ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 59897— 2021

ДАННЫЕ ДЛЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ОБРАЗОВАНИИ

Требования к сбору, хранению, обработке, передаче и защите данных

Издание официальное

Москва
Российский институт стандартизации
2021

Предисловие

1 РАЗРАБОТАН Федеральным государственным автономным образовательным учреждением высшего образования «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ)

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 26 ноября 2021 г. № 1619-ст

4 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)

Содержание

1 Область применения 1
2 Нормативные ссылки 1
3 Термины и определения 1
4 Общие требования 2
5 Структура образовательной деятельности и модель данных 2
6 Источники данных 3
7 Жизненный цикл данных 3
8 Требования к сбору данных 4
9 Требования к хранению данных 5
10 Требования к обработке данных 5
11 Требования к передаче данных 6
12 Требования к защите данных 6
Приложение А (справочное) Примеры полей метаданных 7
Библиография 8

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ДАННЫЕ ДЛЯ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ОБРАЗОВАНИИ

Требования к сбору, хранению, обработке, передаче и защите данных

Data for artificial intelligence systems in education. Requirements for the collection, storage, processing, transmission and protection of data

Дата введения — 2022—03—01

1 Область применения

Настоящий стандарт устанавливает требования к процессам сбора, хранения, обработки, передачи и защиты данных, используемых в образовательных программно-технических системах с алгоритмами искусственного интеллекта.

2 Нормативные ссылки

В настоящем стандарте использована ссылка на следующий стандарт:
ГОСТ Р ИСО 8000-2 Качество данных. Часть 2. Словарь

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 ссылка, внесено изменение, затрагивающее 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 термины по ГОСТ Р ИСО 8000-2, а также следующие термины с соответствующими определениями:

3.1 **образовательный продукт с алгоритмами искусственного интеллекта:** Программно-техническая система, использующая алгоритмы искусственного интеллекта для решения различных задач в области образования.

3.2 **жизненный цикл данных** (data lifecycle): Последовательность этапов, через которые проходят данные от начального этапа формирования до момента уничтожения.

3.3 **основные данные [мастер-данные** (master-data)**]:** Данные, описывающие основные объекты и субъекты, участвующие в образовательной деятельности.

Примечания

1 Помимо информации непосредственно о том или ином объекте, в основные данные входят взаимосвязи между этими объектами и субъектами и иерархии.

2 Каждая организация определяет самостоятельно, какие данные следует считать основными.

Издание официальное

3.4 **транзакционные данные:** Данные, которые образовались в результате выполнения каких-либо операций.

3.5 **очистка данных** (data cleansing): Процесс исправления или удаления неверных, поврежденных, неправильно отформатированных, дублированных или неполных данных в наборе данных.

3.6

персональные данные: Любая информация, прямо или косвенно относящаяся к определенному или определяемому физическому лицу (субъекту персональных данных).

[1, статья 3]

3.7

репозиторий: Место, где хранятся и поддерживаются какие-либо данные вместе с историей их изменения и другой служебной информацией.

[ГОСТ Р 57723—2017, статья 3.1.21]

3.8 **системы управления учебной деятельностью** (LMS-системы): Программно-технические системы для организации учебного процесса и управления образовательными материалами.

4 Общие требования

При организации сбора, хранения, обработки, передачи и защиты данных в образовательных продуктах с алгоритмами искусственного интеллекта:

а) сбор, хранение, обработка и передача персональных данных может осуществляться только с согласия пользователей образовательного продукта. Рекомендуется предусмотреть возможность пользователя ознакомиться с собираемыми о нем данными;

б) должны быть идентифицированы все заинтересованные стороны или их представители, на которых может быть оказано влияние в результате использования данных, определены их интересы и связанные с ними риски;

в) предприняты необходимые действия для минимизации выявленных рисков.

5 Структура образовательной деятельности и модель данных

5.1 Образовательную деятельность с использованием образовательных продуктов с алгоритмами искусственного интеллекта можно представить в виде последовательности действий по планированию, осуществлению и оценке деятельности и ее результатов (см. рисунок 1).

