Международный союз электросвязи

МСЭ-Т

СЕКТОР СТАНДАРТИЗАЦИИ
ЭЛЕКТРОСВЯЗИ МСЭ

X.1148

(09/2020)

СЕРИЯ X: СЕТИ ПЕРЕДАЧИ ДАННЫХ,
ВЗАИМОСВЯЗЬ ОТКРЫТЫХ СИСТЕМ
И БЕЗОПАСНОСТЬ

Безопасные приложения и услуги (1) –
Безопасность веб-среды

Структура процесса деидентификации для поставщиков услуг электросвязи

Рекомендация МСЭ-Т X.1148

ITU-T

Международный союз электросвязи

РЕКОМЕНДАЦИИ МСЭ-Т СЕРИИ X

СЕТИ ПЕРЕДАЧИ ДАННЫХ, ВЗАИМОСВЯЗЬ ОТКРЫТЫХ СИСТЕМ И БЕЗОПАСНОСТЬ

СЕТИ ПЕРЕДАЧИ ДАННЫХ ОБЩЕГО ПОЛЬЗОВАНИЯ	X.1–X.199
ВЗАИМОСВЯЗЬ ОТКРЫТЫХ СИСТЕМ	X.200–X.299
ВЗАИМОДЕЙСТВИЕ МЕЖДУ СЕТЯМИ	X.300–X.399
СИСТЕМЫ ОБРАБОТКИ СООБЩЕНИЙ	X.400–X.499
СПРАВОЧНИК	X.500–X.599
ОРГАНИЗАЦИЯ СЕТИ ВОС И СИСТЕМНЫЕ АСПЕКТЫ	X.600–X.699
УПРАВЛЕНИЕ В ВОС	X.700–X.799
БЕЗОПАСНОСТЬ	X.800–X.849
ПРИЛОЖЕНИЯ ВОС	X.850–X.899
ОТКРЫТАЯ РАСПРЕДЕЛЕННАЯ ОБРАБОТКА	X.900–X.999
БЕЗОПАСНОСТЬ ИНФОРМАЦИИ И СЕТЕЙ	
Общие аспекты безопасности	X.1000–X.1029
Безопасность сетей	X.1030–X.1049
Управление безопасностью	X.1050–X.1069
Телебиометрия	X.1080–X.1099
БЕЗОПАСНЫЕ ПРИЛОЖЕНИЯ И УСЛУГИ (1)	
Безопасность многоадресной передачи	X.1100–X.1109
Безопасность домашних сетей	X.1110–X.1119
Безопасность подвижной связи	X.1120–X.1139
Безопасность веб-среды	**X.1140–X.1149**
Протоколы безопасности (1)	X.1150–X.1159
Безопасность одноранговых сетей	X.1160–X.1169
Безопасность сетевой идентификации	X.1170–X.1179
Безопасность IPTV	X.1180–X.1199
БЕЗОПАСНОСТЬ КИБЕРПРОСТРАНСТВА	
Кибербезопасность	X.1200–X.1229
Противодействие спаму	X.1230–X.1249
Управление определением идентичности	X.1250–X.1279
БЕЗОПАСНЫЕ ПРИЛОЖЕНИЯ И УСЛУГИ (2)	
Связь в чрезвычайных ситуациях	X.1300–X.1309
Безопасность повсеместных сенсорных сетей	X.1310–X.1319
Безопасность "умных" электросетей	X.1330–X.1339
Сертифицированная электронная почта	X.1340–X.1349
Безопасность интернета вещей (IoT)	X.1360–X.1369
Безопасность интеллектуальных транспортных систем (ИТС)	X.1370–X.1379
Безопасность технологии распределенного реестра	X.1400–X.1429
Безопасность технологии распределенного реестра	X.1430–X.1449
Протоколы безопасности (2)	X.1450–X.1459
ОБМЕН ИНФОРМАЦИЕЙ, КАСАЮЩЕЙСЯ КИБЕРБЕЗОПАСНОСТИ	
Обзор кибербезопасности	X.1500–X.1519
Обмен информацией об уязвимости/состоянии	X.1520–X.1539
Обмен информацией о событии/инциденте/эвристических правилах	X.1540–X.1549
Обмен информацией о политике	X.1550–X.1559
Эвристические правила и запрос информации	X.1560–X.1569
Идентификация и обнаружение	X.1570–X.1579
Гарантированный обмен	X.1580–X.1589
БЕЗОПАСНОСТЬ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ	
Обзор безопасности облачных вычислений	X.1600–X.1601
Проектирование безопасности облачных вычислений	X.1602–X.1639
Передовой опыт и руководящие указания в области облачных вычислений	X.1640–X.1659
Обеспечение безопасности облачных вычислений	X.1660–X.1679
Другие вопросы безопасности облачных вычислений	X.1680–X.1699
КВАНТОВАЯ СВЯЗЬ	
Терминология	X.1700–X.1701
Квантовый генератор случайных чисел	X.1702–X.1709
Структура безопасности QKDN	X.1710–X.1711
Проектирование безопасности QKDN	X.1712–X.1719
Методы обеспечения безопасности QKDN	X.1720–X.1729
БЕЗОПАСНОСТЬ ДАННЫХ	
Безопасность больших данных	X.1750–X.1759
БЕЗОПАСНОСТЬ СЕТЕЙ 5G	X.1800–X.1819

Для получения более подробной информации просьба обращаться к перечню Рекомендаций МСЭ-Т.

Рекомендация МСЭ-Т X.1148

Структура процесса деидентификации для поставщиков услуг электросвязи

Резюме

Организации электросвязи осуществляют сбор данных о физических лицах, включая информацию, позволяющую установить личность, управление ими, их использование и обмен такими данными. Для того чтобы защитить эти персональные данные, они используют методы деидентификации (удаления идентификационной информации). В Рекомендации МСЭ-Т X.1148 описана структура процесса деидентификации, включая практические шаги, и на основе модели жизненного цикла данных и ролей участников определены модели публикации данных и этапы данных в процессе деидентификации для поставщиков услуг электросвязи.

Хронологическая справка

Издание	Рекомендация	Утверждение	Исследовательская комиссия	Уникальный идентификатор*
1.0	МСЭ-Т X.1148	03.09.2020 г.	17-я	11.1002/1000/14249

Ключевые слова

Деидентификация, защита PII, субъект данных, процесс деидентификации, модели публикации данных, k-анонимность, l-разнообразие, t-плотность.

* Для доступа к Рекомендации наберите в адресном поле вашего веб-навигатора URL http://handle.itu.int/, после которого укажите уникальный идентификатор Рекомендации. Например: http://handle.itu.int/11.1002/1000/11830-en.

ПРЕДИСЛОВИЕ

Международный союз электросвязи (МСЭ) является специализированным учреждением Организации Объединенных Наций в области электросвязи и информационно-коммуникационных технологий (ИКТ). Сектор стандартизации электросвязи МСЭ (МСЭ-Т) – постоянный орган МСЭ. МСЭ-Т отвечает за изучение технических, эксплуатационных и тарифных вопросов и за выпуск Рекомендаций по ним с целью стандартизации электросвязи на всемирной основе.

На Всемирной ассамблее по стандартизации электросвязи (ВАСЭ), которая проводится каждые четыре года, определяются темы для изучения исследовательскими комиссиями МСЭ-Т, которые, в свою очередь, вырабатывают Рекомендации по этим темам.

Утверждение Рекомендаций МСЭ-Т осуществляется в соответствии с процедурой, изложенной в Резолюции 1 ВАСЭ.

В некоторых областях информационных технологий, которые входят в компетенцию МСЭ-Т, необходимые стандарты разрабатываются на основе сотрудничества с ИСО и МЭК.

