



ПРЕДВАРИТЕЛЬНЫЙ
НАЦИОНАЛЬНЫЙ
СТАНДАРТ
РОССИЙСКОЙ
ФЕДЕРАЦИИ

ПНСТ
847—
2023

Искусственный интеллект
БОЛЬШИЕ ДАННЫЕ
Функциональные требования в отношении
происхождения данных

(ITU-T Y.3602 (2022), NEQ)

Издание официальное

Москва
Российский институт стандартизации
2024

Предисловие

1 РАЗРАБОТАН Научно-образовательным центром компетенций в области цифровой экономики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова» (МГУ имени М.В. Ломоносова) и Обществом с ограниченной ответственностью «Институт развития информационного общества» (ИРИО)

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 13 декабря 2023 г. № 93-пнст

4 Настоящий стандарт разработан с учетом основных нормативных положений международного документа ITU-T Y.3602 (2022) «Большие данные. Функциональные требования в отношении происхождения данных» (Recommendation ITU-T Y.3602 (2022) «Big data — Functional requirements for data provenance», NEQ)

Правила применения настоящего стандарта и проведения его мониторинга установлены в ГОСТ Р 1.16—2011 (разделы 5 и 6).

Федеральное агентство по техническому регулированию и метрологии собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее чем за 4 мес до истечения срока его действия разработчику настоящего стандарта по адресу: 119991 Москва, Ленинские горы, д. 1 и в Федеральное агентство по техническому регулированию и метрологии по адресу: 123112 Москва, Пресненская набережная, д. 10, стр. 2.

В случае отмены настоящего стандарта соответствующая информация будет опубликована в ежемесячном информационном указателе «Национальные стандарты» и также будет размещена на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

© Оформление. ФГБУ «Институт стандартизации», 2024

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

II

Содержание

1 Область применения1

2 Термины и определения.....1

3 Сокращения2

4 Соглашения по терминологии2

5 Введение в происхождение данных2

6 Обзор управления информацией о происхождении больших данных.....4

7 Функциональные требования к управлению информацией о происхождении больших данных. . . .15

8 Требования безопасности17

Приложение А (справочное) Варианты использования информации о происхождении
больших данных18

Приложение Б (справочное) Примеры профилей информации о происхождении26

Библиография41

Введение

В настоящем стандарте описываются модели и операции с информацией о происхождении больших данных. Он также содержит функциональные требования к сервис-провайдеру больших данных в отношении управления информацией о происхождении больших данных. Надежность данных является важным фактором при определении надежности результата анализа. Информация о происхождении данных помогает обеспечить надежность данных посредством обеспечения прозрачности их исторического пути. В среде больших данных сложная обработка данных и их миграция, связанная с жизненным циклом больших данных и их распространением, приводят к разного рода трудностям при управлении информацией о происхождении данных.

В настоящий стандарт включены дополнительные по отношению к рекомендации ITU-T Y.3602 определения и положения из ГОСТ Р ИСО/МЭК 20546—2021 «Информационные технологии. Большие данные. Обзор и словарь» и ИСО/МЭК 22989 «Информационные технологии. Искусственный интеллект. Понятия и терминология искусственного интеллекта». Это позволяет гармонизировать настоящий стандарт с принятыми ранее национальными стандартами и предварительными национальными стандартами в области ИИ.

Доступ к данной рекомендации можно получить по адресу: <http://handle.itu.int/> (уникальный идентификатор — <http://handle.itu.int/11.1002/1000/14682>).

Все рекомендации ITU-T и другие источники могут подвергаться пересмотру, поэтому всем пользователям настоящего стандарта предлагается изучить возможность применения последнего издания рекомендаций и других справочных документов. Перечень действующих в настоящее время рекомендаций ITU-T регулярно публикуется. Ссылка на документ в рамках настоящего стандарта не придает ему как отдельному документу статус рекомендации.

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Искусственный интеллект

БОЛЬШИЕ ДАННЫЕ

Функциональные требования в отношении происхождения данных

Artificial intelligence. Big data. Functional requirements for data provenance

Срок действия — с 2024—02—01
до 2027—02—01

1 Область применения

Настоящий стандарт устанавливает функциональные требования к происхождению данных в экосистеме больших данных (см. [1]). Стандарт вводит понятия происхождения данных, а также происхождения данных в экосистеме больших данных, и описывает концептуальную модель, операции, логические компоненты и функциональные требования, относящиеся к информации о происхождении больших данных. Содержащиеся в настоящем стандарте функциональные требования получены путем анализа вариантов использования.

2 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями:
2.1

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

Примечание — Термин «большие данные» широко применяется в различных значениях, например в качестве наименования технологии масштабирования, используемой для обработки больших массивов данных.

[ГОСТ Р ИСО/МЭК 20546—2021, пункт 3.1.1]

2.2

происхождение (provenance): Информация, относящаяся к какому-либо источнику информации, включая сведения о стороне или сторонах, участвующих в его создании, введении в оборот и/или поручительстве за него.

[[2], пункт 3.2.11]

2.3 происхождение больших данных (big data provenance): Информация, документирующая исторический путь данных в соответствии с операциями жизненного цикла данных в экосистеме больших данных.

Примечания

- 1 Операции жизненного цикла данных (data lifecycle operations) включают генерацию, передачу, хранение, использование и уничтожение (удаление) данных.
- 2 Информация о происхождении данных обеспечивает детальные сведения об источнике данных, такие, например, как сведения о стороне, ответственной за предоставление данных; сведения о примененных к данным функциях, и информацию о вычислительной среде для обработки данных (например, операционная система, описание аппаратного обеспечения, региональные настройки и часовой пояс).

3 Сокращения

В настоящем стандарте применены следующие сокращения:

ПДн	— персональные данные (personally identifiable information);
BD	— большие данные (big data);
BDC	— потребитель услуг больших данных (big data service customer);
BDSP	— сервис-провайдер больших данных (big data service provider);
DB	— брокер данных (data broker);
DP	— провайдер данных (data provider);
DS	— поставщик данных (data supplier);
H/W	— аппаратное обеспечение (hardware);
OS	— операционная система (operating system);
PI	— информация о происхождении (provenance information);
URI	— унифицированный идентификатор ресурса (uniform resource identifier).

4 Соглашения по терминологии

В настоящем стандарте:

ключевые слова «требуется, чтобы» означают требование, которое должно строго соблюдаться и отклонение от которого не допускается, если будет сделано заявление о соответствии настоящему стандарту;

ключевое слово «рекомендуется» означает требование, которое рекомендуется, но не является абсолютно необходимым. Таким образом, это требование не является обязательным для заявления о соответствии настоящему стандарту;

ключевые слова «может опционально» означают необязательное требование, которое является допустимым, но при этом не подразумевается, что оно в каком-либо смысле рекомендуется. Данная формулировка не подразумевает ни обязанности предлагаемого производителем варианта реализации предоставить такую опцию, ни возможности опционального подключения такой функциональной возможности оператором сети/сервис-провайдером. Она означает лишь то, что производитель может опционально предоставить эту функциональную возможность и по-прежнему заявлять о соответствии спецификации.

В основном тексте настоящего стандарта и в приложениях к нему могут иногда встречаться слова «обязан» (shall), «не вправе» (shall not), «следует» (should) и «может» (may), которые в таком случае следует соответственно толковать как «обязан сделать что-либо», «запрещается что-либо», «весьма рекомендуется что-либо» и «опционально может быть сделано что-либо». При появлении таких фраз или ключевых слов в приложении или в материале, которые явно помечены как «справочные», их следует интерпретировать как не имеющее нормативного характера.

5 Введение в происхождение данных

5.1 Общая концепция происхождения данных

Надежность используемых данных является важным фактором при определении достоверности результатов анализа данных. Действительно, возможно манипулирование данными и их преобразование в соответствии с намерениями аналитика, а также их искажение для получения желательного результата. В этом плане сбор и сохранение информации о происхождении данных направлены на обеспечение надежности данных и результатов анализа посредством обеспечения прозрачности исторического пути данных.

Происхождение — это информация, относящаяся к стороне, ответственной за предоставление данных, а в случаях, когда изменяются исходные данные, — сведения о примененных к данным функциям и информация о вычислительной среде для обработки данных (например, операционная система, описание аппаратного обеспечения, региональные настройки и часовой пояс). С точки зрения управления данными, данные продолжают изменяться до тех пор, пока продолжается их использование и в составе информации о происхождении данных сохраняется история этих изменений:

- в информационном продукте (data product).

Примечание — Информационный продукт представляет собой результат производства данных для целей открытого или платного распространения;

- в процессе, обеспечивающем возможность создания данных.

Примечание — Описание процесса включает информацию о примененных к источнику данных функциях, промежуточных результатах, порядке их использования;

- в процессе документирования метаданных, охватывающем информацию о потоке рабочих процессов, аннотации, заметки о процессах;

- в информации, помогающей проследить историю создания информационного продукта начиная с его первоисточников.

Знание происхождения данных полезно:

- для прослеживания истории создания информационного продукта начиная с его первоисточников;
- для удостоверения качества данных на основе сведений об исходных данных и процессах их обработки;
- для отслеживания источников ошибок;
- для поддержки возможности автоматического повторного выполнения процессов обработки для обновления данных;
- для обеспечения атрибуции источников данных.

Для того, чтобы обеспечить соответствие применения настоящего стандарта национальным нормативным правовым требованиям, концепция происхождения данных в стандарте не включает в себя ни идентификацию субъектов персональных данных, ни сбор информации с целью их идентификации.

Информация о происхождении данных (PI) состоит из набора потоков данных, при этом каждый такой поток содержит информацию о процессах (f), источниках данных (d) и об ответственных сторонах (p). Для отражения этого факта используется нотация:

$$PI = \{(f, p), (d, p)\}.$$

Каждый поток данных включает «непосредственно связанный поток» (directly associated flow) и «подчиненно связанный поток» (subordinately associated flow). Например, на рисунке 5.1 информация о происхождении набора данных «Данные d» состоит из множества:

- непосредственно связанный поток: $PI = \{(f2, pC), (Данные\ c, pC)\}$;
- подчиненно связанный поток: $PI(Данные\ c) = \{(f1, pC), ((Данные\ a, pA), (Данные\ b, pB))\}$,

где использованы обозначения:

- pX — ответственная сторона X,
- $f<n>$ — n -й процесс обработки.

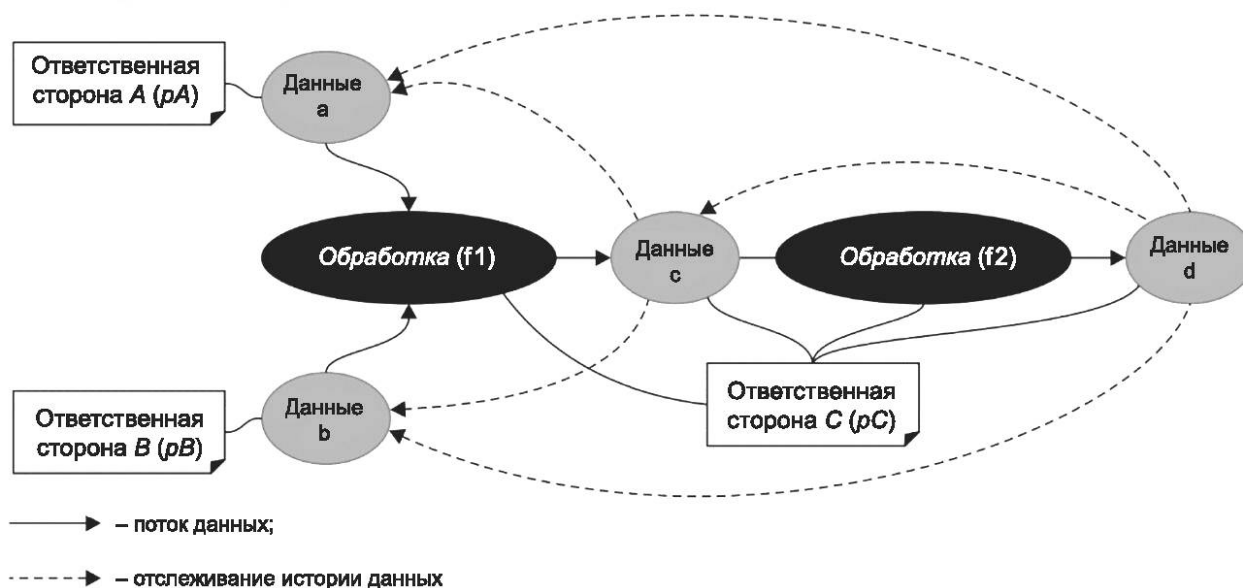


Рисунок 5.1 — Пример информации о происхождении данных PI

5.2 Происхождение данных в среде больших данных

В среде больших данных сложная обработка данных и их миграция, связанная с операциями жизненного цикла больших данных (такими, как создание данных, их передача, хранение, использование, удаление) и с распространением данных, приводят к разного рода трудностям при управлении информацией о происхождении данных. В соответствии с описанием экосистемы больших данных в [1], при управлении информацией о происхождении больших данных следует учитывать:

- огромные объемы неструктурированных, полуструктурированных и структурированных данных;
- описание функций оперирования данными различных типов и форматов;
- историю данных в сферах применения, где для них имеется несколько приложений.