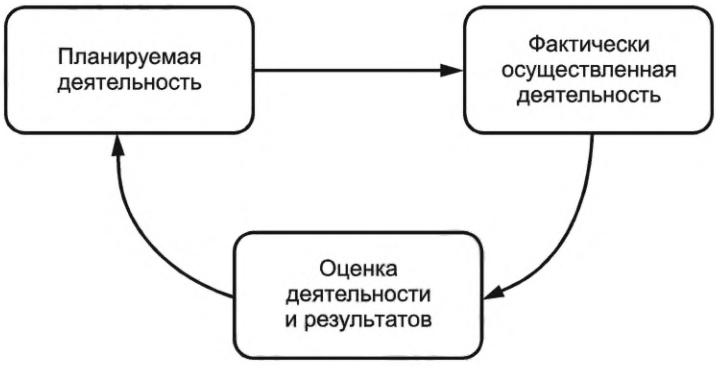

Рисунок 1 — Этапы образовательной деятельности

5.2 Фактически осуществленная образовательная деятельность может отличаться от запланированной в силу объективных и субъективных обстоятельств.

5.3 Модель данных включает в себя основные и транзакционные данные, описывающие участников образовательной деятельности, их планируемую и фактическую деятельность, оценку деятельности обучающегося и результатов обучения.

5.4 Данные о планируемой деятельности включают в себя основные данные о структуре, содержании, планируемых результатах, организационно-педагогических условиях их реализации.

5.5 Выделяют три уровня данных о фактической образовательной деятельности:

- уровень деятельности (например, прохождение образовательной программы или курса в целом);
- уровень отдельного действия, имеющего смысл с точки зрения обучения (например, выполнение отдельной задачи в рамках курса, ответ на отдельный вопрос, просмотр видео и т.п.);
- уровень операции, не имеющей самостоятельного смысла с точки зрения обучения (например, взаимодействие с алгоритмами интерфейса в информационной системе).

5.6 В образовательную деятельность с использованием образовательных продуктов с алгоритмами искусственного интеллекта вовлечены обучающиеся и педагоги, а также инструменты и образовательные материалы продукта. Таким образом данные о действиях в процессе обучения могут относиться к деятельности обучающихся, педагогов, а также использованию образовательных модулей и элементов.

5.7 Отдельные элементы фактически осуществленной деятельности обучающегося описаны следующими параметрами:

участник образовательной деятельности;
цель деятельности;
инструменты, используемые в деятельности;
обстоятельства и условия деятельности;
предметная область деятельности;
состояние участника в момент деятельности;
результаты деятельности: образовательные и предметные (артефакты);
роли участника деятельности (в коллективной деятельности);
продемонстрированные или освоенные знания, умения, навыки.

Данные параметры также могут быть использованы для описания отдельных аспектов планируемой деятельности.

Примечание — Участник деятельности может быть представлен отдельным физическим лицом или группой лиц.

6 Источники данных

6.1 Для эффективного функционирования образовательных продуктов с алгоритмами искусственного интеллекта необходимо обеспечить сбор и использование данных об участниках и объектах образовательного процесса из разнородных источников.

6.2 Источники данных можно разделить на внутренние и внешние.

6.3 Внутренними источниками данных являются отдельные модули и системы, входящие в состав образовательных продуктов с алгоритмами искусственного интеллекта. Внутренние источники обеспечивают сбор транзакционных данных о фактически осуществленной образовательной деятельности. Для успешного использования технологий искусственного интеллекта внутренние источники данных должны обеспечивать сбор разнородных данных, включая видео и аудиоданные, изображения, текстовые и неструктурированные данные.

6.4 Внешними источниками данных являются иные информационные системы, цифровые платформы и технические устройства. Источником внешних данных, необходимых для образовательных продуктов с алгоритмами искусственного интеллекта могут выступать:

- информационные системы образовательных организаций, в которых используются образовательные продукты с алгоритмами искусственного интеллекта;
- государственные информационные системы;
- информационные системы иных организаций, а также отраслевые информационные системы;
- цифровые социальные платформы.

6.5 Для функционирования образовательных продуктов с алгоритмами искусственного интеллекта необходимо использовать только верифицированные источники данных, гарантирующие достоверность, правильность и точность предоставляемой информации.

7 Жизненный цикл данных

7.1 Жизненный цикл данных включает следующие этапы: сбор, хранение, обработка, использование, архивация и уничтожение данных.

7.2 Сбор данных представляет собой формирование новых данных, получаемых от источников данных.

7.3 Хранение данных представляет собой обеспечение сохранности и доступа к данным.

7.4 Обработка данных представляет собой манипуляции с данными на протяжении всего жизненного цикла, обеспечивающие их качество. На этапе обработки данные могут быть очищены, преобразованы, подвергнуты слиянию, улучшены или агрегированы.

7.5 Использование данных представляет собой применение данных для решения различных задач.

7.6 Архивация данных представляет собой копирование данных в специализированную систему (среду), в которой будет обеспечено их долгосрочное хранение, и удаление этих данных из активной системы. Архивация обеспечивает возможность повторного использования исторических данных, если они понадобятся вновь.