ПРИМЕЧАНИЕ

В настоящей Рекомендации термин "администрация" используется для краткости и обозначает как администрацию электросвязи, так и признанную эксплуатационную организацию.

Соблюдение положений данной Рекомендации осуществляется на добровольной основе. Однако данная Рекомендация может содержать некоторые обязательные положения (например, для обеспечения функциональной совместимости или возможности применения), и в таком случае соблюдение Рекомендации достигается при выполнении всех указанных положений. Для выражения требований используются слова "следует", "должен" ("shall") или некоторые другие обязывающие выражения, такие как "обязан" ("must"), а также их отрицательные формы. Употребление таких слов не означает, что от какой-либо стороны требуется соблюдение положений данной Рекомендации.

ПРАВА ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

МСЭ обращает внимание на вероятность того, что практическое применение или выполнение настоящей Рекомендации может включать использование заявленного права интеллектуальной собственности. МСЭ не занимает какую бы то ни было позицию относительно подтверждения, действительности или применимости заявленных прав интеллектуальной собственности, независимо от того, доказываются ли такие права членами МСЭ или другими сторонами, не относящимися к процессу разработки Рекомендации.

На момент утверждения настоящей Рекомендации МСЭ не получил извещения об интеллектуальной собственности, защищенной патентами, которые могут потребоваться для выполнения настоящей Рекомендации. Однако те, кто будет применять Рекомендацию, должны иметь в виду, что вышесказанное может не отражать самую последнюю информацию, и поэтому им настоятельно рекомендуется обращаться к патентной базе данных БСЭ по адресу: http://www.itu.int/ITU-T/ipr/.

СОДЕРЖАНИЕ

Стр.

Введение

Стремительное развитие информационно-коммуникационных технологий и услуг, основанных на интернете, обусловливает генерирование, передачу и хранение больших объемов данных, которые растут быстрыми темпами. Данные генерируются не только датчиками, камерами и сетевыми устройствами, но и веб-страницами, системами электронной почты, социальными сетями и многими другими источниками. Наборы данных становятся настолько большими и сложными и поступают столь быстро, что традиционные методы и инструменты их обработки уже недостаточны. Эффективный анализ данных с допустимой задержкой становится чрезвычайно сложной задачей. Для решения вышеуказанных проблем разрабатывается концептуальная схема, называемая анализом больших данных.

Организации электросвязи осуществляют сбор данных о физических лицах, включая информацию, позволяющую установить личность, управление ими, их использование и обмен такими данными. Для того чтобы защитить эти персональные данные, они используют методы деидентификации. Отношения между сторонами, участвующими в процессе обмена данными, определяют, необходимо ли удалять идентификационную информацию перед сбором данных, после их сбора, но до их сохранения или только непосредственно перед передачей следующей стороне. Соответственно поставщики услуг электросвязи должны своевременно, эффективно и безопасно предоставлять потребителям данных услуги деидентификации данных.

Рекомендация МСЭ-Т X.1148

Структура процесса деидентификации для поставщиков услуг электросвязи

1 Сфера применения

В настоящей Рекомендации дается обзор процесса деидентификации на основе модели жизненного цикла данных и описывается структура этого процесса с указанием рабочих этапов и ролей заинтересованных сторон. Кроме того, в Приложениях и Дополнениях к ней рассматриваются модели публикации данных и этапы данных в процессе деидентификации, а также различные подходы к деидентификации и примеры.

Вопросы, связанные с регулированием, в настоящей Рекомендации не рассматриваются.

2 Справочные документы

Отсутствуют.

3 Термины и определения

3.1 Термины, определенные в других документах

В настоящей Рекомендации используются следующие термины, определенные в других документах:

3.1.1 агрегированные данные (aggregated data) [b-ISO/IEC 20889]: Данные, относящиеся к группе субъектов данных, такие как набор статистических свойств этой группы.

3.1.2 обезличивание (anonymization) [b-ISO/IEC 29100]: Процесс, посредством которого информация, позволяющая установить личность (PII), необратимо изменяется таким образом, что субъект PII уже не может быть прямо или косвенно идентифицирован диспетчером PII, действующим как в одиночку, так и в сотрудничестве с любой другой стороной.

3.1.3 атрибут (attribute) [b-ISO/IEC 20889]: Индивидуальная характеристика.

3.1.4 набор данных (dataset) [b-ISO/IEC 20889]: Собрание данных.

3.1.5 деидентификация; удаление идентификационной информации (de-identification) [b-ISO 25237]: Общий термин для любого процесса удаления ассоциации между набором идентифицирующих данных и субъектом данных (см. пункт 3.2.4).

3.1.6 процесс деидентификации (de-identification process) [b-ISO/IEC 20889]: Процесс удаления ассоциации между набором идентифицирующих атрибутов и субъектом данных.

3.1.7 метод деидентификации (de-identification technique) [b-ISO/IEC 20889]: Метод преобразования набора данных в целях уменьшения степени возможной ассоциации информации с отдельными субъектами данных.

3.1.8 деидентифицированный набор данных (de-identified dataset) [b-ISO/IEC 20889]: Набор данных, полученный в результате применения процесса деидентификации.

3.1.9 деидентифицированная информация (de-identified information) [b-NISTIR 8053]: Запись, в которой удалено или скрыто достаточно информации PII, так что оставшаяся информация не позволяет идентифицировать человека и нет разумных оснований полагать, что эта информация может быть использована для идентификации человека.

3.1.10 дифференциальная конфиденциальность (differential privacy) [b-ISO/IEC 20889]: Формальная модель измерения конфиденциальности, гарантирующая, что вероятностное распределение результатов статистического анализа различается не более чем на заданную величину, независимо от того, представлен ли во входном наборе данных какой-либо конкретный субъект данных.

ПРИМЕЧАНИЕ. – В частности, дифференциальная конфиденциальность обеспечивает:

a) математическое определение конфиденциальности, которое гласит, что для того чтобы результаты любого статистического анализа считались сохраняющими конфиденциальность, результаты анализа исходного набора данных должны быть неотличимы от результатов, полученных после добавления в набор данных или удаления из него какого-либо субъекта данных; и

b) меру конфиденциальности, позволяющую отслеживать совокупную потерю конфиденциальности и устанавливать верхнюю границу (или "бюджет") такой потери. Формальное определение выглядит следующим образом. Пусть ε – положительное действительное число, а M – рандомизированный алгоритм, принимающий входной набор данных. Алгоритм M называется ε-дифференциально конфиденциальным, если для всех наборов данных D1 и D2, различающихся одним элементом (т. е. данными одного субъекта данных), и всех поднаборов S диапазона M mml_m1, где вероятность относится к степени случайности, используемой алгоритмом.

3.1.11 идентификатор (identifier) [b-ISO/IEC 20889]: Набор атрибутов в наборе данных, обеспечивающий возможность уникальной идентификации субъекта данных в определенном рабочем контексте.

ПРИМЕЧАНИЕ. – Соотношение этого определения с определениями, приведенными в других стандартах, рассматривается в Приложении B.

3.1.12 идентифицирующий атрибут (identifying attribute) [b-ISO/IEC 20889]: Атрибут в наборе данных, способствующий уникальной идентификации субъекта данных в определенном рабочем контексте.

3.1.13 лицо, заинтересованное в обеспечении конфиденциальности (privacy stakeholder) [b-ISO/IEC 29100]: Физическое или юридическое лицо, орган государственной власти, агентство или любая другая организация, которые могут повлиять на решение или действие, связанное с обработкой информации, позволяющей установить личность (PII), либо быть затронуты или чувствовать себя затронутыми таким решением или действием.

3.1.14 псевдонимизация (pseudonymization) [b-ISO/IEC 20889]: Метод деидентификации, заменяющий идентификатор (или идентификаторы) субъекта данных псевдонимом в целях сокрытия идентичности этого субъекта данных.

3.1.15 квазиидентификатор (quasi-identifier) [b-ISO/IEC 20889]: Атрибут в наборе данных, который в сочетании с другими атрибутами того же набора выделяет субъекта данных.

3.1.16 запись (record) [b-ISO/IEC 20889]: Набор атрибутов, относящихся к одному и тому же субъекту данных.

3.1.17 реидентификация; **восстановление идентификационной информации (re-identification)** [b-ISO/IEC 20889]: Процесс восстановления ассоциации данных в деидентифицированном наборе данных с исходным субъектом данных.

ПРИМЕЧАНИЕ. – Это определение охватывает процесс, устанавливающий присутствие конкретного субъекта данных в наборе данных.

3.1.18 выделение (single out) [b-ISO/IEC 20889]: Вычленение записи, относящейся к субъекту данных, из набора данных путем выявления набора известных характеристик, уникально идентифицирующих этого субъекта данных.

3.1.19 третья сторона (third party) [b-ISO/IEC 29100]: Любая сторона, заинтересованная в обеспечении конфиденциальности, кроме субъекта информации, позволяющей установить личность (PII), диспетчера PII и обработчика PII, а также физических лиц, уполномоченных обрабатывать данные под непосредственным руководством диспетчера PII или обработчика PII.

3.1.20 доверенная третья сторона (trusted third party) [b-ISO/IEC 18014-1]: Орган обеспечения безопасности или его агент, который пользуется доверием остальных участников процесса в отношении действий, связанных с безопасностью.

3.1.21 k-анонимность (k-anonymity) [b-ISO/IEC 20889]: Формальная модель измерения конфиденциальности, гарантирующая, что для каждого идентификатора из набора данных существует соответствующий класс эквивалентности, содержащий не менее K записей.

3.1.22 l-разнообразие (l-diversity) [b-ISO/IEC 20889]: Формальная модель измерения конфиденциальности, гарантирующая, что для выбранного атрибута каждый класс эквивалентности имеет по крайней мере *L* хорошо представленных значений.

ПРИМЕЧАНИЕ. – L-разнообразие – это свойство набора данных, которое дает гарантированную нижнюю границу *L* разнообразия значений, общих для класса эквивалентности выбранного атрибута.

3.1.23 t-плотность (t-closeness) [b-ISO/IEC 20889]: Формальная модель измерения конфиденциальности, гарантирующая, что расстояние между распределением выбранного атрибута в некотором классе эквивалентности и распределением этого атрибута во всей таблице не превышает порогового значения *T*.

ПРИМЕЧАНИЕ. – Считается, что таблица имеет T-плотность по отношению к выбранному атрибуту, если T-плотность имеют все классы эквивалентности, содержащие этот атрибут.

3.2 Термины, определенные в настоящей Рекомендации

В настоящей Рекомендации определяются следующие термины:

3.2.1 диспетчер данных (data controller): Заинтересованная сторона (или сторона, заинтересованная в обеспечении конфиденциальности), определяющая цели и средства обработки данных, за исключением физических лиц, использующих данные в личных целях.

3.2.2 обработчик данных (data processor): Заинтересованная сторона, обрабатывающая данные от имени диспетчера данных и в соответствии с его инструкциями.

3.2.3 сотрудник по защите данных (data protection officer): Лицо, назначенное диспетчером PII для обеспечения независимым образом соблюдения требований законодательства/регламента конфиденциальности.

ПРИМЕЧАНИЕ. – "Диспетчер PII" (PII controller) и "диспетчер данных" (data controller) – синонимы.

3.2.4 субъект данных (data subject): Лицо, к которому относятся данные.

ПРИМЕЧАНИЕ. – Английский термин "data subject" (субъект данных) является синонимом английских терминов "PII principal" (субъект PII) и "data principal" (субъект данных).

3.2.5 обработка (process): В отношении информации или данных означает получение, запись или хранение информации или данных или выполнение любой операции или набора операций с информацией или данными, включая:

– организацию, адаптацию или изменение информации или данных;

– извлечение информации или данных, обращение к ним или их использование;

– раскрытие информации или данных путем передачи, публикации или предоставления иным способом; или

– согласование, объединение, блокирование, удаление или уничтожение информации или данных.

4 Сокращения и акронимы

В настоящей Рекомендации используются следующие сокращения и акронимы.

DP	Differential Privacy	Дифференциальная конфиденциальность
DPO	Data Protection Officer	Сотрудник по защите данных
PII	Personally Identifiable Information	Информация, позволяющая установить личность
TTP	Trusted Third Party	Доверенная третья сторона

5 Соглашения

Отсутствуют.

6 Обзор процесса деидентификации

Целью процесса деидентификации является обеспечение защиты конфиденциальности данных субъектов. Поскольку до и после анализа данных в целях извлечения значимой информации эти данные могут содержать информацию, позволяющую установить личность (PII), аналитик должен учитывать аспекты безопасности.

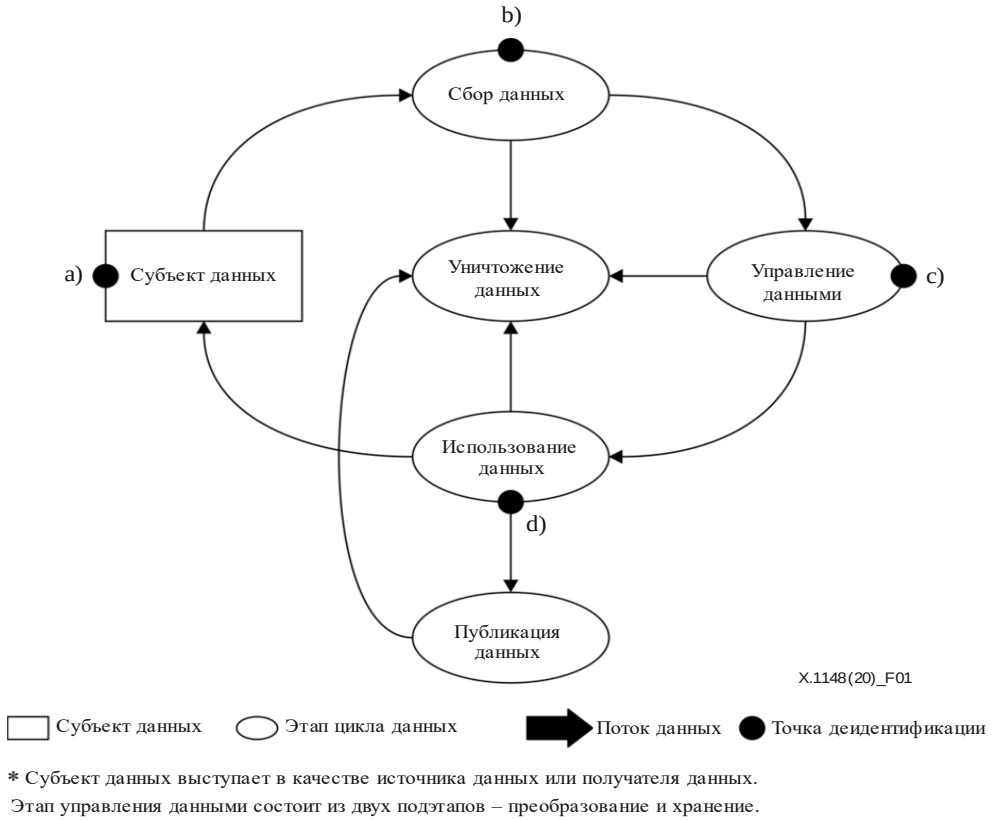
В этом разделе определяются условия анализа данных, модель жизненного цикла данных, роли участников процесса деидентификации и другие аспекты деидентификации.

6.1 Модель жизненного цикла данных и этап деидентификации

Как правило, организация устанавливает задачи деидентификации в целях обеспечения конфиденциальности и безопасности. В этом пункте определяется жизненный цикл данных и указывается, когда следует рассматривать процесс деидентификации на основе этой модели жизненного цикла данных.

Концепция жизненного цикла данных используется для выбора подходящих мер и средств контроля на основе анализа возможности реидентификации. В настоящей Рекомендации жизненный цикл данных определяется согласно описанию в пунктах 6.1.1–6.1.5.

Обзор процесса деидентификации в модели жизненного цикла данных представлен на рисунке 1.



* Субъект данных выступает в качестве источника данных или получателя данных.
Этап управления данными состоит из двух подэтапов – преобразование и хранение.

Рисунок 1 − Процесс деидентификации в модели жизненного цикла данных

6.1.1 Этап сбора данных

Данные собираются от субъектов данных, то есть лиц, к которым относятся данные. Набор данных, созданный в результате этого процесса сбора данных, может включать PII. В результате деидентификации создается новый набор данных, из которого была удалена вся PII. Рекомендуется, чтобы внутри организации вместо исходного набора данных, по возможности, использовались деидентифицированные наборы данных.

Используя эту модель, деидентификацию можно выполнить:

– либо во время сбора данных, то есть в точке b) на рисунке 1;
– либо когда данные собраны, но идентификатор фактически не требуется, то есть в точке a) на рисунке 1.

Идентификаторы, которые не нужны для управления данными (преобразование и хранение данных), вводить не следует.

6.1.2 Этап управления данными

Во избежание архивирования идентификаторов деидентификацию следует применять после преобразования данных и до начала их хранения, то есть в точке c) на рисунке 1. Организациям рекомендуется учитывать возможность реидентификации и установить четкие меры и средства контроля доступа, максимальные сроки хранения и правила удаления данных, максимально ограничивающие возможность установления взаимосвязей (ассоциируемость) между деидентифицированными данными. Организациям рекомендуется рассмотреть методы обезличивания, такие как агрегирование данных, если предполагаемая цель использования это позволяет.

6.1.3 Этап использования данных

Если в организации для управления данными необходима PII, рекомендуется деидентифицировать эти данные перед их публикацией в виде набора данных для обмена данными, то есть в точке d) на рисунке 1.

6.1.4 Этап публикации данных

Данные могут передаваться третьим сторонам, связанным дополнительными административными мерами контроля, такими как соглашение об обмене данными. Также возможна публикация деидентифицированных наборов данных. Публикация деидентифицированных данных осуществляется в соответствии с тремя моделями – открытой, полуоткрытой и закрытой. Необходимая степень деидентификации может варьироваться в зависимости от выбранной модели публикации.

6.1.5 Этап уничтожения данных

Уничтожение данных может быть выполнено на любом этапе, то есть на этапе сбора данных, управления данными, использования данных или публикации данных. Данные должны уничтожаться проверенными методами, исключающими возможность их восстановления. В частности, следует рассмотреть необходимость уничтожения данных при обнаружении возможности реидентификации.

6.2 Аспекты деидентификации

Применение деидентификации на всем протяжении жизненного цикла данных повышает ее эффективность. Однако характер отношений между сторонами, участвующими в потоке данных, влияет на выбор момента деидентификации: до сбора данных, то есть в точке a) на рисунке 1; после их сбора, то есть в точке b) на рисунке 1, но до начала их хранения, то есть в точке c) на рисунке 1; или только перед передачей данных следующей стороне в потоке данных, то есть в точке d) на рисунке 1. Это решение, в свою очередь, влияет на осуществимость мер безопасности и других организационных мер, направленных на повышение эффективности конкретного метода деидентификации в каждом случае использования. Хотя деидентификация может быть полезным методом защиты конфиденциальности данных субъектов в тех случаях, когда цель использования не допускает применения методов обезличивания, ее самой по себе недостаточно для защиты данных субъектов, и она должна рассматриваться как часть комплексной системы защиты данных. В этом разделе описываются особенности и аспекты каждого этапа.

6.2.1 Сбор данных

Наиболее распространенным подходом является локальная деидентификация (или деидентификация в источнике), которая позволяет отдельному лицу (или диспетчеру данных этого лица) удалить всю PII перед публикацией данных для анализа.

Одним из аспектов деидентификации, непосредственно связанным с этапом сбора данных, является минимизация данных. Каждый диспетчер данных, собирающий данные субъектов, должен точно

определить, какие данные строго необходимы для предполагаемой цели использования, и ограничить сбор данных только этими определенными параметрами.

Следует предусмотреть специальные процессы, исключающие ненужную PII из процесса сбора или передачи данных, чтобы уменьшить поля данных.

Другим аспектом деидентификации является агрегирование данных. Диспетчеры данных должны рассматривать возможность агрегирования данных во всех случаях, когда цель использования не предполагает строгого требования выделения отдельных субъектов данных.

6.2.2 Управление данными

6.2.2.1 Преобразование данных

Этап преобразования данных может включать применение методов деидентификации, таких как агрегирование данных, статистическое ограничение раскрытия данных, шифрование и т. д. Преобразование данных может применяться на одном или нескольких этапах, в том числе непосредственно после сбора данных и перед длительным хранением, после периода продолжительного хранения и перед доступом, или его можно объединить с процессом доступа.

Общее преобразование в виде компоновки или агрегирования данных можно использовать в любое время после сбора данных и вплоть до их публикации. Если компоновка или агрегирование применяются сразу после сбора данных, они могут уменьшить потенциальный вред для субъектов данных в случае их утечки; однако это также ограничивает возможность связывания, слияния или обновления данных после компоновки.

Метод преобразования данных следует выбирать после тщательного рассмотрения потенциального вреда для субъектов данных в случае их утечки. Принимая решение по преобразованию, следует также учитывать последующий анализ, который впоследствии должен выполняться в соответствии с целью использования данных, поскольку методы снижения рисков раскрытия могут влиять на возможность последующего использования и анализа.

6.2.2.2 Хранение данных

Хранение данных определяется как процесс сохранения данных, включая PII, в любой форме на энергонезависимом носителе диспетчером данных или стороной, действующей под руководством диспетчера. Меры и средства управления информационной безопасностью и конфиденциальностью уже сфокусированы на этапе хранения данных, поэтому в настоящем пункте они перечислены без подробного рассмотрения [b-ISO/IEC 27001]. На этапе хранения используется ряд общих мер и средств управления информационной безопасностью и конфиденциальностью, таких как управление доступом, техническое сопровождение, оценка безопасности, процедуры аутентификации, контроль за инцидентами и реагирование на них, а также аудиторские проверки.

В частности, организации должны соблюдать максимальные сроки хранения и правила удаления данных, чтобы гарантировать, что данные хранятся не дольше, чем это строго необходимо для достижения цели их использования, и что по истечении этого максимального срока хранения данные полностью уничтожаются. Например, в соглашениях об обмене данными часто указывается, что получатель должен уничтожить данные в течение указанного периода времени, например через год после получения, и такое положение договора может требоваться по закону.

6.2.3 Использование данных

Деидентифицированные данные могут собираться, храниться или распространяться для ряда целей и применений, зависящих от определенных свойств данных, сохранившихся после деидентификации. Одной из основных причин публикации деидентифицированных наборов данных является предоставление другим лицам возможности изучения значений и свойств исходных данных в исследовательских целях [b-ISO/IEC 20889]. Поэтому при деидентификации следует также стремиться сохранить как можно больше полезной информации, защищая при этом конфиденциальность физических лиц. Эта двоякая цель деидентификации делает такой подход важным при рассмотрении вопроса об использовании данных в ряде контекстов, включая модели публикации данных.

При публикации деидентифицированных данных организация, как правило, на уровне экспертного комитета, в состав которого входит широкий круг заинтересованных сторон, должна принять решение

о рассмотрении потенциальных последствий такой публикации для затрагиваемых ею субъектов данных. Для такой оценки и определения подходящего механизма публикации, который уменьшит риск реидентификации, часто используются оценки риска и контрольные списки.

Выбор методов деидентификации зависит от степени их применимости или "полезности" в конкретном случае использования.

7 Структура процесса деидентификации

В данном разделе описывается структура процесса деидентификации для получения деидентифицированной PII в четыре этапа, как показано на рисунке 2 [b-KOREA].

Этап 1 − Предварительный обзор

Этап 1 заключается в проверке принадлежности целевых данных к категории PII. Если данные содержат PII, то переходят к этапу 2. Необходима деидентификация.

Этап 2 − Деидентификация

Этап 2 предусматривает применение деидентификации данных для предотвращения возможности извлечения из целевого набора данных конкретной индивидуальной информации. На этом этапе применяются методы полного или частичного удаления или преобразования элементов PII. К элементам PII относятся идентификаторы, квазиидентификаторы и конфиденциальные атрибуты.

Этап 3 − Оценка адекватности

Этап 3 заключается в оценке адекватности деидентифицированного набора данных, включая элементы PII. К числу рассматриваемых аспектов относятся вопрос о содержании в целевом наборе данных остаточной PII, возможность прямой реидентификации и возможность установления взаимосвязей, которые могут привести к реидентификации.

Этап 4 − Последующее управление

Этап 4 предусматривает измерение организационной и технической безопасности для предотвращения реидентификации.

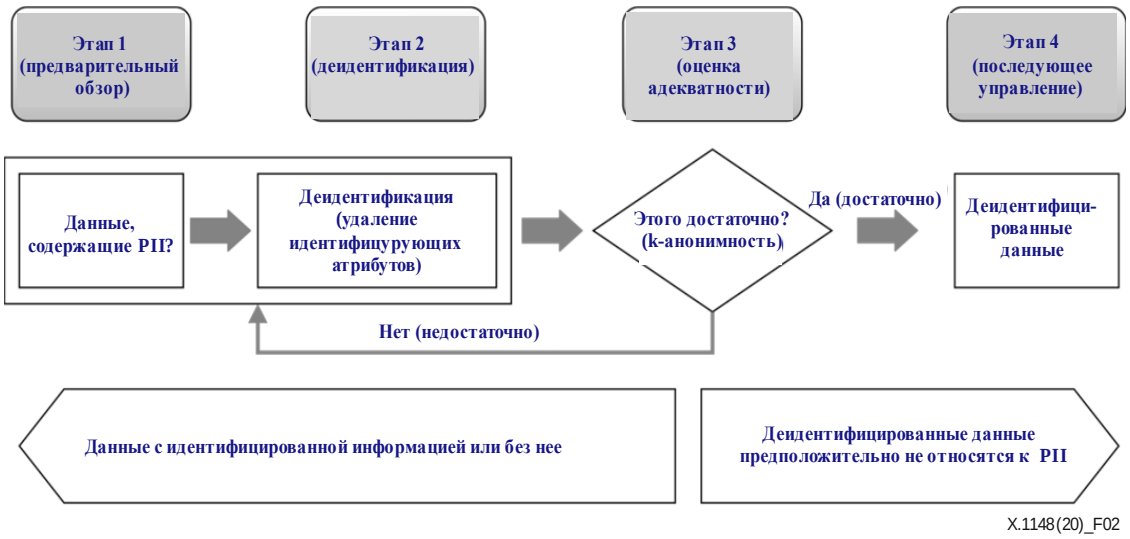


Рисунок 2 − Процесс деидентификации

Описание каждого из этих этапов приведено в пунктах 7.1–7.4.

7.1 Этап 1 − Предварительный обзор

Организациям, которые намереваются использовать или предоставлять данные для различных целей, сначала следует определить свою политику и стандарты. Рекомендуется, чтобы политика и стандарты отвечали на следующие вопросы.

– Какова цель и предполагаемое назначение деидентифицированной информации?
– Из каких атрибутов данных состоят деидентифицированные данные?
– Какие методы используются для деидентификации?
– Каковы уровни риска реидентификации и ее неблагоприятные последствия?
– Какие имеются решения на случай реидентификации конкретного человека?
– Как оценивается уровень реидентификации?
– Как определяется численность необходимого персонала и стоимость деидентификации?

Конкретные аспекты, составляющие предварительный анализ, могут различаться в зависимости от типа данных и предполагаемой цели их использования. Тем не менее рекомендуется установить набор стандартов.

Организации, намеревающиеся обрабатывать данные для нескольких целей, должны проверить по соответствующим стандартам, относятся ли конкретные данные к PII или нет. Даже если данные не определены как PII, организация должна рассмотреть любой риск установления взаимосвязей между имеющимися данными и принять соответствующие меры для минимизации этого риска. Если данные относятся к PII, то необходим этап деидентификации.

Примеры критериев принятия решений о принадлежности к PII:

– отсутствуют какие бы то ни было специальные ограничения на тип, форму, характеристики и формат данных;
– если диспетчер данных может с помощью данных установить личность человека, то такие данные считаются относящимися к PII;
– данные должны относиться к отдельному человеку. Статистическая ценность группы, состоящей из нескольких человек, не относится к PII;
– данные, по которым можно установить личность человека, скомбинировав их с дополнительной информацией, относятся к PII. Под дополнительной информацией обычно понимается общедоступная/легко доступная информация.

7.2 Этап 2 − Применение деидентификации

7.2.1 Удаление идентификаторов

Идентификатор – это информация, такая как значение или имя, однозначно присвоенные человеку или предмету, связанному с человеком. В общем случае набор идентификаторов должен быть сведен к минимуму, и любые идентификаторы, включенные в наборы данных, должны быть удалены.

Однако следующие виды данных могут относиться к идентификаторам, абсолютно необходимым для предполагаемой цели:

– уникальный идентификатор (гражданский регистрационный номер, номер социального страхования (SSN), номер паспорта, идентификационный номер иностранца, номер водительского удостоверения и т. д.);
– имя (китайские иероглифы, имя на английском языке и т. д.);
– подробный адрес (номер дома, улица и т. д.);
– дата (дата рождения, годовщина (свадьбы и т. д.), дата выдачи свидетельства и т. д.);
– номер телефона (мобильного, домашнего, рабочего, факса и т. д.);
– номер медицинской карты, номер государственного медицинского страхования, номер получателя социальной помощи и т. д.;
– номер банковского счета, номер кредитной карты и т. д.;

– фотоматериалы (фотографии, видеозаписи, записи системы видеонаблюдения (CCTV) и т. д.);
– биометрические данные (отпечатки пальцев, голос, рисунок радужной оболочки глаза и т. д.);
– адрес электронной почты, IP-адрес, адрес управления доступом к среде передачи (MAC), унифицированный указатель ресурса (URL) домашней страницы и т. д.;
– идентификационный код (номер сотрудника, номер клиента и т. д.);
– другой уникальный идентификационный номер (номер военнослужащего, регистрационный номер предприятия и т. д.).