Примечания

1 Сфера применения (application domain) — область знаний или деятельности в одной конкретной экономической, коммерческой, социальной или административной сфере [3].

2 Примерами сфер применения могут служить транспорт, здравоохранение или система государственного управления.

Кроме того, среда больших данных создает для управления информацией о происхождении данных ряд проблем вычислительного характера, таких как:

- обеспечение эффективного механизма хранения информации о происхождении. Объем информации о происхождении может оказаться больше, чем объем исходных данных, что может привести к существенным накладным расходам на хранение;
- минимизация накладных расходов при сборе информации о происхождении. В среде распределенных систем документирование происхождения и затраты на соответствующие вычисления являются важными аспектами, которые следует рассматривать совместно;
- обеспечение возможности воспроизведения обработки на основе информации о происхождении. Для некоторых приложений больших данных информация об их вычислительной среде (например, информация об аппаратном обеспечении и о конфигурации параметров машин обработки больших данных) является важным фактором, который необходимо учитывать в случае, если ставится задача воспроизведения обработки на основе информации о происхождении данных.

Примерами применения информации о происхождении больших данных и получаемой от нее отдачи являются:

- совместный анализ больших данных. Знание происхождения больших данных делает возможным проведение совместного анализа больших данных, связанных с несколькими различными областями применения или приложениями, посредством использования информации об источниках данных и этапах их обработки;
- повторное использование процесса анализа больших данных. Обычно процесс анализа больших данных включает сложные этапы. На основе информации о происхождении может быть получена четко определенная модель анализа, полезная для повторного использования в других случаях обработки больших данных.

Примечание — Обработка данных в системе обработки данных означает последовательность событий, происходящих в соответствии с намеченной целью воздействия;

- автоматизация процесса анализа больших данных. Знание происхождения данных обеспечивает контекст при использовании данных и позволяет автоматически проверять и пересматривать производные данные в случае обновления первичных данных;
- аудит и защита интеллектуальной собственности. Информация о происхождении данных обеспечивает «родословную» данных и позволяет проводить аудит и отслеживание цифровых прав на смешанные данные.

6 Обзор управления информацией о происхождении больших данных

В данном разделе представлен обзор управления информацией о происхождении больших данных. В нем описывается роль и место происхождения данных в экосистеме больших данных, концептуальная модель, операции с информацией о происхождении и логические компоненты для управления информацией о происхождении больших данных.

6.1 Экосистема больших данных и происхождение данных

Согласно [1], сервис-провайдер больших данных (BDSP) осуществляет поддержку информации о происхождении данных в качестве одного из аспектов управления данными, управляя информацией об источнике и методах генерации данных, включая сведения о стороне или сторонах, участвующих в процессах генерации, ввода в оборот и/или объединения данных.

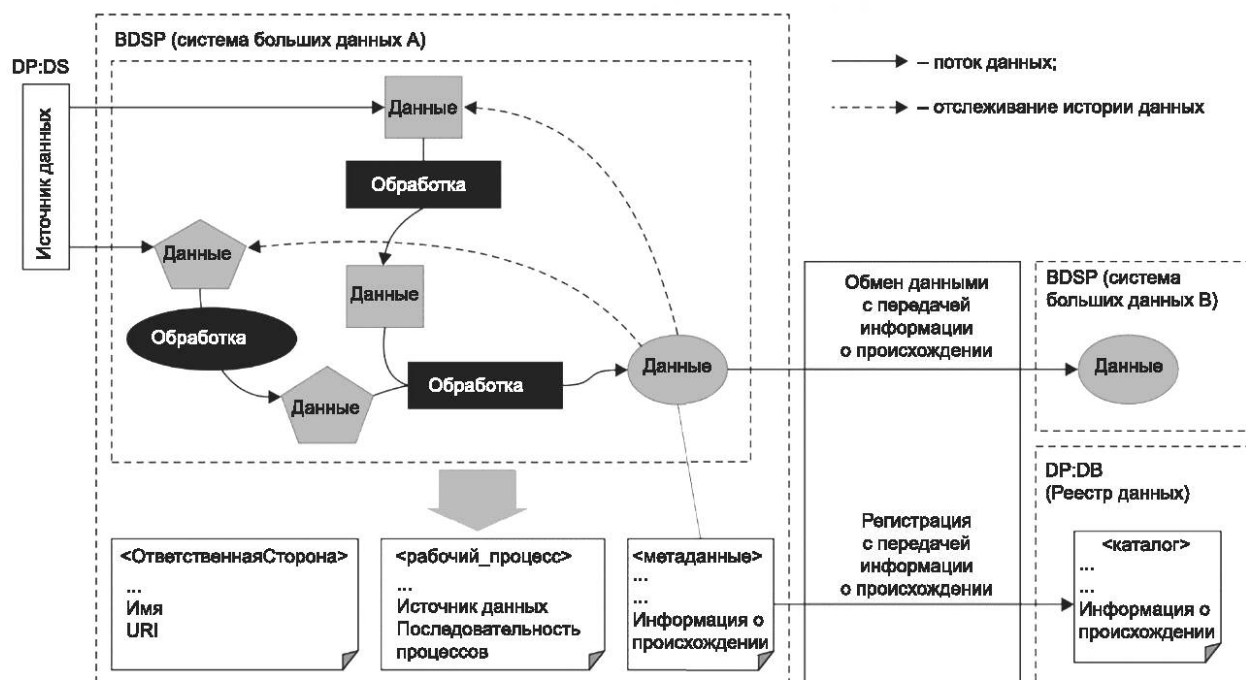


Рисунок 6.1 — Использование информации о происхождении данных в экосистеме больших данных

На рисунке 6.1 показано использование информации о происхождении данных в экосистеме больших данных:

- когда данные импортируются из внешнего источника данных поставщиком данных (DP:DS) и сохраняются, BDSP (система больших данных A) генерирует метаданные на основе контекста импортирования (включающие, например, информацию об ответственной стороне, времени, объеме) и использует элементы метаданных для формирования информации о происхождении данных;
- BDSP A ведет мониторинг и хранит информацию о процессах анализа данных в форме информации о происхождении данных, в интересах обеспечения надежности качества данных и воспроизводимости результатов анализа;
- когда BDSP A экспортирует данные для BDSP B или регистрирует информацию в каталоге данных (data catalogue) в составе реестра данных на торговой площадке данных для брокера данных (DP:DB), BDSP A также предоставляет соответствующую информацию о происхождении данных.

Примечание — Что касается данных, экспортируемых или регистрируемых BDSP, то уровень детализации предоставляемой BDSP информации о происхождении данных зависит от применяемой им политики в отношении данных или предоставления сервисов.

6.2 Концептуальная модель информации о происхождении больших данных

Описание в данном разделе информации о происхождении больших данных основано на расширении общей концепции информации о происхождении данных, приведенной в 5.1.

6.2.1 На рисунке 6.2 показана концептуальная модель информации о происхождении больших данных. Информация о происхождении больших данных (*BDP_ProvenanceInformation*) представляет собой множество блоков происхождения больших данных (*BDP_ProvenanceUnit*).

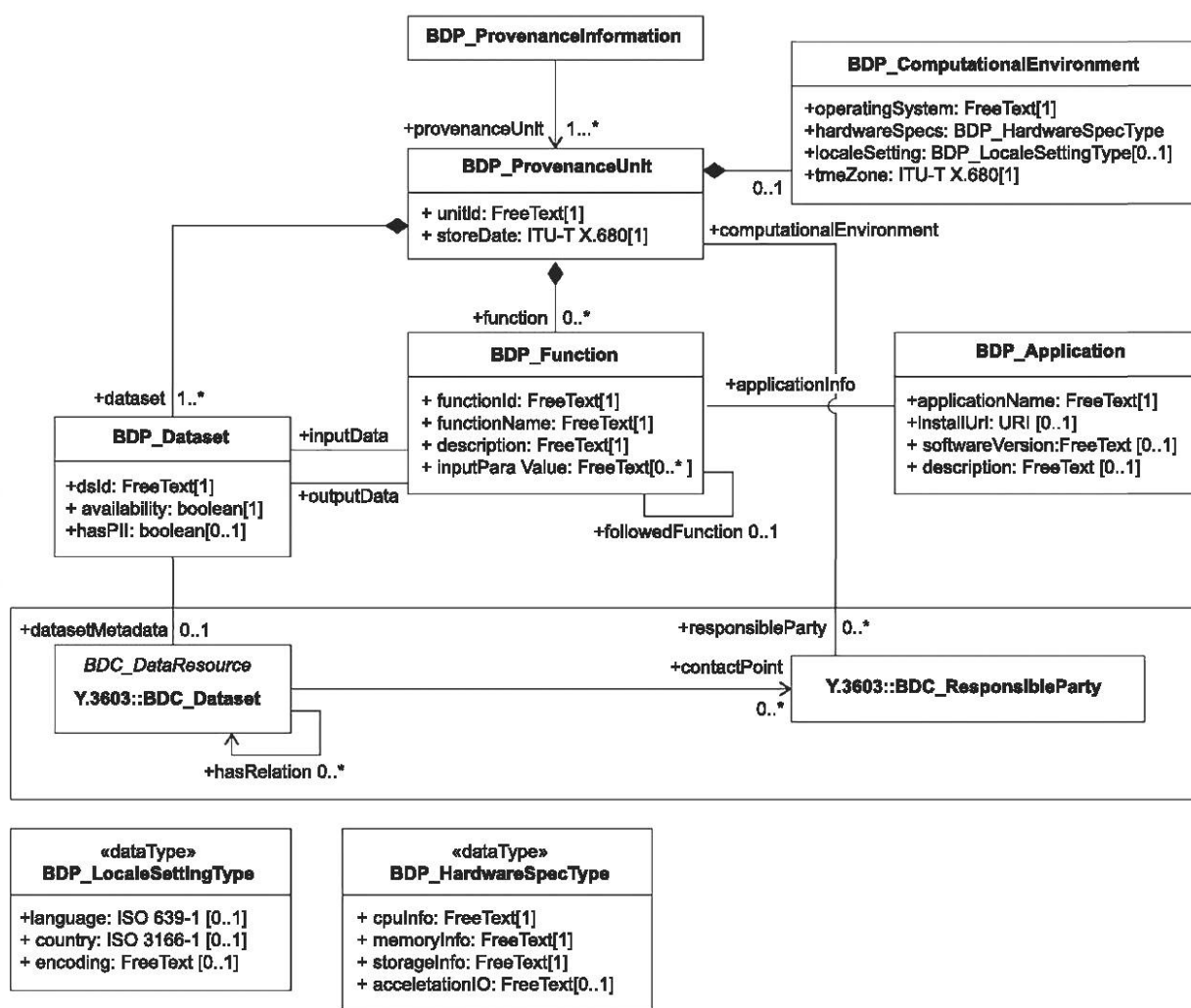


Рисунок 6.2 — Концептуальная модель информации о происхождении больших данных (см. [4])

Каждый блок происхождения больших данных соответствует минимальному набору информации о происхождении больших данных. Он содержит информацию о правах обладания данными и/или полномочиях ими распоряжаться (*BDC_ResponsibileParty*), о среде обработки данных (*BDP_ComputationalEnvironment*) и о последовательности выполнения функций (*BDP_Function*), вместе со сведениями о входных и выходных данных (*BDP_Dataset*), которые использованы в процессах очистки, подготовки, объединения и/или анализа данных.

Чтобы обеспечить соответствие применения настоящего стандарта национальным нормативным правовым требованиям, использование в нем для целей управления информацией о происхождении атрибута *hasPII* в классе *BDP_Dataset* не предполагает идентификацию субъектов персональных данных или сбор информации с целью их идентификации.

Примечания

1 Рабочий процесс показывает фактическую последовательность выполнения используемых функций в процессе обработки данных. В модели информации о происхождении больших данных рабочий процесс может быть определен посредством использования связи (*+followedFunction*), которая позволяет описать последовательность применения используемых в рабочем процессе функций (*BDP_Function*).

2 Класс *BDP_Dataset* хранит информацию о том, имеется ли набор данных в хранилище и содержит ли он персональные данные. Данный класс также ссылается на метаданные *BDC_Dataset*, которые состоят из идентифицируемой для набора данных информации (например, информации о доступе, типе и формате данных, объеме данных, дате).

3 Класс *BDC_ResponsibileParty* определяется в [4], см. раздел 8.

6.2.2 Классы и типы данных

Для объяснения словарей данных в последующих разделах используются следующие описательные элементы (дескрипторы):

- название/роль: название класса (соответствующие строки в таблицах выделены серым фоном), связи (обозначается префиксом «role:») или атрибута;
- описание: краткое описание объекта, указанного в поле «Название/роль»;
- степень обязательности (M/R/O): обязательный (mandatory)/рекомендуемый (recommended)/опциональный (optional) — необходимость каждого класса, связи и атрибута;
- кратность (cardinality): максимальное количество экземпляров, которое может иметь объект или элемент метаданных. Если может быть только один экземпляр, ставится отметка «1»; если экземпляров может быть несколько, ставится отметка «N»;
- допустимые значения: указываются допустимые значения или возможность помещения в поле произвольного текста. Отметка «произвольный текст» означает, что на содержимое поля не накладываются никаких ограничений;
- код допустимого значения: установленный код, используемый взамен детального описания.