7.7 Уничтожение данных представляет собой необратимое удаление данных, исключающее их использование и восстановление.

7.8 Допускается возможность перехода данных на новый жизненный цикл при их модификации и обогащении новыми данными.

8 Требования к сбору данных

8.1 До проведения мероприятий по сбору данных для использования в образовательных продуктах с алгоритмами искусственного интеллекта необходимо определить:

- цели и задачи, являющиеся основанием для сбора данных;
- перечень и объем собираемых данных;
- методы сбора данных.

Также целесообразно определить гипотезу, которая может быть подтверждена или опровергнута в ходе исследования собираемых данных.

8.2 Для эффективного функционирования образовательных продуктов с алгоритмами искусственного интеллекта необходимо установить уровень качества собираемых данных и соответствующие требования к его определению.

8.3 Качество данных оценивают по следующим критериям:

а) точность — соответствие данных реальному состоянию исследуемых объектов;

б) полнота — данные отражают все ожидаемые характеристики исследуемых объектов в ожидаемом объеме;

в) согласованность — в данных отсутствуют внутренние противоречия, идентичные данные из различных источников совпадают;

г) целостность — данные не были изменены при выполнении какой-либо операции (передача, хранение или отображение);

д) обоснованность — собранные данные отвечают поставленным целям и задачам;

е) расхождение во времени — соответствие собираемых данных времени их возникновения;

ж) уникальность — в данных отсутствуют дубликаты;

и) валидность — данные соответствуют ожидаемому формату, значения находятся в ожидаемых диапазонах и имеют ожидаемую точность.

8.4 Если данные не соответствуют установленному уровню качества, необходимо провести мероприятия по его повышению.

8.5 Для удобства дальнейшего использования данных, в том числе для их последующей очистки, по результатам оценки качества данных может быть сформирован соответствующий отчет.

8.6 Для собираемых данных необходимо предварительно определить основные параметры жизненного цикла данных, включая продолжительность хранения, сроки архивации и уничтожения данных.

8.7 Сбор данных для использования в образовательных продуктах с алгоритмами искусственного интеллекта может проходить в несколько этапов:

- подготовительный этап, на котором данные собирают для настройки и тестирования моделей машинного обучения (элементов искусственного интеллекта), используемых в образовательном продукте;
- формирующий этап — первичный сбор данных об участнике образовательной деятельности, осуществляемый при его первоначальном обращении к образовательному продукту;
- реализующий этап — текущий сбор данных, осуществляемый регулярно для реализации функций адаптивной обучающей системы (например, сбор цифрового следа обучающегося).

8.8 Для наиболее эффективного функционирования образовательных продуктов с алгоритмами искусственного интеллекта необходимо обеспечить сбор данных обо всех действиях участников образовательного процесса.

9 Требования к хранению данных

9.1 Хранимые данные должны иметь определенный набор метаданных. Метаданные подразделяются на три основных категории:

- описательные метаданные, описывающие содержание и состояние данных;
- технические метаданные, описывающие технические подробности хранения данных;
- операционные метаданные описывают процессы обработки данных и доступа к ним.

Примеры полей метаданных приведены в приложении А.

9.2 Для управления метаданными необходимо установить требования к:

- частоте обновления метаданных;
- необходимости хранения исторических метаданных;
- правам доступа к метаданным;
- степени интеграции метаданных из различных источников;
- процессам и правилам обновления метаданных;
- ролям и обязанностям по управлению метаданными;
- качеству метаданных.

9.3 Должно быть реализовано управление основными данными, включающее следующие мероприятия:

- установление сущностей и атрибутов основных данных;
- создание идентификаторов и перекрестных ссылок для интеграции данных из разных источников;
- объединение данных из различных источников для устранения несоответствий;
- дополнение и обновление репозитория основных данных.

9.4 Для обеспечения возможности замены программного обеспечения, осуществляющего обработку и анализ данных, рекомендуется использовать открытые форматы хранения данных.

10 Требования к обработке данных

10.1 Если собранные данные являются первичными, то сначала необходимо провести их очистку.

10.2 Данные после очистки и достижения установленного уровня качества форматируют в наборы данных для дальнейшего использования в образовательных продуктах с алгоритмами искусственного интеллекта.

Примечание — На данном этапе могут возникнуть производные атрибуты или новые записи, а также данные, интегрированные из других источников.

10.3 Для каждой отдельной разновидности организации данных должен быть определен формат набора данных.

10.4 Поскольку качество данных может пострадать на любом этапе жизненного цикла, необходимо планировать меры по обеспечению качества данных в расчете на весь жизненный цикл данных.