7.2.2 Удаление квазиидентификаторов и атрибутов с высоким потенциалом идентифицируемости

Квазиидентификаторы, включенные в наборы данных, как правило, удаляются, если они не имеют отношения к цели использования данных. Если квазиидентификатор, относящийся к использованию данных, имеет идентифицируемые элементы, должны применяться методы деидентификации, такие как псевдонимизация и агрегирование.

К данным с высоким потенциалом идентифицируемости, таким как поведенческая информация, должны применяться методы деидентификации и, по возможности, методы обезличивания.

7.2.3 Методы деидентификации

Может использоваться целый ряд методов, включая псевдонимизацию, агрегирование, скрытие и маскирование данных, по отдельности или в сочетании друг с другом. Применения одного только метода псевдонимизации может быть недостаточно для деидентификации.

Каждый из методов может быть реализован с помощью различных легкодоступных приемов. Следует выбрать наиболее подходящий метод и использовать его исходя из назначения данных, а также из достоинств и недостатков каждого конкретного метода. После завершения деидентификации можно переходить к следующему этапу.

7.3 Этап 3 − Оценка адекватности процесса деидентификации

При недостаточной деидентификации личность человека можно установить, объединив данные с другими или используя различные методы логического вывода.

Чтобы снизить риск реидентификации, перед использованием данных необходимо провести оценку адекватности деидентифицированных данных. Это предполагает, в частности, ответы на следующие вопросы.

– Какова цель данного запроса на деидентификацию?
– Какие типы атрибутов данных участвуют в деидентификации (включены ли идентификаторы)?
– Каков надлежащий уровень деидентификации?

Эту оценку адекватности может выполнить сотрудник по защите данных (DPO), уполномоченная доверенная третья сторона (TTP) или внешняя группа по оценке.

В числе других моделей защиты конфиденциальности при оценке адекватности используется модель k-анонимности. Эта модель служит базовым средством оценки. При необходимости могут применяться дополнительные модели оценки (l-разнообразие, t-плотность, дифференцированная конфиденциальность (DP) и т. д.).

Дополнительная информация по оценке адекватности приведена в Приложении A.

7.4 Этап 4 − Последующее управление

7.4.1 Меры защиты деидентифицированных данных

Меры защиты применяются для предотвращения возможности реидентификации деидентифицированных данных в случае их утечки и/или их объединения с другими данными. К ним относятся, в частности, следующие меры:

– организационные меры защиты – назначение должностного лица, ответственного за файлы деидентифицированных данных, решение о передаче деидентифицированных данных и уничтожение данных по достижении цели их использования;
– технические меры защиты – ограничение доступа к файлам деидентифицированных данных, управление регистрацией доступа, а также установка и эксплуатация программ безопасности.

Кроме того, в число мер безопасности входят защитные меры, принимаемые в случае утечки деидентифицированных данных. К ним относятся, в частности, следующие меры:

– анализ причин утечки и принятие как организационных, так и технических мер безопасности для предотвращения дополнительной утечки;
– изъятие и уничтожение похищенных деидентифицированных данных.

7.4.2 Контроль возможностей реидентификации

Диспетчер данных, намеревающийся использовать деидентифицированные данные или предать их третьей стороне, должен проводить регулярный контроль возможностей реидентификации.

При обнаружении возможности реидентификации необходимо обратиться к диспетчеру данных, предоставившему деидентифицированные данные, с запросом на приостановку обработки, изъятие и уничтожение этих данных.

7.4.3 Требования к договору с третьей стороной

При предоставлении или передаче деидентифицированных данных третьей стороне для их использования в соответствующем договоре должно быть предусмотрено управление риском реидентификации. К управлению риском реидентификации относятся:

– уведомление субъектов данных о раскрытии данных третьим лицам;
– предоставление третьим лицам по возможности обезличенных данных;
– запрет на реидентификацию – диспетчеру данных, которому поручена обработка деидентифицированных данных, должно быть запрещено реидентифицировать эти данные путем их объединения с другими данными;
– ограничение передачи или перепоручения – при передаче деидентифицированных данных или поручении их обработки в договоре должен быть указан разрешенный объем передачи или перепоручения;
– уведомление о рисках реидентификации – должно быть принято обязательство прекратить обработку данных в случае реидентификации или повышения ее возможности и проинформировать о проблеме отправителя и получателя данных.

7.4.4 Контрмеры против реидентификации

В случае реидентификации деидентифицированных данных обработку данных следует прекратить и принять необходимые меры для предотвращения утечки PII.

Реидентифицированные данные подлежат немедленному уничтожению.

8 Полезность деидентифицированных данных

8.1 Этапы деидентификации данных

В этом разделе определяются этапы деидентификации данных, которые могут быть представлены в виде типов данных для описания того, в какой мере личность прямо идентифицируется данными и в какой мере она ассоциируется с характеристиками (атрибутами), содержащимися в данных. Спецификация данных в контексте использования или обработки данных должна включать в себя не только тип данных, но и описание того, в какой мере данные могут идентифицировать человека или ассоциировать его личность с содержащимся в этих данных набором характеристик.

На рисунке 3 представлены этапы перехода от данных, содержащих идентификационную информацию, к деидентифицированным данным в процессе деидентификации. На каждом этапе имеется разная возможность реидентификации, что определяет спектр риска. Тип данных

характеризует конкретные стадии, которые должен пройти набор данных, предоставляя все меньше и меньше возможностей для реидентификации.

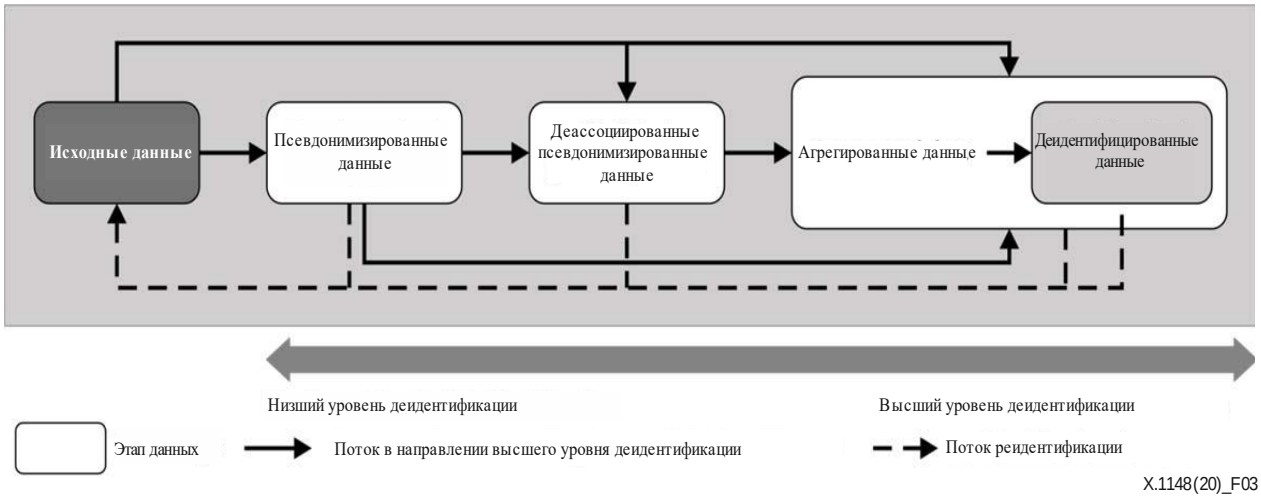


Рисунок 3 − Этапы деидентификации данных

Как показано на рисунке 3, на этапе деидентификации присутствуют все данные. Справа (высший уровень деидентификации) находятся деидентифицированные данные, не относящиеся к отдельным лицам (например, данные о погоде за прошлые периоды) и поэтому не несущие в себе риска нарушения конфиденциальности. На левом конце (низший уровень деидентификации) находятся данные, содержащие идентификационную информацию, напрямую ассоциируемые с конкретными людьми. Между этими двумя этапами данных находятся данные, которые можно связать с отдельными людьми, затратив некоторые усилия, данные, которые можно связать только с группами людей, и данные, которые относятся к отдельным людям, но не допускают обратного прослеживания до них. В общем случае процессы деидентификации направлены на продвижение данных вправо при сохранении некоторой желаемой степени полезности, снижении риска распространения деидентифицированных данных среди более широкой аудитории или их превращения в общедоступные данные.

8.1.1 Этап исходных данных

На этапе исходных данных, содержащих идентифицированную информацию, данные могут быть однозначно связаны с конкретным человеком, поскольку по этой информации прослеживается личность. Руководство по вопросу о том, что может считаться идентификаторами, содержится в пункте 4.4.1 [b-ISO/IEC 29100].

8.1.2 Этап псевдонимизированых данных

На этапе псевдонимизированных данных никто, кроме стороны, присвоившей псевдоним, не может разумными усилиями восстановить данные, поскольку все идентификаторы заменены псевдонимами. Однако идентификационная информация все еще может быть восстановлена из псевдонимизированных данных через возможность установления взаимосвязей с другими данными.

Это соответствует данным, прошедшим процесс, которые определен в пункте 3.1.14 как "псевдонимизация".

8.1.3 Этап деассоциированных псевдонимизированных данных

На этапе деассоциированных псевдонимизированных данных все идентификаторы стираются или заменяются псевдонимами, функция присвоения которых стирается или является необратимой, так что никто не может восстановить ассоциацию никакими разумными усилиями, включая сторону, производившую эту обработку данных. Однако идентификационная информация все еще может быть восстановлена из деассоциированных псевдонимизированных данных через возможность установления взаимосвязей с другими данными.

8.1.4 Этап агрегированных данных

На этапе агрегированных данных данные представляют собой информацию о достаточно разных физических лицах, в которой атрибуты индивидуального уровня являются невосстанавливаемыми обобщенными статистическими данными, не содержащими записей индивидуального уровня. При использовании методов агрегирования все агрегированные данные не достигают степени идентифицируемости ниже порогового уровня, если размер ячейки для данного пересечения некоторой комбинации переменных может привести к тому, что кто-то идентифицирует конкретного человека.

Это соответствует данным, определенным в пункте 3.1.1 как "агрегированные данные".

8.1.5 Этап деидентифицированных данных

На этапе деидентифицированных данных данные не ассоциированы и атрибуты изменены (например, значения атрибутов рандомизированы или обобщены) таким образом, что существует разумный уровень уверенности в том, что человек не может быть прямо или косвенно идентифицирован с помощью только этих данных или при их сочетании с другими данными.

8.2 Модели публикации данных

Модели публикации деидентифицированных данных классифицируются по трем категориям в зависимости от аналитического контекста данных [b-UKAN].

Существуют три модели публикации деидентифицированных данных – открытая, полуоткрытая и закрытая.

Каждая модель публикации предусматривает разные уровни доступности и защиты информации. В зависимости от цели и/или законодательных требований, предъявляемых к публикации данных, пригодность каждой модели может варьироваться. Модель публикации данных играет важную роль в процессе деидентификации, поскольку необходимая степень деидентификации может варьироваться в зависимости от выбранной модели публикации.

Каждая из трех моделей публикации рассматривается в пунктах 8.2.1–8.2.3.

8.2.1 Модель открытой публикации данных

При традиционной открытой публикации данных любой человек может получить доступ к данным без регистрации или каких-либо условий. Примерами такой публикации могут служить общедоступные данные организаций и данные, размещенные в хранилище данных открытого доступа, например на веб-портале. Организации активно публикуют наборы данных и делают их свободно доступными каждому для использования и тиражирования.

При открытой публикации данных на информацию обычно налагается как можно меньше ограничений, включая ограничения, касающиеся того, кто и как может получить к ней доступ. Таким образом, когда лица, загружающие набор данных, не могут быть идентифицированы, такое раскрытие информации следует рассматривать как открытую публикацию данных.

Однако в случае доступа к информации по запросам, как указано пункте 8.2.2, публикацию следует считать открытой в тех случаях, когда от лица, запрашивающего информацию, не требуется согласие с условиями или положениями, касающимися обработки, конфиденциальности или безопасности информации.

8.2.2 Модель полуоткрытой публикации данных

Модель полуоткрытой публикации данных является более ограничительной, чем модель открытой публикации, и имеет место тогда, когда присутствует процесс формального запроса и разрешения на получение доступа к данным. В этом случае получатель данных может согласиться с некоторыми условиями использования или подписать контракт нажатием кнопки. Такие контракты представляют собой условия использования в онлайновом режиме, которые могут налагать ограничения на то, что можно делать с данными и как с ними обращаться. Несмотря на это, такие данные может загрузить любой.

Деидентификация также может быть полезна при ответе на запросы доступа к наборам данных. Используя деидентификацию, организации могут отвечать на такие запросы безопасным для

конфиденциальности образом, сохраняя при этом полезность информации. Для соблюдения некоторых ограничений при публикации данных через информационную систему организации могут использовать средства управления доступом, например:

– требовать от всех пользователей регистрации и предоставления контактной информации перед получением доступа к данным;

– использовать протоколы аутентификации для проверки личности человека;

– использовать многоуровневые системы доступа для предоставления разных уровней доступа разным лицам, например в зависимости от принадлежности к той или иной организации или от полномочий.

В таких информационных системах научному сообществу может предлагаться система интерактивных запросов, и исходные данные будут предоставляться лишь небольшому количеству аналитиков, прошедших тщательный отбор.

Кроме того, возможен случай доступа к данным, не требующий публикации, когда диспетчер данных сам может выполнять анализ по поручению аналитиков. В этом случае публикация данных организацией может отсутствовать.

8.2.3 Модель закрытой публикации данных

Обмен наборами данных, содержащими PII, может иметь место внутри организаций и между ними только в том случае, если раскрытие таких данных разрешено нормативными актами страны. Если же раскрытие не разрешено, а учреждения желают обмениваться наборами данных, то любую PII необходимо удалить. Закрытая публикация данных обеспечивает наименьшую доступность, но более высокий уровень защиты, требуя меньшей работы по деидентификации.

При обмене информацией между организациями, поскольку доступ к набору данных ограничен организацией, требования в отношении конфиденциальности и безопасности информации могут быть установлены и их выполнение обеспечено посредством соглашения об обмене данными. Стороны должны заключить между собой такое соглашение, чтобы публикация данных считалась закрытой. Соглашение об обмене данными – важная часть стратегии снижения рисков при такой публикации; в нем используются некоторые общепринятые условия, например:

– перечень лиц, которым разрешен доступ (контроль получателей);

– требования к безопасности данных (контроль инфраструктуры);

– ограничения на использование, в частности запрет на ссылки на другие файлы и запрет на преднамеренную реидентификацию (контроль других видов данных и управления);

– требование уничтожать данные по завершении использования (контроль управления).

Соглашение об обмене данными преследует троякую цель:

– устанавливает четкое различие между теми лицами или организациями, которым диспетчер данных доверяет, и теми, кому он не доверяет;

– будучи рамочным соглашением, устанавливает условия, при соблюдении которых может происходить доступ;

– может определять санкции или штрафы за нарушение этих условий доступа физическим лицом/организацией.