6.2.2.1 Класс: BDP_ProvenanceInformation

В таблице 6.1 приведен словарь данных класса *BDP_ProvenanceInformation*.

Т а б л и ц а 6.1 — Обзор информации о происхождении

Номер	Название/роль	Описание	M/R/O	Кратность	Допустимые значения/диапазон
1	BDP_Provenance Information	Корневая структура, определяющая PI	M	1	Строка 2
2	role: provenanceUnit	Блок происхождения, содержащий PI	M	N	BDP_Provenance Unit

6.2.2.2 Класс: BDP_ProvenanceUnit

В таблице 6.2 приведен словарь данных класса *BDP_ProvenanceUnit*.

Т а б л и ц а 6.2 — Блок происхождения

Номер	Название/роль	Описание	M/R/O	Кратность	Допустимые значения/диапазон
3	BDP_Provenance Unit	Запись о последних изменениях в данных	M	N	Строки с 4 по 9
4	role: computational Environment	Информация о вычислительной среде для блока происхождения	O	1	BDP_Computational Environment
5	role: responsibleParty	Информация об ответственной стороне для каталога данных	O	N	BDC_Responsible Party
6	role: dataset	Информация о наборе данных для каталога данных	M	1	BDP_Dataset
7	role: function	Информация о функции для каталога данных	R	N	BDP_Function
8	unitId	Уникальный идентификатор блока происхождения	M	1	Произвольный текст
9	storedDate	Дата сохранения блока происхождения	R	N	См. [5]

П р и м е ч а н и е — Согласно [5] поддерживаются все форматы даты и времени, специфицированные в [6].

6.2.2.3 Класс: BDP_ComputationalEnvironment

В таблице 6.3 приведен словарь данных класса *BDP_ComputationalEnvironment*.

Таблица 6.3 — Вычислительная среда

Номер	Название/роль	Описание	M/R/O	Кратность	Допустимые значения/диапазон
10	BDP_Computational Environment	Информация о вычислительной среде	O	1	Строки с 11 по 14
11	operatingSystem	Информация об операционной системе	M	1	Произвольный текст
12	hardwareSpecs	Информация о спецификациях оборудования	M	N	BDP_Hardware Spec-Type
13	localeSetting	Информация о языке, форматах отображения времени, даты, валюты и т. д.	O	1	BDP_Locale Setting-Type
14	timeZone	Информация о часовом поясе	M	N	ITU-T X.680
15	BDP_Hardware Spec-Type	Информация, необходимая для описания аппаратной среды	M	1	Строки с 16 по 19
16	cpuInfo	Информация о скорости и количестве центральных процессоров	M	1	Произвольный текст
17	memoryInfo	Информация об объеме и быстродействии оперативной памяти	M	1	Произвольный текст
18	storageInfo	Информация об объеме и быстродействии системы хранения	M	1	Произвольный текст
19	accelerationIO	Информация об аппаратном ускорителе (например, GPU)	O	1	Произвольный текст
20	BDP_Locale Setting-Type	Информация, необходимая для описания региональных настроек	O	1	Строки с 21 по 23
21	language	Естественный язык, используемый при работе с оборудованием	O	1	См. [7]
22	country	Код страны	O	1	См. [8]
23	encoding	Название используемой кодировки символов	O	1	Произвольный текст (например, «UTF-8»)

6.2.2.4 Класс: BDC_ResponsibleParty

В таблице 6.4 приведен словарь данных класса *BDC_ResponsibleParty*. Более детальная информация приведена в [4], 8.2.2.

Таблица 6.4 — Информация об ответственной стороне

Номер	Название/роль	Описание	M/R/O	Кратность	Допустимые значения/диапазон
24	BDC_Responsible Party	Информация об имени и должности лица или о названии организации, ответственных за каталог данных или набор данных	O	N	См. [4]

6.2.2.5 Класс: BDP_Dataset

В таблице 6.5 приведен словарь данных класса *BDP_Dataset*.

Таблица 6.5 — Набор данных

Номер	Название/роль	Описание	M/R/O	Кратность	Допустимые значения/ диапазон
25	BDP_Dataset	Набор данных, информация о котором записана в блоке происхождения	M	N	Строки с 26 по 30
26	role: datasetMetadata	Метаданные набора данных	O	1	BDC_Dataset
27	dsId	Уникальный идентификатор набора данных	M	1	Произвольный текст
28	availability	Информация о том, существуют ли данные в хранилище или были удалены	M	1	Логическое значение
29	hasPII	Информация о том, присутствуют ли персональные данные в наборе данных	R	1	Логическое значение
30	BDC_Dataset	Набор данных, который является доступным и/или может быть скачан	O	1	См. [4]

6.2.2.6 Класс: BDP_Function

В таблице 6.6 приведен словарь данных класса *BDP_Function*.

Таблица 6.6 — Функция обработки

Номер	Название/роль	Описание	M/R/O	Кратность	Допустимые значения/ диапазон
31	BDP_Function	Информация о примененной функции	O	N	Строки с 32 по 39
32	role: inputData	Информация о входном наборе данных	O	N	BDP_Dataset
33	role: outputData	Информация о выходном наборе данных	O	N	BDP_Dataset
34	role: followedFunction	Информация о функции, примененной следующей	O	1	BDP_Function
35	role: applicationInfo	Информация о программном обеспечении, обеспечивающем выполнение примененной функции	M	1	BDP_Application
36	functionId	Уникальное название функции	M	N	Произвольный текст
37	functionName	Альтернативное название функции	O	1	Произвольный текст
38	description	Краткое описание функции	M	1	Произвольный текст
39	inputParaValue	Значение входного параметра	O	N	Произвольный текст

6.2.2.7 Класс: BDP_Application

В таблице 6.7 приведен словарь данных класса *BDP_Application*.

Таблица 6.7 — Программное приложение

Номер	Название/роль	Описание	M/R/O	Кратность	Допустимые значения/ диапазон
40	BDP_Application	Информация о программном обеспечении, обеспечивающем выполнение примененной функции	M	1	Строки с 41 по 44
41	applicationName	Название программного приложения	M	1	Произвольный текст

Окончание таблицы 6.7

Номер	Название/роль	Описание	M/R/O	Кратность	Допустимые значения/диапазон
42	installUri	Унифицированный идентификатор ресурса URI, с которого может быть установлено приложение	R	1	URI
43	softwareVersion	Версия программного обеспечения	R	1	Произвольный текст (старшая_версия. младшая_версия. патч)
44	description	Дополнительная информация о программном приложении	O	1	Произвольный текст

6.2.3 Примеры информации о происхождении

На рисунке 6.3 показан пример захвата блока происхождения для набора данных «Данные С».

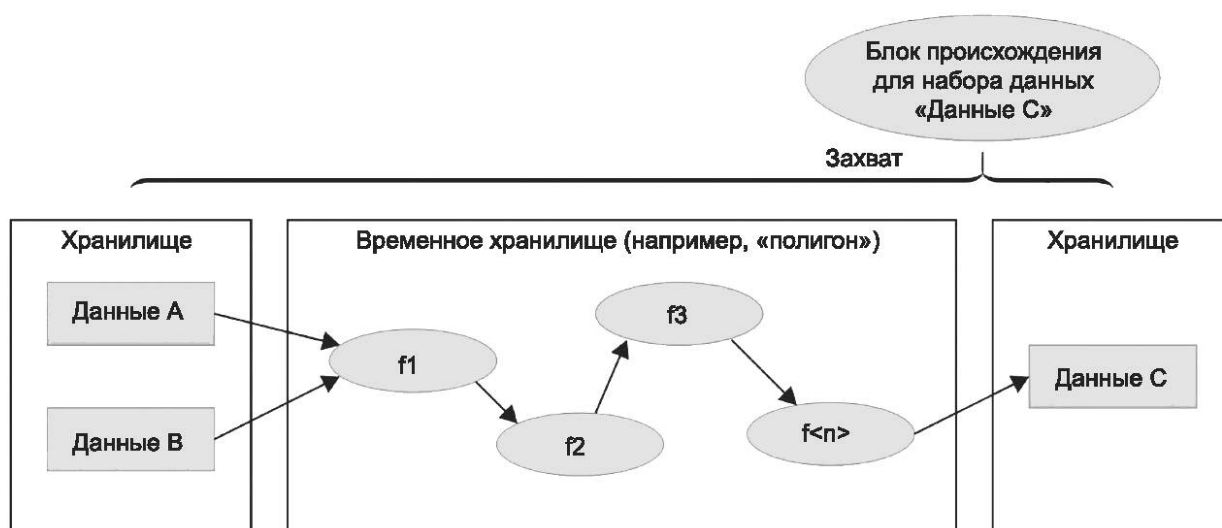


Рисунок 6.3 — Пример захвата блока происхождения для набора данных

Захват блока происхождения происходит одновременно с сохранением набора данных «Данные С» в хранилище данных. Несмотря на то, что все функции имеют свои входные и выходные данные, в блоке происхождения документируется информация о первых входных данных и о последних выходных данных, при этом шаги процесса описываются в виде последовательности функций.

Примечание — На рисунке 6.3 наборы данных «Данные А» и «Данные В» являются входными данными, а набор данных «Данные С» — выходными данными.

На рисунке 6.4 показан пример информации о происхождении больших данных, представленный в виде графа. Информация о происхождении набора данных «Данные D» представляет собой объединение блока происхождения наборов данных «Данные С» с блоком происхождения набора данных «Данные D».

6.3 Операции с информацией о происхождении

При изменении состояния данных в случае, например, их сохранения, обновления или удаления осуществляется запись, сохранение, комбинирование или удаление блока происхождения. На рисунке 6.5 показана взаимосвязь между изменением состояния данных и операциями с информацией о происхождении.

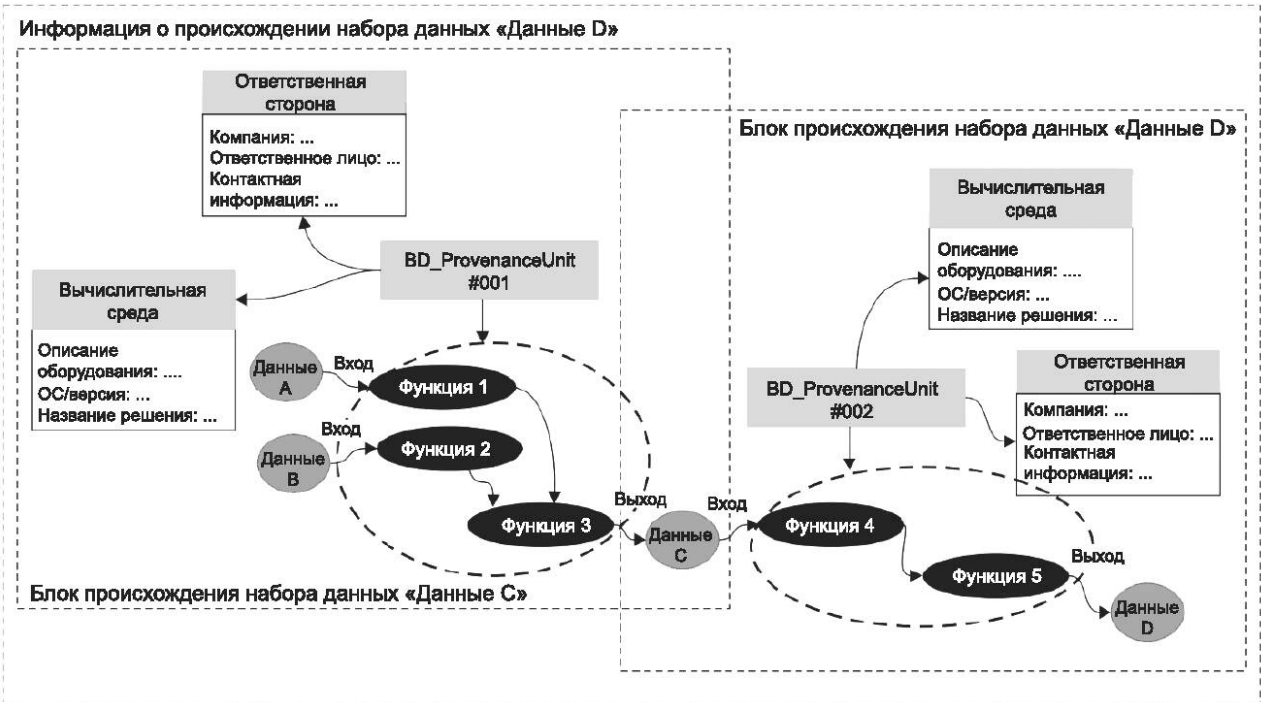


Рисунок 6.4 — Пример информации о происхождении больших данных



Рисунок 6.5 — Операции с информацией о происхождении, выполняемые в соответствии с изменением состояния данных

- Операциями по управлению информацией о происхождении являются:
- запись блока происхождения (см. 6.3.1);
 - сохранение блока происхождения (см. 6.3.2.1);
 - объединение блоков происхождения (см. 6.3.2.2);
 - удаление блока происхождения (см. 6.3.2.3);
 - извлечение информации о происхождении (см. 6.3.3).

6.3.1 Запись блока происхождения

На рисунке 6.6 показан пример записи блока происхождения в случае сохранения и обновления данных:

- запись блока происхождения. Когда сохраняется набор данных «Данные 1», BDSP записывает Блок происхождения 1. Далее набор данных «Данные 1» обновляются сначала до набора данных «Данные 2», затем до набора данных «Данные 3», а BDSP последовательно записывает Блок происхождения 2, а затем Блок происхождения 3.

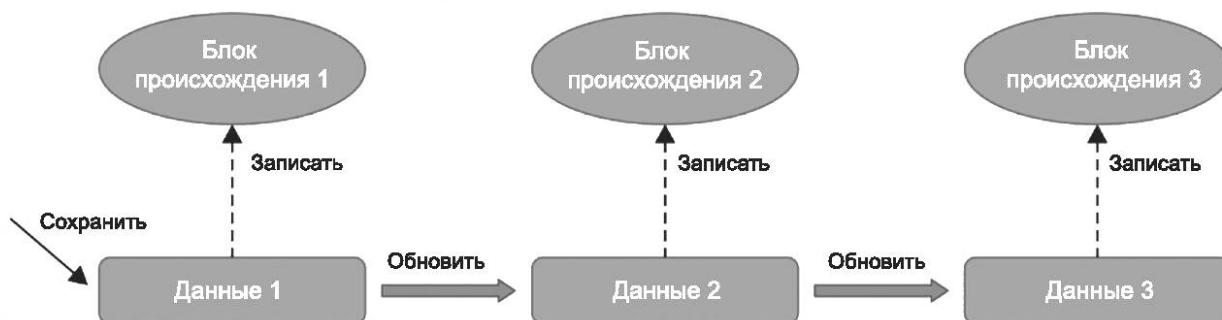


Рисунок 6.6 — Запись блоков происхождения

6.3.2 Операции с информацией о происхождении в случае удаления данных

В случае удаления данных система управления информацией о происхождении BDSP действует тремя способами, описанными в 6.3.2.1—6.3.2.3.

6.3.2.1 Сохранение блока происхождения

Когда набор данных «Данные 2» удаляется из хранилища, система управления информацией о происхождении сохраняет соответствующий Блок происхождения 2 с целью сохранения информации о происхождении данных в процессе обработки данных. На рисунке 6.7 показан пример, когда блоки происхождения сохраняются после удаления данных.



Рисунок 6.7 — Вариант сохранения блока происхождения

6.3.2.2 Объединение блоков происхождения

Как показано в примере на рисунке 6.8, когда набор данных «Данные 2» удаляется из хранилища, система управления информацией о происхождении объединяет его Блок происхождения 2 с ближайшим последующим блоком происхождения в рамках этапов процесса обработки (в данном примере это Блок происхождения 3).

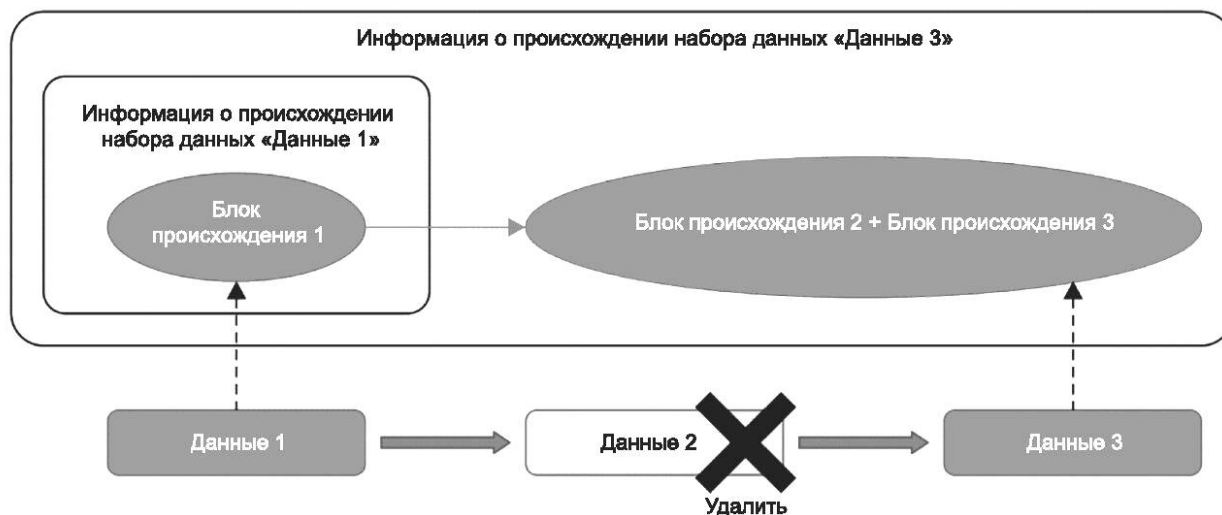


Рисунок 6.8 — Вариант объединения блоков происхождения

6.3.2.3 Удаление блока происхождения

Как показано в примере на рисунке 6.9, когда набор данных «Данные 3» удаляется из хранилища, система управления информацией о происхождении удаляет соответствующий Блок происхождения 3 вместе с набором данных, поскольку они были получены в ходе процесса последними.

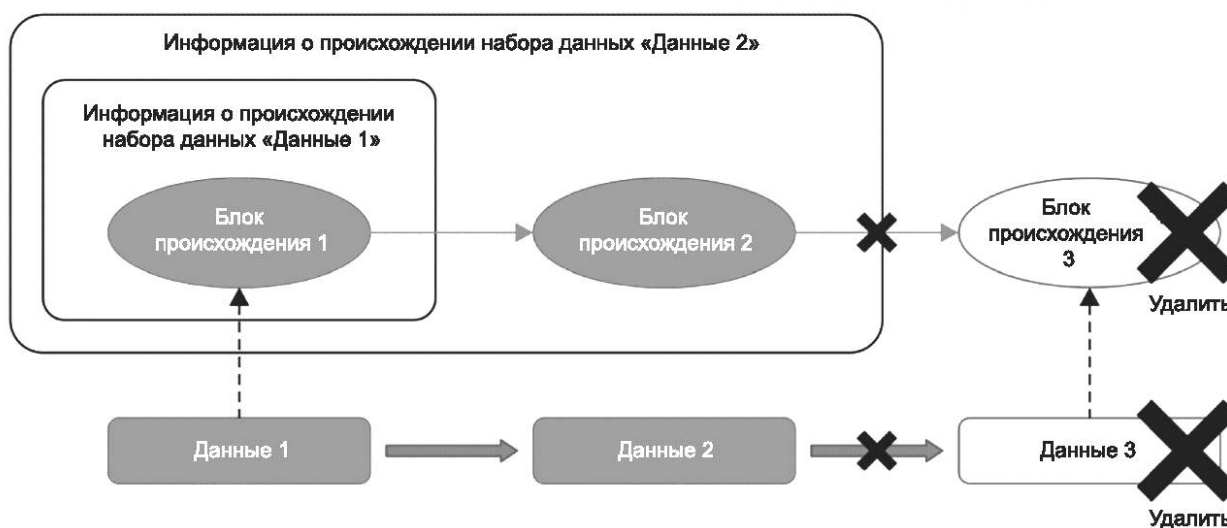


Рисунок 6.9 — Вариант удаления блока происхождения

6.3.3 Извлечение информации о происхождении

На рисунке 6.10 показан пример извлечения информации о происхождении. Вначале из приложения запрашивается информация о происхождении набора данных «Данные 3» (шаг 1), затем BDSP последовательно прослеживает историю происхождения набора данных на основе каждого выявленного блока происхождения (шаг 2). После этого BDSP агрегирует выявленные блоки происхождения (шаг 3) и предоставляет агрегированную информацию о происхождении набора данных «Данные 3» обратно запрашивающему приложению (шаг 4).

6.4 Логические компоненты для управления информацией о происхождении больших данных

На рисунке 6.11 показаны логические компоненты для управления информацией о происхождении больших данных.

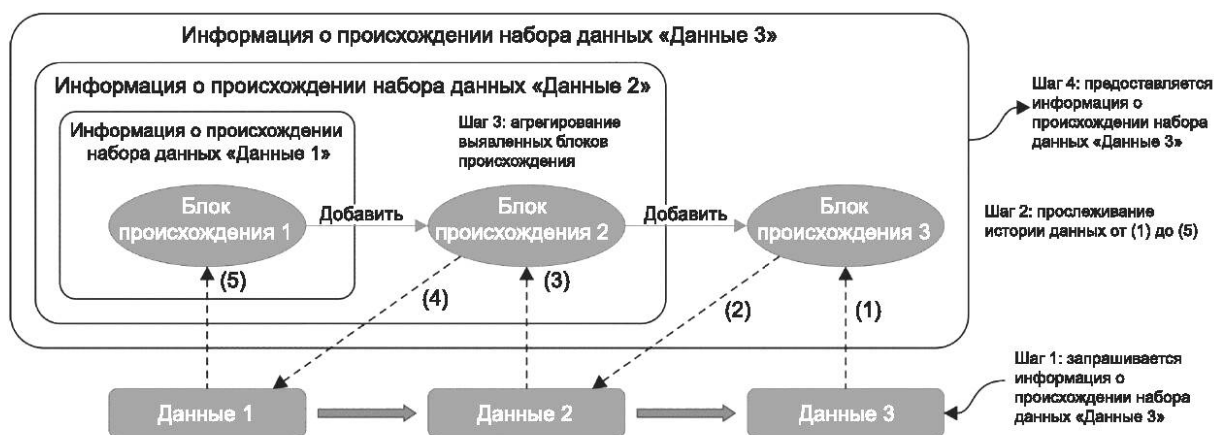


Рисунок 6.10 — Извлечение информации о происхождении



Рисунок 6.11 — Конфигурация логических компонентов для управления информацией о происхождении больших данных

На рисунке 6.11 показаны следующие логические компоненты:

- управление моделью информации о происхождении. Данный логический компонент обеспечивает совместимость информации о происхождении у различных BDSP. Данный логический компонент проводит валидацию полученной извне информации о происхождении на основе модели информации о происхождении больших данных (см. 6.2). Прошедшая валидацию информация о происхождении затем кодируется в соответствии с общей моделью и передается в логический компонент управления жизненным циклом информации о происхождении, который ее сохраняет;
- управление жизненным циклом информации о происхождении. Данный логический компонент выполняет запись и удаление информации о происхождении в соответствии с осуществляемым сохранением, обновлением и удалением данных (см. 6.3.1 и 6.3.2). Данный логический компонент поддерживает извлечение информации о происхождении (см. 6.3.3);
- поддержка анализа. Данный логический компонент извлекает из информации о происхождении рабочие процессы и сохраняет их. Из сохраненных рабочих процессов данный логический компонент извлекает рабочие процессы, являющиеся кандидатами на проведение анализа исходя из информации об имеющихся у BDSP данных и функциях для их анализа. В случае запроса информации о происхождении или рабочего процесса из другой системы (например, у внешнего BDSP) этот логический компонент может проверить возможность адаптации вычислительной среды и подобрать эквивалентные

функции в своей системе. Данный логический компонент также поддерживает автоматизацию процесса анализа данных на основе обновления данных, добавления пользовательских аннотаций к информации о происхождении и управления взаимосвязями между имеющимися у BDSP функциями и данными.

Примечание — На основе зафиксированной в информации о происхождении взаимосвязи между функциями и данными можно запросить список имеющихся данных, к которым могут применяться такие же функции, и список функций, применимых к соответствующим данным;

- управление политиками совместного использования информации о происхождении. Данный логический компонент управляет несколькими политиками совместного использования информации о происхождении. При экспорте информации о происхождении BDSP сверяется с политикой совместного использования и может упростить информацию о происхождении перед отправкой ее другому BDSP.

Примечание — Информация о происхождении может содержать конфиденциальную информацию (например, ключевые идеи конкретного процесса анализа). По этой причине BDSP формулирует политику и реализует процессы, обеспечивающие надлежащее и ответственное совместное использование информации о происхождении. Политика совместного использования может реализовываться посредством обобщения на различных уровнях определенных деталей в составе информации о происхождении;

- управление персональными данными. Данный логический компонент в момент записи блока происхождения проверяет, содержит ли экземпляр данных персональные данные. Этот логический компонент также запрашивает применение имеющегося у BDSP механизма защиты к содержащей персональные данные информации о происхождении;

- мониторинг. Данный логический компонент осуществляет мониторинг изменений в вычислительной среде и в информации об ответственных сторонах по сравнению со сведениями, используемыми при формировании информации о происхождении. При обнаружении таких изменений этот логический компонент обновляет соответствующие сведения.

7 Функциональные требования к управлению информацией о происхождении больших данных

7.1 Требования к жизненному циклу информации о происхождении

Требования к жизненному циклу информации о происхождении включают:

- описание модели информации о происхождении. Требуется, чтобы BDSP поддерживал модель информации о происхождении больших данных.

Примечания

1 Модель информации о происхождении больших данных включает информацию о названиях функций и их использовании, о вычислительной среде, о типах и форматах входных и выходных данных, о входных параметрах, об ответственных сторонах и т. д.

2 Примером информации о вычислительной среде является информация об операционной системе, региональных настройках, часовом поясе; описание оборудования и т. д.;

- использование распространенного формата для обмена. Рекомендуется, чтобы BDSP поддерживал кодирование и декодирование информации о происхождении в распространенном формате для ее использования в разных системах.

Примечание — В настоящем стандарте под кодированием понимается процесс преобразования информации о происхождении в специализированный формат, а под декодированием — процесс, обратный кодированию;

- начало документирования происхождения. Требуется, чтобы BDSP записывал блок происхождения при сохранении данных.

Примечание — Информация, содержащаяся в метаданных (полученных от брокера данных DP:DB или созданных BDSP), может использоваться для формирования блока происхождения;

- сохранение блоков происхождения. Требуется, чтобы BDSP поддерживал экономически эффективный механизм сохранения блоков происхождения.

Примечание — В случае записи информации о происхождении потоковых данных в рамках эффективного механизма хранения следует установить предопределенный интервал времени, по истечении которого записывается очередной блок происхождения вместо записи блока происхождения всякий раз, когда данные сохраняются. Можно также рассмотреть возможность использования методов сжатия данных;

- хранение информации о происхождении. BDSP может опционально поддерживать опережающее сохранение информации о происхождении с целью своевременного удовлетворения запросов пользователя на предоставление информации о происхождении;
- поиск блоков происхождения. Требуется, чтобы BDSP поддерживал поиск блоков происхождения;
- объединение блоков происхождения. Требуется, чтобы BDSP поддерживал объединение блоков происхождения.

Примечание — Объединение блоков происхождения требуется в случае уничтожения (удаления) данных (см. 6.3.2.2);

- агрегирование информации о происхождении. Требуется, чтобы BDSP поддерживал агрегирование блоков происхождения;
- удаление блоков происхождения. Требуется, чтобы BDSP поддерживал удаление блоков происхождения.

Примечания

1 В случае удаления данных BDSP действует в соответствии с одним из трех механизмов (сохранения, объединения или удаления блока происхождения), в зависимости от обстоятельств (см. 6.3.2).

2 BDSP может продолжить хранение блока происхождения после удаления связанных с ним данных. Этот вопрос регулируется политикой управления информацией о происхождении.

7.2 Требования к поддержке анализа

Требования к поддержке анализа включают:

- выделение рабочих процессов. Требуется, чтобы BDSP обеспечивал извлечение информации о рабочем процессе из информации о происхождении;
- сохранение рабочего процесса. Рекомендуется, чтобы BDSP поддерживал сохранение информации о рабочем процессе.

Примечание — Рабочий процесс сохраняется в виде графа, который организован с учетом частоты использования функций анализа и взаимосвязей между ними, определяющих последовательность их использования;

- извлечение рабочего процесса. Рекомендуется, чтобы BDSP поддерживал извлечение информации о рабочем процессе;
- предоставление списка данных по функции. Рекомендуется, чтобы BDSP предоставлял список данных, связанных с определенной функцией, использование которой было зафиксировано в информации о конкретном рабочем процессе;
- предоставление списка функций по данным. Рекомендуется, чтобы BDSP предоставлял список функций, связанных с определенными данными, использование которых было зафиксировано в информации о конкретном рабочем процессе;
- автоматизация анализа данных. Рекомендуется, чтобы BDSP поддерживал автоматизацию анализа данных на основе информации о рабочем процессе;
- пользовательские аннотации. BDSP опционально может поддерживать аннотирование информации о происхождении;
- эквивалентные функции для шагов процесса. Рекомендуется, чтобы BDSP подбирал эквивалентные функции для шагов процесса на основе информации о происхождении, поступающей из другой системы.

Примечания

1 Для подбора подобной эквивалентной функции могут использоваться название функции, формат и структура входных и выходных данных этой функции, частота использования функций анализа и взаимосвязи между ними.

2 Результатом подбора эквивалентной функции может быть такая же функция с другим именем или же комбинация функций, позволяющая получить такой же результат;

- адаптивность вычислительной среды. Рекомендуется, чтобы BDSP обеспечивал определение характеристик вычислительной среды на основе информации о происхождении, поступающей из другой системы.

7.3 Требования к мониторингу

Требования к мониторингу включают:

- мониторинг вычислительной среды. Требуется, чтобы BDSP проводил мониторинг изменений, касающихся вычислительной среды;

- мониторинг ответственных сторон. Требуется, чтобы BDSP проводил мониторинг изменений, касающихся информации об ответственных сторонах;
- применение результатов мониторинга. Требуется, чтобы BDSP использовал результаты мониторинга для обновления информации, используемой при записи блоков происхождения.

Примечание — Результаты мониторинга включают информацию об изменениях в вычислительной среде и в сведениях об ответственных сторонах.

7.4 Требования к управлению политикой

Требования к управлению политикой включают:

- проверка на наличие персональных данных. Требуется, чтобы BDSP проводил проверку на наличие персональных данных в экземпляре данных при записи соответствующего блока происхождения.

Примечания

1 Проверка на присутствие персональных данных осуществляется в соответствии с политикой BDSP в отношении персональных данных.

2 В блоке происхождения в составе информации об экземпляре данных (BDP_Dataset) имеется информация о том, содержит ли тот персональные данные (см. 6.2);

- защита персональных данных. Требуется, чтобы BDSP обеспечил наличие механизма защиты персональных данных в источниках данных.

Примечание — Если персональные данные присутствуют в источниках данных, то BDSP принимает решение о том, исключать их или нет из состава предоставляемых пользователю данных, исходя из наличия у пользователя полномочий на доступ к таким данным;

- упрощение информации о происхождении. Рекомендуется, чтобы BDSP поддерживал упрощение информации о происхождении на основе политики совместного использования.

Примечания

1 Методы упрощения информации о происхождении включают ряд уровней детализации и форматов кодирования и т. д..

2 О «политике совместного использования» см. примечание 2 в 6.4;

- уровень совместного использования информации о происхождении. Требуется, чтобы BDSP поддерживал политику совместного использования, предусматривающую различные уровни детальности информации о происхождении.

Примечания

1 Уровень детальности раскрываемой информации о происхождении определяется политикой совместного использования. Информация о происхождении включает сведения об этапах процесса, применяемых функциях, промежуточных данных и информацию об ответственных сторонах. При передаче другой стороне информацию о происхождении можно упростить в соответствии с политикой совместного использования.

2 Об уровне детальности информации о происхождении см. также примечание 2 в 6.4.

8 Требования безопасности

Необходимо принимать во внимание соответствующие требования по безопасности из [9], [10] и применимых Рекомендаций ITU-T серий X, Y и M, включая требования в отношении управления доступом, аутентификации, конфиденциальности данных, политики хранения данных, сетевой безопасности, целостности данных, доступности и защиты персональных данных.

Приложение А
(справочное)

Варианты использования информации о происхождении больших данных

Т а б л и ц а А.1 — Вариант использования — начало документирования

Название	Начало документирования происхождения
Описание	BDSP запрашивает данные у провайдера данных (DP) и получает их. В этот момент BDSP или сохраняет данные как есть, или модифицирует исходные данные таким образом, чтобы их можно было включить в его собственную базу данных. В ходе этого процесса BDSP начинает документировать информацию о происхождении, включая описание с помощью метаданных происхождения данных и функциональные процессы, которые были применены для модификации данных, сохраняя их затем в распределенной базе данных
Роли/суброли	DP:DS DP:DB BDSP
Рисунок (опционально)	
Предварительные условия (опционально)	Поставщик данных DP:DS публикует метаданные для брокера данных DP:DB; BDSP ведет поиск по данным, полученным от брокера данных DP:DB и либо запрашивает данные у поставщика данных DP:DS, либо ведет поиск по его данным
Постусловия (опционально)	—
Производные требования	Начало документирования происхождения (см. 7.1); описание модели информации о происхождении (см. 7.1); сохранение блоков происхождения (см. 7.1); пользовательские аннотации (см. 7.2); уровень совместного использования информации о происхождении (см. 7.4)

Т а б л и ц а А.2 — Вариант использования — обновление данных и управление информацией о происхождении

Название	Обновление данных и управление информацией о происхождении в системе больших данных
Описание	<Добавление информации о происхождении в связи с изменением статуса источника данных> - BDSP регулярно использует данные от поставщика данных DP:DS. Когда DP:DS обновляет статус данных (например, обновляется версия схемы, вносятся изменения в информацию об ответственном лице), BDSP обновляет информацию об источнике данных; <Управление информацией о происхождении в случае удаления или архивации данных из локального хранилища данных> - BDSP удаляет либо архивирует хранившиеся данные с целью повышения эффективности хранения, а также по иным причинам управленческого характера. BDSP ведет мониторинг информации о происхождении и решает, удалять ее или нет

Окончание таблицы А.2

Название	Обновление данных и управление информацией о происхождении в системе больших данных
Роли/суброли	DP:DS BDSP
Рисунок (опционально)	Обновление версии данных Регулярные запросы данных Поставщик данных (DP:DS) Сервис-провайдер больших данных (BDSP) Обновление информации об источнике данных Проверка доступности существующих данных. Уничтожение либо архивация данных Данные версии 1 Данные версии 2
Предварительные условия (опционально)	BDSP сохранил данные в формате XML с внешним унифицированным идентификатором ресурса (URI), полученным от DP:DS
Постусловия (опционально)	—
Производные требования	Поиск блоков происхождения (см. 7.1); объединение блоков происхождения (см. 7.1); удаление блоков происхождения (см. 7.1); мониторинг вычислительной среды (см. 7.3); мониторинг ответственных сторон (см. 7.3); применение результатов мониторинга (см. 7.3)

Таблица А.3 — Вариант использования — совместное использование и агрегирование информации о происхождении

Название	Совместное использование и агрегирование информации о происхождении
Описание	Два сотрудничающих BDSP делятся друг с другом информацией о происхождении. Когда BDSP В использует данные BDSP А: - BDSP В запрашивает и получает данные от BDSP А; - BDSP В инициирует документирование информации о происхождении при сохранении данных, полученных от BDSP А. Если BDSP В требуется дополнительная информация о данных, полученных от BDSP А (например, история данных), то он: - запрашивает у BDSP А его информацию о происхождении полученных данных; - агрегирует полученную от BDSP А информацию о происхождении с локальной информацией, созданной при сохранении полученных от BDSP А данных
Роли/суброли	BDSP
Рисунок (опционально)	Сервис-провайдер больших данных А Сервис-провайдер больших данных В Интеграция данных. Анализ данных Поиск информации о происхождении. Кодирование информации о происхождении Декодирование информации о происхождении. Агрегирование информации о происхождении (предварительный) запрос больших данных Большие данные Запрос информации о происхождении Информация с происхождением

Окончание таблицы А.3

Название	Совместное использование и агрегирование информации о происхождении
Предварительные условия (опционально)	BDSP В запросил большие данные у BDSP А и получил их
Постусловия (опционально)	BDSP В сохранил агрегированную информацию о происхождении
Производные требования	Поиск блоков происхождения (см. 7.1); использование распространенного формата для обмена (см. 7.1); агрегирование информации о происхождении (см. 7.1); упрощение информации о происхождении (см. 7.4)

Т а б л и ц а А.4 — Вариант использования — повторное использование методов обработки данных

Название	Повторное использование и автоматизация применения методов обработки данных на основе информации о происхождении больших данных
Описание	Аналитик данных (например, потребитель сервиса больших данных BDC, осуществляющий деятельность по анализу данных, см. [1]), на основе существующих результатов анализа больших данных готовит эксперимент, в котором будет использоваться другой источник данных. С этой целью аналитик данных использует предоставляемые BDSP функции анализа информации о происхождении для извлечения процесса анализа данных и его применения: - аналитик данных изучает результаты анализа; - аналитик данных выбирает результат анализа с целью повторного использования примененных в нем методов обработки; - аналитик данных извлекает из информации о происхождении для выбранного результата анализа сведения о методах обработки данных и взаимосвязанных данных, а затем модифицирует их так, чтобы они подходили для нового эксперимента; - аналитик данных применяет метод обработки данных к новым данным. Аналитик данных использует автоматизацию анализа обновленных данных на основе информации о происхождении. Аналитик данных настраивает периодическое выполнение анализа обновленных данных с учетом периода обновления данных
Роли/суброли	DP:DS BDSP BDC
Рисунок (опционально)	< Повторное использование методов обработки данных >

Окончание таблицы А.4

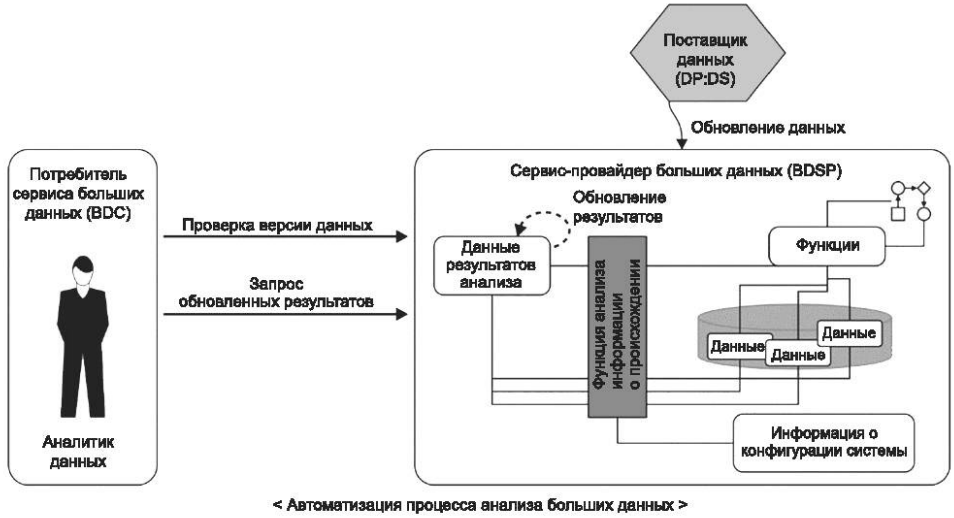
Название	Повторное использование и автоматизация применения методов обработки данных на основе информации о происхождении больших данных
Рисунок (опционально)	
Предварительные условия (опционально)	BDSP хранит информацию о происхождении данных, представляющих собой результаты анализа
Постусловия (опционально)	—
Производные требования	Выделение рабочих процессов (см. 7.2); извлечение рабочего процесса (см. 7.2); автоматизация анализа данных (см. 7.2); предоставление списка данных по функции (см. 7.2); предоставление списка функций по данным (см. 7.2)