10.5 Для корректного функционирования образовательных продуктов с алгоритмами искусственного интеллекта необходимо проводить мероприятия по повышению качества данных.

10.6 Мероприятия по повышению качества данных разделяют на две категории: предупредительные и корректирующие.

Примечания

1 Примеры предупредительных мероприятий по повышению качества данных: проверка данных на несоответствие на входе, повышение квалификации сотрудников, ответственных за сбор данных, определение правил в части качества данных, использование источников с высококачественными данными и определение должностных лиц, ответственных за качество данных.

2 Примеры корректирующих мероприятий по повышению качества данных: автоматическое исправление данных по известным шаблонам, исправление автоматизированными инструментами с ручной проверкой и ручное исправление.

11 Требования к передаче данных

11.1 Для организации взаимодействия образовательных продуктов с алгоритмами искусственного интеллекта с другими образовательными продуктами и иными информационными системами, в том числе с системами управления учебной деятельностью (LMS-системами), возможна организация разовой или регулярной передачи данных.

11.2 Данные могут передаваться как в исходном, так и в обезличенном формате.

11.3 Данные, собираемые и хранимые в образовательных продуктах с алгоритмами искусственного интеллекта, необходимо передавать в формате набора данных с соответствующим набором метаданных.

11.4 Должна быть обеспечена надежность передачи данных, уменьшающая риск снижения качества данных в процессе передачи, а также исключающая копирование данных в другие системы.

12 Требования к защите данных

12.1 Права доступа к данным, содержащимся в образовательных продуктах с алгоритмами искусственного интеллекта, должны быть описаны для каждой целевой группы пользователей.

12.2 Для каждого набора данных, а также метаданных, хранящихся в образовательных продуктах с алгоритмами искусственного интеллекта, на уровне метаданных должен быть установлен уровень конфиденциальности.

12.3 Данные, содержащиеся в образовательных продуктах с алгоритмами искусственного интеллекта, должны быть защищены от потери (например, с помощью резервного копирования или реплицирования).

Приложение А
(справочное)

Примеры полей метаданных

А.1 Примеры описательных метаданных

Примерами описательных метаданных служат:
- определения и описания данных, сущностей, атрибутов;
- правила использования данных;
- уровень качества данных;
- расписание обновления данных;
- происхождение данных;
- допустимые ограничения значений;
- необходимая контактная информация;
- уровень конфиденциальности данных;
- известные проблемы с данными;
- примечания.

А.2 Примеры описательных метаданных для образовательных данных

Примерами описательных метаданных для образовательных данных служат:
- контекст сбора данных (название образовательной программы, учебного курса, мероприятия);
- характер данных (описание и план деятельности, фактически осуществленная деятельность, оценка деятельности);
- источник данных (учащийся, фасилитатор);
- вид данных (данные о характеристиках учащегося или фасилитатора, данные о деятельности);
- уровень данных о деятельности (деятельность, действие, операция).

А.3 Примеры технических метаданных:

Примерами технических метаданных служат:
- имена таблиц и столбцов баз данных и их свойства;
- права доступа к данным, группы и роли;
- правила создания, замены, обновления и удаления данных;
- физические модели данных (имена таблиц, ключей и т.п.);
- перечень используемых справочников и классификаторов с указанием их версий;
- определение формата хранения данных;
- информация о происхождении, включая информацию о версиях;
- описание используемых программ и приложений;
- правила восстановления и резервного копирования.

А.4 Примеры операционных метаданных:

Примерами операционных метаданных служат:
- журналы выполнения заданий;
- история использования;
- журнал ошибок;
- отчеты о запросах, включая частоту и время выполнения;
- план и текущее состояние обслуживания и обновления;
- информация о резервном копировании;
- правила архивирования и хранения данных, связанные архивы;
- критерии очистки;
- правила обмена данными;
- технические роли и обязанности, контактная информация.

Библиография

[1] Федеральный закон от 27 июля 2006 г. № 152-ФЗ «О персональных данных»

УДК 004.896:004.624:006.354 ОКС 35.240.90

Ключевые слова: данные для систем искусственного интеллекта в образовании, сбор данных, хранение данных, обработка данных, передача данных, защита данных

Редактор *Н.А. Аргунова*
Технический редактор *И.Е. Черепкова*
Корректор *А.С. Черноусова*
Компьютерная верстка *Г.Р. Арифулина*

Сдано в набор 29.11.2021. Подписано в печать 22.12.2021. Формат 60 × 84 1/8. Гарнитура Ариал.
Усл. печ. л. 1,40. Уч.-изд. л. 1,26.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «РСТ»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru