

Артефакты верификации и валидации включают:

a) документирование результатов верификации и валидации на всех стадиях жизненного цикла;

b) документирование проблем и ограничений верификации и валидации;

c) предложения по улучшению.

8.4.2 Артефакты управления конфигурацией

Артефакты управления конфигурацией включают:

a) документированный план управления конфигурацией;

b) документированную стратегию конфигурации;

c) документацию по допустимым конфигурациям и связанными с ними требованиями к качеству.

8.4.3 Артефакты управления изменениями

Артефакты управления изменениями включают:

а) план управления изменениями;

б) запросы на изменение;

в) анализ воздействия;

г) отчеты об изменениях.

9 Управление качеством данных в цепочках поставок

9.1 Цель

Целью требований к управлению цепочками поставок является обеспечение того, чтобы любые выбранные поставщики предоставляли данные, отвечающие требованиям организации. Требования этого документа также применяются к любым поставщикам, которые не предоставляют данные, но предоставляют услуги по обработке данных (например, вычислительные услуги, хранение, архивирование, транспортные услуги).

9.2 Требования и рекомендации

При выборе поставщика необходимо учитывать возможности поставщика по предоставлению данных качество которых соответствует разделам 6–8 настоящего документа.

Требования к запросу цен:

a) Запрос цен поставщику должен содержать:

1) требования к соблюдению настоящего документа;

2) соответствующие спецификации данных;

3) соответствующие требования к качеству.

b) Пользователь и поставщик в соглашении о взаимодействии должны указать:

1) назначение менеджеров по качеству от пользователя и поставщика;

2) соответствие настоящему документу (см. пункт 5);

3) действия в рамках жизненного цикла данных, которые должны выполняться каждой стороной;

4) совместно используемая информация и рабочие продукты;

5) обязанности каждой стороны по каждому виду деятельности;

6) требования к качественным и количественным характеристикам (см. пункт b) пункта 6.3.10);

7) связанные с интерфейсом процессы, методы и инструменты;

8) деятельность по оценке качества;

9) планирование отчета об оценке качества поставщика;

10) соглашение, которое позволяет аудитору, назначенному пользователем, получать доступ ко всем необходимым ресурсам для целей аудита качества;

11) обязанности и действия во время эксплуатации и вывода из эксплуатации (см. Этапы 8–10 в ИСО/МЭК 5259-1:—, рисунок 2);

12) требования к передаче информации, проблемы и аномалии, влияющие на качество данных или риск нарушения соглашения об интерфейсе разработки.

9.3 Артефакты

Артефакты управления качеством данных в цепочках поставок должны включать:

а) отчет о выборе поставщика;

б) соглашение об интерфейсе разработки;

в) отчет об оценке качества.

10 Управление средствами обработки данных

10.1 Цель

Цель управления инструментами обработки данных состоит в том, чтобы гарантировать, что данные соответствуют требованиям организации всякий раз, когда инструмент обработки данных используется или применяется к данным в течение жизненного цикла данных.

10.2 Требования и рекомендации

Организация должна учитывать влияние на качество данных, по крайней мере, для следующих категорий средств обработки данных:

a) комплектование данных (например, веб-скраперы, снимки экрана, анкеты, датчики);

b) предварительная обработка данных (например, очистка данных, преобразование данных);

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

г) маркировка данных;

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

f) передача данных (например, сети, шины данных);

g) безопасность (например, брандмауэры, шифрование).

10.3 Артефакты

Артефакты управления средствами обработки данных должны включать:

а) документацию по любым используемым инструментам обработки данных;

б) документирование любого воздействия средств обработки данных на качество данных.

11 Управление зависимостями качества данных

11.1 Цель

Целью управления зависимостями качества данных является обеспечение того, чтобы зависимости не нарушали требования организации к данным. Зависимости могут быть внутренними (например, обработка данных, выполняемая под непосредственным контролем организации) или внешними (например, облачные службы, маркировка данных по контракту).

11.2 Требования и рекомендации

Организация должна рассмотреть как минимум следующие вопросы:

а) любая внутренняя зависимость, которая может повлиять на качество данных.

б) любая внешняя зависимость, которая может повлиять на качество данных;

с) любые меры по сокращению последствий, которые организация может предпринять для устранения влияния любой зависимости на качество данных.

11.3 Артефакты

Артефакты для управления зависимостью качества данных должны включать:

а) документирование всех внутренних зависимостей и их потенциального влияния на качество данных;

б) документирование всех внешних зависимостей и их потенциального влияния на качество данных;

с) документацию по смягчению последствий, использованных для обеспечения соответствия данных требованиям организации.

12 Управление качеством данных по конкретному проекту

12.1 Цель

Целью управления качеством данных по конкретному проекту для учреждения является обеспечение того, чтобы требования к качеству данных по конкретному проекту учитывались способом, подходящим и достаточным для этого проекта.

12.2 Требования и рекомендации

12.2.1 Контекст и предполагаемое использование

Контекст требований и рекомендаций по управлению качеством данных — это проект аналитики и машинного обучения для конкретной системы искусственного интеллекта, как указано организацией. Проект аналитики и машинного обучения может состоять из нескольких наборов данных, алгоритма машинного обучения или модели машинного обучения. Требования и рекомендации по управлению качеством данных для конкретных проектов касаются зависящего от системы ИИ компонента качества данных.

12.2.1.1 Цель

Цель этих требований и рекомендаций — гарантировать, что данные, используемые в указанном проекте аналитики и машинного обучения, соответствуют требованиям.

12.2.1.2 Требования и рекомендации

Организация должна:

- указать масштаб проекта;
- документировать мотивы использования данных;
- задокументировать условия использования данных;
- задокументировать предполагаемое использование.

Организация должна документировать известные ограничения из-за известных несовместимостей или конфликтов:

- юридические;
- договорные соглашения;
- период обслуживания и продолжительность использования;
- технические (например, алгоритмы, библиотеки);
- область использования;
- рыночные условия;
- неудачное использование данных.

Список известных ограничений должен быть обновлен и сообщен, если и когда станут известны новые ограничения.

Модель МО обычно взаимодействует с несколькими другими компонентами системной архитектуры, например, пользовательским интерфейсом, базой данных и другими пользователями результатов модели МО. Если такие интерфейсы 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

12.3.2 Требования и рекомендации

Организация должна предоставить документацию по требованиям и рекомендациям по качеству данных. Организация должна предоставить рекомендации по формулировке требований к качеству данных, таких как естественный язык, онтологии или математические спецификации. Такие требования должны быть выражены точно и недвусмысленно, чтобы обеспечить их проверяемость.

Требования к качеству данных должны:

а) быть идентифицированным как таковые;

б) присваиваться удовлетворяющим им данным;

с) быть частью управления конфигурацией.

г) быть проверяемыми.

Требования к качеству данных должны иметь уникальный идентификатор, статус и быть прослеживаемыми на данных.

Проверка может осуществляться различными способами, включая, но не ограничиваясь:

- проверка экспертами в предметной области;
- проверка привлеченными разработчиками;
- формальная проверка (если применимо).

12.4 Управление рисками

12.4.1 Цель

Цель управления рисками состоит в том, чтобы обеспечить учет, оценку и, при необходимости, снижение всех рисков для качества данных.

12.4.2 Требования и рекомендации

Организация должна проводить оценку рисков и вести документацию в отношении рисков, связанных с качеством данных. Организация должна управлять и поддерживать соответствующие процессы для минимизации выявленных рисков.

Класс риска данных должен быть указан, обоснован и задокументирован как метка риска в контексте предполагаемого использования системы ИИ. Выбор соответствующих показателей качества данных должен обновляться и управляться системой управления качеством данных на основе связанного риска. Организация должна сообщать о рисках, связанных с данными, соответствующим заинтересованным сторонам.

ПРИМЕЧАНИЕ. Чтобы определить процессы управления рисками, подходящие для ИИ, организации могут рассмотреть ИСО/МЭК 23894:—1) [5].

12.4.3 Артефакты

Артефакты оценки риска должны включать:

а) документацию по рискам качества данных;

б) план обработки риска.

12.5 Роли и обязанности в управлении качеством данных

12.5.1 Цель

Цель управления ролями и обязанностями по управлению качеством данных состоит в том, чтобы гарантировать, что соответствующие роли качества данных сформулированы, и каждой роли назначены необходимые обязанности. Персонал, назначенный на данные роли, должен иметь соответствующую профессиональную квалификацию и иметь необходимые ресурсы для выполнения своих обязанностей.

Цель управления ролями и обязанностями по управлению качеством данных состоит в том, чтобы гарантировать, что соответствующие роли качества данных сформулированы, и каждой роли назначены необходимые обязанности. Персонал, назначенный на данные роли, должен иметь соответствующую профессиональную квалификацию и иметь необходимые ресурсы для выполнения своих обязанностей.

12.5.2 Требования и рекомендации

Организация должна назначить:

а) менеджера проекта в начале проекта, обладающего необходимыми полномочиями для обеспечения того, чтобы:

1) выполнять обязанности по обеспечению качества;

2) достигнуть соответствия этому документу;

3) назначить менеджера по качеству;

б) менеджера по качеству.

12.5.3 Артефакты

Артефакты для ролей и обязанностей в управлении качеством данных должны включать:

а) задокументированные описания ролей для управления качеством данных, включая:

1) руководителя проекта;

2) менеджера по качеству.

b) документированные обязанности для каждой роли управления качеством данных.

12.6 Адаптация мероприятий по обеспечению качества данных

Мероприятия по обеспечению качества в отношении жизненного цикла данных, используемые в этом документе, могут быть адаптированы, если они документированы в плане качества и дано их обоснование.

12.7 Планирование и координация деятельности по обеспечению качества данных

12.7.1 Общие положения

Деятельность по управлению качеством данных должна включать:

а) планирование и координацию мероприятий по обеспечению качества;

б) обеспечение выполнения плана качества данных;

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

d) распределение обязанностей по обеспечению качества и информирование об ответственности за их выполнение.

12.7.2 План качества данных

Все вовлеченные организации должны поддерживать план качества данных. План качества данных должен:

а) быть непосредственно включен или упомянут в плане проекта. Деятельность в области качества должна быть отмечена как таковая;

б) содержать перечень мероприятий по достижению качества данных;

c) регулярно пересматриваться и обновляться, включая все артефакты.

12.7.3 Планирование процессов

Планирование деятельности по обеспечению качества должно включать:

а) критерии качества данных;

б) зависимости от других процессов или информации;

в) лицо, ответственное за выполнение процесса;

г) необходимые ресурсы для выполнения процесса;

е) начальная или конечная точка во времени и продолжительность;

f) идентификацию соответствующего артефакта.

12.8 Развитие жизненного цикла качества данных

Действия по управлению качеством данных должны следовать жизненному циклу качества данных, используемому организацией:

а) Стадии следует начинать только в том случае, если недостающая информация на предыдущих стадиях наложенных рисков обрабатывается в рамках деятельности по управлению рисками организации.

б) Рабочие продукты плана качества должны быть частью управления конфигурацией и изменениями.

12.9 Обоснование качества данных

При необходимости организация должна представить убедительное обоснование заявленного качества данных. Обоснование качества данных должно:

a) предоставить аргументацию для достижения качества данных;

b) содержать все артефакты, произведенные в течение жизненного цикла качества данных.

12.10 Выпуск

Перед выпуском продукта или услуги, для которых осуществляется управление качеством данных, организация должна гарантировать, что:

a) обоснование качества данных (если применяется), включая меры по его подтверждению, завершено до этапа развертывания системы (см. этап 7 в ИСО/МЭК 5259-1:—, рисунок 2);

b) элементы предоставления данных, перечисленные в пункте 7.3.7.2, предоставляются (см. этап 7 на рисунке 1) при условии наличия достаточных доказательств их качества.

12.11 Артефакты

Артефакты управления качеством данных для конкретного проекта должны включать:

a) требования к качеству данных;

б) план качества;

c) обоснование качества (если применимо);

d) отчеты о мерах подтверждения;

e) отчет о выпуске.

Библиография

[1]	ISO/IEC TR 4027:2021	Information technology Artificial Intelligence (AI) Bias in AI systems and AI-aided decision making
[2]	ISO/IEC TR 24368:—	Information technology — Artificial intelligence — Overview of ethical and societal concerns
[3]	ISO/IEC 42001:—	Information technology — Artificial intelligence — Management system
[4]	ISO 9001:2015	Quality management systems — Requirements
[5]	ISO 9001:2015	Quality management systems — Requirements

Приложение А
(справочное)
Сведения о соответствии ссылочных международных стандартов национальным стандартам

Таблица А

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	NEQ*	ПНСТ 553–2021** Информационные технологии. Искусственный интеллект. Концепции искусственного интеллекта и терминология (ISO/IEC DIS 22989, NEQ)
*NEQ — неэквивалентный стандарт. ** В России принят предварительный национальный стандарт, основанный на ранней редакции международного стандарта		

ГОСТ (Р) _____—_____.

(проект)

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: информационные технологии, искусственный интеллект, качество данных, характеристики качества данных, большие данные, набор данных, подготовка наборов данных, вывод данных из эксплуатации, управление качеством данных, аномалии качества данных, интеграция системы управления качеством данных, жизненный цикл управления качеством данных, свойства данных, комплектование данных, артефакты управления качеством данных, контрольные точки жизненного цикла качества данных, верификация и валидация данных, управление средствами обработки данных, роли и обязанности в управлении качеством данных, план качества данных.

Председатель совета директоров ООО «Институт развития информационного общества», руководитель разработки стандарта		Ю.Е. Хохлов
Руководитель департамента развития технологий больших данных ООО «Институт развития информационного общества» профессор, Санкт-Петербургский государственный университет, разработчик стандарта		Н.Л. Щеголева