Таблица А.5 — Вариант использования — управление персональными данными

Название	Управление персональными данными
Описание	Аналитик данных (например, потребитель сервиса больших данных BDC, выполняющий деятельность по анализу данных, см. [1]), запрашивает у BDSP информацию о происхождении набора данных «Данные 1». В таком случае BDSP: - прослеживает историю набора данных «Данные 1»; - проверяет, содержит ли какой-либо из наборов данных персональные данные, и выясняет, что набор данных «Данные 3» такие данные содержит; - проверяет право доступа BDC к набору данных «Данные 3»
Описание	Если BDC имеет право доступа к набору данных «Данные 3», то BDSP агрегирует блоки происхождения и возвращает их BDC. Если у BDC нет такого права, то BDSP либо возвращает информацию о происхождении за исключением Блока происхождения 3, либо осуществляет обобщение информации о происхождении на основе предопределенной политики и возвращает результат запрашивающей стороне
Роли/суброли	BDC BDSP

Окончание таблицы А.5



Таблица А.6 — Вариант использования — повторное использование информации о происхождении из другой аналитической системы

Название	Повторное использование информации о происхождении из другой аналитической системы
Описание	Аналитик данных (например, потребитель сервиса больших данных BDC, осуществляющий деятельность по анализу данных, см. [1]), готовит эксперимент со своими данными с использованием полученной от BDSP A информации о происхождении
Описание	С этой целью аналитик данных использует предоставляемые BDSP A функции анализа информации о происхождении для извлечения рабочего процесса анализа данных и модифицирует этот рабочий процесс так, чтобы он соответствовал среде для анализа BDSP B: - аналитик данных запрашивает у BDSP A информацию о происхождении BDSP B; - BDSP B декодирует информацию о происхождении; - BDSP B извлекает рабочий процесс из информации о происхождении; - BDSP B проверяет способность рабочего процесса к адаптации и преобразует его в доступный для использования: 1) BDSP B проверяет возможность адаптации у себя вычислительной среды BDSP A, 2) BDSP B отображает извлеченные из информации о происхождении шаги процесса и соответствующие функции в поддерживаемые им функции, 3) если корректно сопоставить функции не удастся, BDSP B проверяет наличие альтернативных функций, и аналитик данных выбирает функции из их числа, 4) аналитик данных применяет шаги процесса к своим собственным данным
Роли/суброли	BDSP BDC

Окончание таблицы А.6

Название	Повторное использование информации о происхождении из другой аналитической системы
Рисунок (опционально)	The diagram illustrates the reuse of origin information from another analytical system. It is divided into several sections: Top Section: Shows the interaction between 'Сервис-провайдер больших данных А' (Big Data Service Provider A) and 'Поставщик услуг больших данных В' (Big Data Service Provider B). A 'Запрос информации о происхождении' (Request for origin information) is sent from A to B, and 'Предоставление информации о происхождении' (Provision of origin information) is sent from B to A. Inside B, there is a 'Функция анализа информации о происхождении' (Origin information analysis function) that processes 'Данные результатов анализа' (Analysis results data) and 'Информация о конфигурации системы' (System configuration information) to produce 'Данные' (Data) and 'Функции' (Functions). A 'Потребитель сервиса больших данных' (Big Data Service Consumer) is shown applying 'Применение шагов процесса обработки' (Application of processing process steps). Bottom Section: A flowchart titled 'Шаги процесса, извлеченные из полученной информации о происхождении' (Process steps extracted from the received origin information). It shows a sequence of 'Данные' (Data) and 'Функция' (Function) nodes. A legend indicates three types of functions: 'Функция' (black oval) for 'соответствие найдено' (match found), 'Функция' (grey oval) for 'альтернативная функция' (alternative function), and 'Функция' (white oval) for 'нет соответствия' (no match). A large arrow labeled 'Шаги процесса отображения' (Process steps display) points to a 'Реконфигурирование шагов процесса для BDSP В' (Reconfiguration of process steps for BDSP B) block.
Предварительные условия (опционально)	—
Постусловия (опционально)	—
Производные требования	Описание модели информации о происхождении (см. 7.1); использование распространенного формата для обмена (см. 7.1); выделение рабочих процессов (см. 7.2); адаптивность вычислительной среды (см. 7.2); эквивалентные функции для шагов процесса (см. 7.2)

Таблица А.7 — Вариант использования — сбор информации о происхождении данных в ходе рабочего процесса научного исследования

Название	Сбор информации о происхождении данных в ходе рабочего процесса научного исследования и ее использование
Описание	Научный исследовательский процесс — типичная прикладная система, способствующая функционированию «электронной науки». Исследователи моделируют, проектируют, исполняют, отлаживают, реконфигурируют и повторно анализируют. Информация о происхождении данных в системе поддержки научного процесса очень полезна для исследователей, помогая им самим интерпретировать результаты своих исследований, а их коллегам — устанавливать доверие к результатам экспериментов. BDSP (система поддержки научного процесса) автоматически начинает документирование информации о происхождении и сохраняет ее в своей базе данных. BDC (научные работники, которые осуществляют деятельность по анализу данных, см. [1]), должны извлекать информацию о происхождении для подтверждения источника научных данных, использованного в процессе экспериментирования или анализа. BDSP автоматически сохраняет информацию о происхождении при изменении статуса данных с целью поддержки частого обращения к этой информации

Окончание таблицы А.7

Название	Сбор информации о происхождении данных в ходе рабочего процесса научного исследования и ее использование
Роли/суброли	BDSP BDC
Рисунок (опционально)	
Предварительные условия (опционально)	—
Постусловия (опционально)	—
Производные требования	Агрегирование информации о происхождении (см. 7.1); хранение информации о происхождении (см. 7.1)

Таблица А.8 — Вариант использования — извлечение рабочего процесса анализа из аккумулированной информации о происхождении

Название	Извлечение рабочего процесса анализа из аккумулированной информации о происхождении
Описание	BDSP A собирает информацию о происхождении у различных BDSP с целью накопления рабочих процессов анализа и их повторного использования. BDSP C хочет через BDSP A найти методы анализа, которые можно было бы применить с использованием функций и данных, имеющихся в его системе. Соответственно: - BDSP A получает информацию о происхождении от BDSP B; - BDSP A выделяет рабочие процессы из информации о происхождении; - BDSP A хранит рабочие процессы вместе с интегрированным графом, организованным на основе частоты использования функций анализа и взаимосвязей, отражающих последовательность их применения; - BDSP C запрашивает рабочий процесс со списком собственных функций и данных, а также сведения об операционной системе, региональных настройках, описание оборудования и т. п.; - BDSP A извлекает рабочий процесс на основе информации, полученной от BDSP C; - BDSP A отправляет список рабочих процессов-кандидатов BDSP C; - BDSP C выбирает рабочий процесс, который удовлетворяет цели анализа; - BDSP A реконструирует рабочий процесс в форме, в которой он может быть выполнен у BDSP C, и отправляет его BDSP C; - BDSP C использует рабочий процесс

Окончание таблицы А.8

Название	Извлечение рабочего процесса анализа из накопленной информации о происхождении
Роли/суброли	BDSP
Рисунок (опционально)	Сервис-провайдер больших данных В-1 Сервис-провайдер больших данных В-2 Сервис-провайдер больших данных В-3 Информация о происхождении Информация о происхождении Информация о происхождении Сервис-провайдер больших данных А Функция анализа информации о происхождении Выделение рабочих процессов из информации о происхождении. Сохранение рабочих процессов вместе с интегрированным графом Извлечение соответствующих условиям рабочих процессов. Отправка списка рабочих процессов-кандидатов. Реконструкция рабочих процессов, в т.ч. сопоставление функций для BDSP С. Экспорт рабочих процессов Функции Запрос рабочего процесса (вместе со списком собственных функций и данных, описанием вычислительной среды BDSP С) Список рабочих процессов-кандидатов Выбранные рабочие процессы Выполнимые у BDSP С рабочие процессы Поставщик услуг больших данных С
Предварительные условия (опционально)	—
Постусловия (опционально)	—
Производные требования	Описание модели информации о происхождении (см. 7.1); выделение рабочих процессов (см. 7.2); извлечение рабочего процесса (см. 7.2); сохранение рабочего процесса (см. 7.2); эквивалентные функции для шагов процесса (см. 7.2); адаптивность вычислительной среды (см. 7.2)

Приложение Б (справочное)

Примеры профилей информации о происхождении

В данном приложении для каталога данных используется профиль метаданных из [4]. Словарь каталогов данных W3C (DCAT) (см. [11]) используется в XML-профиле, в профиле Turtle (от Terse RDF Triple Language — «компактный синтаксис записи RDF-триплетов») и в сериализации на основе языка JSON для связанных данных (JSON-LD). Пространством имен для DCAT является «<http://www.w3.org/ns/dcat#>», однако DCAT широко использует термины из других словарей. Расширенным префиксом и пространством имен для этого приложения являются соответственно «bdp» и «<http://www.itu.int/xml-namespaces/itu-t/Y.3602/bigdataprovence#>».

Б.1 Профиль RDF

В данном разделе представлена XML-схема, у которой формат среды описания ресурсов (RDF) соответствует UML-модели в разделе 6.

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-
schema#" xmlns:bdp="http://www.itu.int/xml-namespaces/itu-t/Y.3602/bigdataprovence#" xmlns:bdc="http://www.itu.int/
xml-namespaces/itu-t/Y.3603/bigdatacatalogue#" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dct="http://purl.org/dc/terms/" xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:xsd="http://www.
w3.org/2001/XMLSchema#">
```

```
<rdfs:Class rdf:about="bdp:Application">
<rdfs:comment>information about the software providing the function used</rdfs:comment>
<rdfs:subClassOf rdf:resource="owl:Thing"/>
</rdfs:Class>

<rdfs:Class rdf:about="bdp:ComputationalEnvironment">
<rdfs:comment>information about the computing environment</rdfs:comment>
<rdfs:subClassOf rdf:resource="owl:Thing"/>
</rdfs:Class>

<rdfs:Class rdf:about="bdp:Dataset">
<rdfs:comment>a collection of data which is recorded in a provenance unit</rdfs:comment>
<rdfs:subClassOf rdf:resource="owl:Thing"/>
</rdfs:Class>

<rdfs:Class rdf:about="bdp:Function">
<rdfs:comment>information about the used function</rdfs:comment>
<rdfs:subClassOf rdf:resource="owl:Thing"/>
</rdfs:Class>

<rdfs:Class rdf:about="bdp:HardwareSpecType">
<rdfs:comment>information required to describe the hardware environment</rdfs:comment>
<rdfs:subClassOf rdf:resource="owl:Thing"/>
</rdfs:Class>

<rdfs:Class rdf:about="bdp:LocaleSettingType">
<rdfs:comment>information required to describe locale setting value</rdfs:comment>
<rdfs:subClassOf rdf:resource="owl:Thing"/>
</rdfs:Class>

<rdfs:Class rdf:about="bdp:ProvenanceInformation">
<rdfs:comment>root entity which defines PI</rdfs:comment>
<rdfs:subClassOf rdf:resource="owl:Thing"/>
</rdfs:Class>

<rdfs:Class rdf:about="bdp:ProvenanceUnit">
<rdfs:comment>a record of the most recent changes to the data</rdfs:comment>
<rdfs:subClassOf rdf:resource="owl:Thing"/>
</rdfs:Class>
```



```

<rdf:Property rdf:about="bdp:accelerationIO">
<rdfs:comment>information about hardware acceleration unit (e.g., GPU)</rdfs:comment>
<rdfs:domain rdf:resource="bdp:HardwareSpecType"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:applicationInfo">
<rdfs:comment>information about the software providing the function used</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Function"/>
<rdfs:range rdf:resource="bdp:Application"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:applicationName">
<rdfs:comment>name of software application</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Application"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:availability">
<rdfs:comment>information about whether the data exists in the storage or has been deleted</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Dataset"/>
<rdfs:range rdf:resource="xsd:boolean"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:computationalEnvironment">
<rdfs:comment>computational environment information for the provenance unit</rdfs:comment>
<rdfs:domain rdf:resource="bdp:ProvenanceUnit"/>
<rdfs:range rdf:resource="bdp:ComputationalEnvironment"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:country">
<rdfs:comment>a country code</rdfs:comment>
<rdfs:domain rdf:resource="bdp:LocaleSettingType"/>
<rdfs:range rdf:resource="dct:ISO3166"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:cpuInfo">
<rdfs:comment>information about speed and number of units of CPU</rdfs:comment>
<rdfs:domain rdf:resource="bdp:HardwareSpecType"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:dataset">
<rdfs:comment>dataset information for the data catalogue</rdfs:comment>
<rdfs:domain rdf:resource="bdp:ProvenanceUnit"/>
<rdfs:range rdf:resource="bdp:Dataset"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:datasetMetadata">
<rdfs:comment>dataset metadata for the dataset</rdfs:comment>

<rdfs:domain rdf:resource="bdp:Dataset"/>
<rdfs:range rdf:resource="bdp:Dataset"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:dsId">
<rdfs:comment>a unique identifier of the dataset</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Dataset"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:encoding">

```

```

<rdfs:comment>the value for character encoding setting</rdfs:comment>
<rdfs:domain rdf:resource="bdp:LocaleSettingType"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:followedFunction">
<rdfs:comment>information about the function used after</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Function"/>
<rdfs:range rdf:resource="bdp:Function"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:function">
<rdfs:comment>function information for the data catalogue</rdfs:comment>
<rdfs:domain rdf:resource="bdp:ProvenanceUnit"/>
<rdfs:range rdf:resource="bdp:Function"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:functionId">
<rdfs:comment>a unique function name</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Function"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:functionName">
<rdfs:comment>the alternative name of the function</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Function"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:hardwareSpecs">
<rdfs:comment>information about the hardware specification</rdfs:comment>
<rdfs:domain rdf:resource="bdp:ComputationalEnvironment"/>
<rdfs:range rdf:resource="bdp:HardwareSpecType"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:hasPII">
<rdfs:comment>information about whether the data includes PII or not</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Dataset"/>
<rdfs:range rdf:resource="xsd:boolean"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:inputParaValue">
<rdfs:comment>the input parameter value</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Function"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:installUri">
<rdfs:comment>URI at which the app may be installed</rdfs:comment>

<rdfs:domain rdf:resource="bdp:Application"/>
<rdfs:range rdf:resource="xsd:anyURI"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:language">
<rdfs:comment>a natural language used for hardware</rdfs:comment>
<rdfs:domain rdf:resource="bdp:LocaleSettingType"/>
<rdfs:range rdf:resource="dct:RFC4646"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:localeSetting">
<rdfs:comment>information about language, display formats of time, date, currency, etc.</rdfs:comment>

```

```

<rdfs:domain rdf:resource="»bdp:ComputationalEnvironment»"/>
<rdfs:range rdf:resource="bdp:LocaleSettingType"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:memoryInfo">
<rdfs:comment>information about size and speed of memory</rdfs:comment>
<rdfs:domain rdf:resource="bdp:HardwareSpecType"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:operatingSystem">
<rdfs:comment>information about the operating system</rdfs:comment>
<rdfs:domain rdf:resource="»bdp:ComputationalEnvironment»"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:provenanceUnit">
<rdfs:comment>provenance unit for the PI</rdfs:comment>
<rdfs:domain rdf:resource="bdp:ProvenanceInformation"/>
<rdfs:range rdf:resource="bdp:ProvenanceUnit"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:responsibleParty">
<rdfs:comment>responsible party information for the data catalogue</rdfs:comment>
<rdfs:domain rdf:resource="bdp:ProvenanceUnit"/>
<rdfs:range rdf:resource="bdc:ResponsibleParty"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:softwareVersion">
<rdfs:comment>version of software</rdfs:comment>
<rdfs:domain rdf:resource="bdp:Application"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:storageInfo">
<rdfs:comment>information about size and speed of storage</rdfs:comment>
<rdfs:domain rdf:resource="bdp:HardwareSpecType"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:storedDate">
<rdfs:comment>the date of storing the provenance unit</rdfs:comment>
<rdfs:domain rdf:resource="»bdp:ProvenanceUnit»"/>
<rdfs:range rdf:resource="dct:W3CDTF"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:timeZone">
<rdfs:comment>information of time zone</rdfs:comment>

<rdfs:domain rdf:resource="bdp:ComputationalEnvironment"/>
<rdfs:range rdf:resource="dct:W3CDTF"/>
</rdf:Property>

<rdf:Property rdf:about="bdp:unitId">
<rdfs:comment>a unique provenance unit name</rdfs:comment>
<rdfs:domain rdf:resource="»bdp:ProvenanceUnit»"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Property>

<rdf:Description rdf:about="dc:description">
<rdfs:comment>a short introduction to the function or the application </rdfs:comment>
<rdfs:domain rdf:resource="bdp:Application"/>

```



```
<rdfs:domain rdf:resource="bdp:Function"/>
<rdfs:range rdf:resource="xsd:string"/>
</rdf:Description>
```

```
</rdf:RDF>
```

Б.2 Профиль Turtle

В данном разделе представлена схема формата Turtle, соответствующая UML-модели в разделе 6.

```
# baseURI: http://www.itu.int/xml-namespace/itu-t/Y.3602/bigdataprovence
# imports: http://purl.org/dc/elements/1.1/
# imports: http://purl.org/dc/terms/
# imports: https://www.w3.org/2002/07/owl#
# imports: http://www.itu.int/xml-namespace/itu-t/Y.3603/bigdatacatalogue
# prefix: bdp

@prefix bdp: <http://www.itu.int/xml-namespace/itu-t/Y.3602/bigdataprovence#> .
@prefix bdc: <http://www.itu.int/xml-namespace/itu-t/Y.3603/bigdatacatalogue#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

bdp:Application rdf:type rdfs:Class ;
rdfs:comment "information about the software providing the function used" ; rdfs:subClassOf owl:Thing ;
.
bdp:ComputationalEnvironment rdf:type rdfs:Class ;
rdfs:comment "information about the computing environment" ; rdfs:subClassOf owl:Thing ;
.
bdp:Dataset
rdf:type rdfs:Class ;
rdfs:comment "a collection of data which is recorded in a provenance unit" ; rdfs:subClassOf owl:Thing ;
.
bdp:Function rdf:type rdfs:Class ;
rdfs:comment "information about the used function " ; rdfs:subClassOf owl:Thing ;
.
bdp:HardwareSpecType rdf:type rdfs:Class ;
rdfs:comment "information required to describe the hardware environment" ; rdfs:subClassOf owl:Thing ;
.
bdp:LocaleSettingType rdf:type rdfs:Class ;
rdfs:comment "information required to describe locale setting value " ; rdfs:subClassOf owl:Thing ;
.
bdp:ProvenanceInformation rdf:type rdfs:Class ;
rdfs:comment "root entity which defines PI" ; rdfs:subClassOf owl:Thing ;
.
bdp:ProvenanceUnit rdf:type rdfs:Class ;
rdfs:comment "a record of the most recent changes to the data" ; rdfs:subClassOf owl:Thing ;
.
bdp:accelerationIO rdf:type rdf:Property ;
rdfs:comment "information about hardware acceleration unit (e.g., GPU)" ; rdfs:domain bdp:HardwareSpecType ;
rdfs:range xsd:string ;
.
bdp:applicationInfo rdf:type rdf:Property ;
rdfs:comment "information about the software providing the function used" ; rdfs:domain bdp:Function ;
rdfs:range bdp:Application ;
.
bdp:applicationName rdf:type rdf:Property ;
rdfs:comment "name of software application" ; rdfs:domain bdp:Application ;
rdfs:range xsd:string ;
.
```

```

bdp:availability rdf:type rdf:Property ;
rdfs:comment "information about whether the data exists in the storage or has been deleted" ; rdfs:domain
bdp:Dataset ;
rdfs:range xsd:boolean ;
.
bdp:computationalEnvironment rdf:type rdf:Property ;
rdfs:comment "computational environment information for the provenance unit" ; rdfs:domain bdp:ProvenanceUnit ;
rdfs:range bdp:ComputationalEnvironment ;
.
bdp:country
rdf:type rdf:Property ; rdfs:comment "a country code" ;
rdfs:domain bdp:LocaleSettingType ; rdfs:range dct:ISO3166 ;
.
bdp:cpuInfo
rdf:type rdf:Property ;
rdfs:comment "information about speed and number of units of CPU" ; rdfs:domain bdp:HardwareSpecType ;
rdfs:range xsd:string ;
.
bdp:dataset
rdf:type rdf:Property ;
rdfs:comment "dataset information for the data catalogue" ; rdfs:domain bdp:ProvenanceUnit ;
rdfs:range bdp:Dataset ;
.
bdp:datasetMetadata rdf:type rdf:Property ;
rdfs:comment "dataset metadata for the dataset" ; rdfs:domain bdp:Dataset ;
rdfs:range bdc:Dataset ;
.
bdp:dslId
rdf:type rdf:Property ;
rdfs:comment "a unique identifier of the dataset" ; rdfs:domain bdp:Dataset ;
rdfs:range xsd:string ;
.
bdp:encoding
rdf:type rdf:Property ;
rdfs:comment "the value for character encoding setting" ; rdfs:domain bdp:LocaleSettingType ;
rdfs:range xsd:string ;
.
bdp:followedFunction rdf:type rdf:Property ;
rdfs:comment "information about the function used after" ; rdfs:domain bdp:Function ;
rdfs:range bdp:Function ;
.
bdp:function
rdf:type rdf:Property ;
rdfs:comment "function information for the data catalogue" ; rdfs:domain bdp:ProvenanceUnit ;
rdfs:range bdp:Function ;
.
bdp:functionId rdf:type rdf:Property ;
rdfs:comment "a unique function name" ; rdfs:domain bdp:Function ;
rdfs:range xsd:string ;
.
bdp:functionName rdf:type rdf:Property ;
rdfs:comment "the alternative name of the function" ; rdfs:domain bdp:Function ;
rdfs:range xsd:string ;
.
bdp:hardwareSpecs rdf:type rdf:Property ;
rdfs:comment "information about the hardware specification" ; rdfs:domain bdp:ComputationalEnvironment ;
rdfs:range bdp:HardwareSpecType ;
.
bdp:hasPII
rdf:type rdf:Property ;
rdfs:comment "information about whether the data includes PII or not" ; rdfs:domain bdp:Dataset ;

```

```

    rdfs:range xsd:boolean ;
.
    bdp:inputParaValue rdf:type rdf:Property ;
    rdfs:comment "the input parameter value" ; rdfs:domain bdp:Function ;
    rdfs:range xsd:string ;
.
    bdp:installUri
    rdf:type rdf:Property ;
    rdfs:comment "URI at which the app may be installed" ; rdfs:domain bdp:Application ;
    rdfs:range xsd:anyURI
.
    bdp:language
    rdf:type rdf:Property ;
    rdfs:comment "a natural language used for hardware" ; rdfs:domain bdp:LocaleSettingType ;
    rdfs:range dct:RFC4646 ;
.
    bdp:localeSetting rdf:type rdf:Property ;
    rdfs:comment "information about language, display formats of time, date, currency, etc." ; rdfs:domain
    bdp:ComputationalEnvironment ;
    rdfs:range bdp:LocaleSettingType ;
.
    bdp:memoryInfo rdf:type rdf:Property ;
    rdfs:comment "information about size and speed of memory" ; rdfs:domain bdp:HardwareSpecType ;
    rdfs:range xsd:string ;
.
    bdp:operatingSystem rdf:type rdf:Property ;
    rdfs:comment "information about the operating system" ; rdfs:domain bdp:ComputationalEnvironment ;
    rdfs:range xsd:string ;
.
    bdp:provenanceUnit rdf:type rdf:Property ;
    rdfs:comment "provenance unit for the PI" ; rdfs:domain bdp:ProvenanceInformation ; rdfs:range bdp:ProvenanceUnit ;
.
    bdp:responsibleParty rdf:type rdf:Property ;
    rdfs:comment "responsible party information for the data catalogue" ; rdfs:domain bdp:ProvenanceUnit ;
    rdfs:range bdc:ResponsibleParty ;
.
    bdp:softwareVersion rdf:type rdf:Property ;
    rdfs:comment "version of software" ; rdfs:domain bdp:Application ; rdfs:range xsd:string ;
.
    bdp:storageInfo rdf:type rdf:Property ;
    rdfs:comment "information about size and speed of storage" ; rdfs:domain bdp:HardwareSpecType ;

    rdfs:range xsd:string ;
.
    bdp:storedDate rdf:type rdf:Property ;
    rdfs:comment "the date of storing the provenance unit" ; rdfs:domain bdp:ProvenanceUnit ;
    rdfs:range dct:W3CDTF ;
.
    bdp:timeZone
    rdf:type rdf:Property ;
    rdfs:comment "information of time zone" ; rdfs:domain bdp:ComputationalEnvironment ; rdfs:range dct:W3CDTF ;
.
    bdp:unitId
    rdf:type rdf:Property ;
    rdfs:comment "a unique provenance unit name" ; rdfs:domain bdp:ProvenanceUnit ;
    rdfs:range xsd:string ;
.
    dc:description
    rdfs:comment "a short introduction to the function or the application " ; rdfs:domain bdp:Application ;
    rdfs:domain bdp:Function ; rdfs:range xsd:string ;

```


Б.3 Профиль JSON-LD

В данном разделе представлена схема формата JSON-LD, соответствующая UML-модели в разделе 6.

```
{
  "@context": {
    "bdp": "http://www.itu.int/xml-namespace/itu-t/Y.3602/bigdataprovence#",
    "bdc": "http://www.itu.int/xml-namespace/itu-t/Y.3603/bigdatacatalogue#",
    "dc": "http://purl.org/dc/elements/1.1/",
    "dct": "http://purl.org/dc/terms/",
    "owl": "http://www.w3.org/2002/07/owl#",
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "xsd": "http://www.w3.org/2001/XMLSchema#"
  },
  "@graph": [
    {
      "@id": "bdp:Application",
      "rdf:type": {
        "@id": "rdfs:Class"
      },
      "rdfs:comment": "information about the software providing the function used",
      "rdfs:subClassOf": {
        "@id": "owl:Thing"
      }
    },
    {
      "@id": "bdp:ComputationalEnvironment",
      "rdf:type": {
        "@id": "rdfs:Class"
      },
      "rdfs:comment": "information about the computing environment",
      "rdfs:subClassOf": {
        "@id": "owl:Thing"
      }
    },
    {
      "@id": "bdp:Dataset",
      "rdf:type": {
        "@id": "rdfs:Class"
      },
      "rdfs:comment": "a collection of data which is recorded in a provenance unit",
      "rdfs:subClassOf": {
        "@id": "owl:Thing"
      }
    },
    {
      "@id": "bdp:Function",
      "rdf:type": {
        "@id": "rdfs:Class"
      },
      "rdfs:comment": "information about the used function",
      "rdfs:subClassOf": {
        "@id": "owl:Thing"
      }
    },
    {
      "@id": "bdp:HardwareSpecType",
      "rdf:type": {
        "@id": "rdfs:Class"
      },
      "rdfs:comment": "information required to describe the hardware environment",
      "rdfs:subClassOf": {
        "@id": "owl:Thing"
      }
    },
    {
      "@id": "bdp:LocaleSettingType",
      "rdf:type": {
        "@id": "rdfs:Class"
      },
      "rdfs:comment": "information required to describe locale setting value",
      "rdfs:subClassOf": {
        "@id": "owl:Thing"
      }
    }
  ]
}
```

```

    }
  },
  {
    "@id": "bdp:ProvenanceInformation", "rdf:type": {
      "@id": "rdfs:Class"
    },
    "rdfs:comment": "root entity which defines PI", "rdfs:subClassOf": {
      "@id": "owl:Thing"
    }
  },
  {
    "@id": "bdp:ProvenanceUnit", "rdf:type": {
      "@id": "rdfs:Class"
    },
    "rdfs:comment": "a record of the most recent changes to the data", "rdfs:subClassOf": {
      "@id": "owl:Thing"
    }
  },
  {
    "@id": "bdp:accelerationIO", "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "information about hardware acceleration unit (e.g., GPU)", "rdfs:domain": {
      "@id": "bdp:HardwareSpecType"
    },
    "rdfs:range": {
      "@id": "xsd:string"
    }
  },
  {
    "@id": "bdp:applicationInfo", "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "information about the software providing the function used", "rdfs:domain": {
      "@id": "bdp:Function"
    },
    "rdfs:range": {
      "@id": "bdp:Application"
    }
  },
  {
    "@id": "bdp:applicationName", "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "name of software application", "rdfs:domain": {
      "@id": "bdp:Application"
    },
    "rdfs:range": {
      "@id": "xsd:string"
    }
  },
  {
    "@id": "bdp:availability", "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "information about whether the data exists in the storage or has been deleted", "rdfs:domain": {
      "@id": "bdp:Dataset"
    },
    "rdfs:range": {
      "@id": "xsd:boolean"
    }
  }
}

```

```

},
{
  "@id": "bdp:computationalEnvironment", "rdf:type": {
    "@id": "rdf:Property"
  },
  "rdfs:comment": "computational environment information for the provenance unit", "rdfs:domain": {
    "@id": "bdp:ProvenanceUnit"
  },
  "rdfs:range": {
    "@id": "bdp:ComputationalEnvironment"
  },
},
{
  "@id": "bdp:country",
  "rdf:type": {
    "@id": "rdf:Property"
  },
  "rdfs:comment": "a country code", "rdfs:domain": {
    "@id": "bdp:LocaleSettingType"
  },
  "rdfs:range": {
    "@id": "dct:ISO3166"
  },
},
{
  "@id": "bdp:cpuInfo",
  "rdf:type": {
    "@id": "rdf:Property"
  },
  "rdfs:comment": "information about speed and number of units of CPU", "rdfs:domain": {
    "@id": "bdp:HardwareSpecType"
  },
  "rdfs:range": {
    "@id": "xsd:string"
  },
},
{
  "@id": "bdp:dataset",
  "rdf:type": {
    "@id": "rdf:Property"
  },
  "rdfs:comment": "dataset information for the data catalogue", "rdfs:domain": {
    "@id": "bdp:ProvenanceUnit"
  },
  "rdfs:range": {
    "@id": "bdp:Dataset"
  },
},
{
  "@id": "bdp:datasetMetadata", "rdf:type": {
    "@id": "rdf:Property"
  },
  "rdfs:comment": "dataset metadata for the dataset", "rdfs:domain": {
    "@id": "bdp:Dataset"
  },
  "rdfs:range": {
    "@id": "bdc:Dataset"
  },
},
{

```



```

"@id": "bdp:dsId",
"rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "a unique identifier of the dataset", "rdfs:domain": {
"@id": "bdp:Dataset"
},
"rdfs:range": {
"@id": "xsd:string"
}
},
{
"@id": "bdp:encoding",
"rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "the value for character encoding setting", "rdfs:domain": {
"@id": "bdp:LocaleSettingType"
},
"rdfs:range": {
"@id": "xsd:string"
}
},
{
"@id": "bdp:followedFunction", "rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "information about the function used after", "rdfs:domain": {
"@id": "bdp:Function"
},
"rdfs:range": {
"@id": "bdp:Function"
}
},
{
"@id": "bdp:function",
"rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "function information for the data catalogue", "rdfs:domain": {
"@id": "bdp:ProvenanceUnit"
},
"rdfs:range": {
"@id": "bdp:Function"
}
},
{
"@id": "bdp:functionId",
"rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "a unique function name", "rdfs:domain": {
"@id": "bdp:Function"
},
"rdfs:range": {
"@id": "xsd:string"
}
},
{
"@id": "bdp:functionName", "rdf:type": {
"@id": "rdf:Property"

```

```

},
"rdfs:comment": "the alternative name of the function", "rdfs:domain": {
"@id": "bdp:Function"
},
"rdfs:range": {
"@id": "xsd:string"
},
},
{
"@id": "bdp:hardwareSpecs", "rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "information about the hardware specification", "rdfs:domain": {
"@id": "bdp:ComputationalEnvironment"
},
"rdfs:range": {
"@id": "bdp:HardwareSpecType"
},
},
{
"@id": "bdp:hasPII",
"rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "information about whether the data includes PII or not", "rdfs:domain": {
"@id": "bdp:Dataset"
},
"rdfs:range": {
"@id": "xsd:boolean"
},
},
{
"@id": "bdp:inputParaValue", "rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "the input parameter value", "rdfs:domain": {
"@id": "bdp:Function"
},
"rdfs:range": {
"@id": "xsd:string"
},
},
{
"@id": "bdp:installUri",
"rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "URI at which the app may be installed", "rdfs:domain": {
"@id": "bdp:Application"
},
"rdfs:range": {
"@id": "xsd:anyURI"
},
},
{
"@id": "bdp:language",
"rdf:type": {
"@id": "rdf:Property"
},
"rdfs:comment": "a natural language used for hardware", "rdfs:domain": {
"@id": "bdp:LocaleSettingType"
}
}

```

```

    },
    "rdfs:range": {
      "@id": "dct:RFC4646"
    }
  },
  {
    "@id": "bdp:localeSetting", "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "information about language, display formats of time, date, currency, etc.", "rdfs:domain": {
      "@id": "bdp:ComputationalEnvironment"
    },
    "rdfs:range": {
      "@id": "bdp:LocaleSettingType"
    }
  },
  {
    "@id": "bdp:memoryInfo",
    "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "information about size and speed of memory", "rdfs:domain": {
      "@id": "bdp:HardwareSpecType"
    },
    "rdfs:range": {
      "@id": "xsd:string"
    }
  },
  {
    "@id": "bdp:operatingSystem", "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "information about the operating system", "rdfs:domain": {
      "@id": "bdp:ComputationalEnvironment"
    },
    "rdfs:range": {
      "@id": "xsd:string"
    }
  },
  {
    "@id": "bdp:provenanceUnit", "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "provenance unit for the PI", "rdfs:domain": {
      "@id": "bdp:ProvenanceInformation"
    },
    "rdfs:range": {
      "@id": "bdp:ProvenanceUnit"
    }
  },
  {
    "@id": "bdp:responsibleParty", "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "responsible party information for the data catalogue", "rdfs:domain": {
      "@id": "bdp:ProvenanceUnit"
    },
    "rdfs:range": {
      "@id": "bdc:ResponsibleParty"
    }
  },

```



```

{
  "@id": "bdp:softwareVersion", "rdf:type": {
    "@id": "rdf:Property"
  },
  "rdfs:comment": "version of software", "rdfs:domain": {
    "@id": "bdp:Application"
  },
  "rdfs:range": {
    "@id": "xsd:string"
  },
  },
  {
    "@id": "bdp:storageInfo", "rdf:type": {
      "@id": "rdf:Property"
    },
    "rdfs:comment": "information about size and speed of storage", "rdfs:domain": {
      "@id": "bdp:HardwareSpecType"
    },
    "rdfs:range": {
      "@id": "xsd:string"
    },
    },
    {
      "@id": "bdp:storedDate",
      "rdf:type": {
        "@id": "rdf:Property"
      },
      "rdfs:comment": "the date of storing the provenance unit",
      "rdfs:domain": {
        "@id": "bdp:ProvenanceUnit"
      },
      "rdfs:range": {
        "@id": "dct:W3CDTF"
      },
      },
      {
        "@id": "bdp:timeZone",
        "rdf:type": {
          "@id": "rdf:Property"
        },
        "rdfs:comment": "information of time zone", "rdfs:domain": {
          "@id": "bdp:ComputationalEnvironment"
        },
        "rdfs:range": {
          "@id": "dct:W3CDTF"
        },
        },
        {
          "@id": "bdp:unitId",
          "rdf:type": {
            "@id": "rdf:Property"
          },
          "rdfs:comment": "a unique provenance unit name", "rdfs:domain": {
            "@id": "bdp:ProvenanceUnit"
          },
          "rdfs:range": {
            "@id": "xsd:string"
          },
          },
          {
            "@id": "dc:description",

```

```

"rdfs:comment": "a short introduction to the function or the application ", "rdfs:domain": [
{
"@id": "bdp:Application"
},
{
"@id": "bdp:Function"
}
],
"rdfs:range": {
"@id": "xsd:string"
}
}
]
}

```

Библиография

- [1] ITU-T Y.3600 (2015) Большие данные. Требования и возможности на базе облачных вычислений (Big data — Cloud computing based requirements and capabilities)
- [2] ITU-T X.1255 (2013) Структура обнаружения информации по управлению определением идентичности (Framework for discovery of identity management information)
- [3] ITU-T Y.4100 / Y.2066 (2014) Общие требования к интернету вещей (Common requirements of the Internet of things)
- [4] ITU-T Y.3603 (2019) Большие данные. Требования и концептуальная модель метаданных для каталога данных (Big data — Requirements and conceptual model of metadata for data catalogue)
- [5] ITU-T X.680 (2021) Информационные технологии. Абстрактная синтаксическая нотация версии 1. Спецификация базовой нотации (Information technology — Abstract Syntax Notation One (ASN.1): Specification of basic notation)
- [6] ИСО 8601-2:2019 Дата и время. Представление для обмена информацией. Часть 2. Расширения (Date and time — Representations for information interchange — Part 2: Extensions)
- [7] ИСО 639-1:2002 Коды для представления названий языков. Часть 1. Двухбуквенный код (Codes for the representation of names of languages — Part 1: Alpha-2 code)
- [8] ИСО 3166-1:2020 Коды для представления названий стран и единиц их административно-территориального деления. Часть 1. Коды стран (Codes for the representation of names of countries and their subdivisions — Part 1: Country code)
- [9] ITU-T Y.2201 (2009) Требования МСЭ-Т к сетям последующих поколений и возможности этих сетей (Requirements and capabilities for ITU-T NGN)
- [10] ITU-T Y.2701 (2007) Требования к безопасности для сетей последующих поколений версии 1 (Security requirements for NGN release 1)
- [11] W3C DCAT Рекомендация Консорциума Всемирной паутины (20200214) «Словарь каталогов данных. Версия 2» (W3C Recommendation (20200214) «Data catalog vocabulary (DCAT) version 2»)

Ключевые слова: информационные технологии; искусственный интеллект; большие данные; происхождение больших данных, провайдер данных, поставщик данных, брокер данных, сервис-провайдер больших данных, жизненный цикл больших данных

Редактор *Л.В. Коретникова*
Технический редактор *В.Н. Прусакова*
Корректор *Е.Д. Дульнева*
Компьютерная верстка *Е.О. Асташина*

Сдано в набор 14.12.2023. Подписано в печать 12.01.2024. Формат 60×84%. Гарнитура Ариал.
Усл. печ. л. 5,12. Уч.-изд. л. 4,10.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

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