8.2.4 Сравнение моделей публикации данных

В условиях потока данных одним из способов ограничения возможности реидентификации является установление контроля за способами получения и использования данных. Меры и средства такого контроля можно классифицировать в соответствии с различными моделями публикации данных, у каждой из которых имеются свои преимущества и свои риски. Организации также могут выбрать подход многоуровневого доступа, при котором сочетаются несколько таких моделей для различных случаев использования и угроз конфиденциальности. Кроме того, модели публикации должны учитывать возможность нескольких или периодических публикаций. Здесь перечислены несколько моделей публикации данных – от неограничивающих до моделей с жесткими ограничениями. Их сравнение приведено в таблице 2.

Таблица 2 − Сравнение моделей публикации данных

	Модель открытой публикации данных	**Модель полуоткрытой публикации данных**	**Модель закрытой публикации данных**
Права доступа	• Каждый имеет свободный доступ к опубликованным данным	• Доступ к опубликованным данным (или их части) предоставляется ограниченному кругу лиц или организаций	• Доступ к опубликованным данным имеет узкий круг лиц или организаций
Сценарии использования	• Неограниченный доступ к данным через веб-портал, то есть свободный доступ для всех	• Обеспечение безопасности на своей территории • Предоставляемый доступ • Дистанционный виртуальный доступ • Доступ через аналитический сервер	• Обмен внутри и между организациями
Права	• Неограниченные права на многократное использование и распространение данных	• Имеются у уполномоченного лица или организации	• Повторное использование, тиражирование и распространение данных запрещены
Попытка реидентификации	• Демонстрационная атака для рекламы	• Умышленная внутренняя атака • Непреднамеренное опознавание конкретного лица в наборе данных знакомым • Утечка данных	

8.3 Связь между моделью публикации данных и этапом деидентификации данных

8.3.1 Модель закрытой публикации данных

При передаче данных из источника данных в модель закрытой публикации требуется деидентификация данных. В обычных обстоятельствах несмотря на модель закрытой публикации используются деассоциированные псевдонимизированные данные и данные с высоким уровнем деидентификации. В этом случае могут использоваться такие средства деидентификации, как псевдонимизация, шифрование, синтез, скрытие и т. д.

Однако если существует специальный договор между двумя сторонами или закон, то для анализа и хранения данных на этом этапе могут использоваться псевдонимизированные данные.

8.3.2 Модель полуоткрытой публикации данных

При передаче данных из источника данных в модель полуоткрытой публикации требуется деидентификация данных более высокого уровня, чем для модели закрытой публикации. Для предотвращения реидентификации выполняется статистическая обработка данных. После этого агрегированные данные и данные с более высоким уровнем деидентификации могут быть переданы в модель полуоткрытой публикации. Конкретнее, могут использоваться такие средства деидентификации, как статистическая обработка, рандомизация и т. д.

Как показано в таблице 2, может быть разрешена деидентификация относительно более низкого уровня, чем в модели открытой публикации, поскольку доступ к данным может получить лишь ограниченный круг лиц или организаций.

8.3.3 Модель открытой публикации данных

При передаче данных из источника данных в модель открытой публикации требуется деидентификация данных более высокого уровня, чем в модели полуоткрытой публикации. Выполняется процесс деидентификации данных, результаты которого могут использоваться для модели открытой публикации, как показано в таблице 2.

Приложение A

Процедуры оценки адекватности

(Данное Приложение является неотъемлемой частью настоящей Рекомендации.)

В данном Приложении представлена модель осуществления процедуры оценки адекватности [b-KOREA]. См. рисунок A.1.

Ниже приведено описание процедуры оценки адекватности по шагам.

– Подготовка базовых документов. Диспетчер данных подготавливает базовые документы, необходимые для оценки адекватности, такие как описание данных, статус деидентификации и уровень управления организациями-пользователями. Организация-пользователь – это организация, которая намеревается использовать данные после деидентификации.

– Формирование группы по оценке. Сотрудник по вопросам конфиденциальности может сформировать группу по оценке или обратиться к DPO или TTP для проведения оценки.

– Проведение оценки. Группа по оценке оценивает адекватность уровня деидентификации, используя базовые документы, подготовленные ответственным за PII.

– Осуществление дополнительной деидентификации. Если по результатам оценки деидентификация признана неадекватной, диспетчер данных осуществляет дополнительный процесс деидентификации, отражающий мнения участников оценки.

– Использование данных. Если по результатам оценки деидентификация признана адекватной, данные могут использоваться или предоставляться в таких целях, как анализ больших данных.

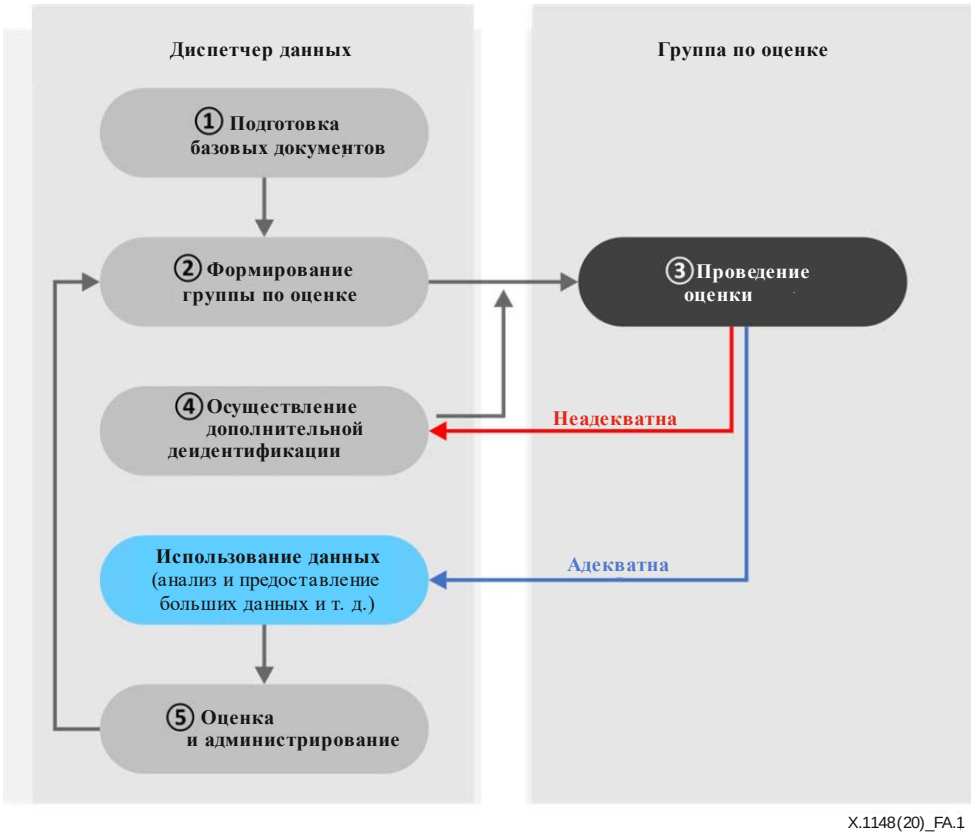


Рисунок A.1 − Процедура оценки адекватности деидентификации

A.1 Подготовка базовых документов

Диспетчер данных подготавливает базовые документы, необходимые для оценки адекватности, такие как описание предмета оценки, статус деидентификации и уровень управления организацией-пользователем.

A.2 Формирование группы по оценке

Сотрудник по вопросам конфиденциальности может сформировать группу по оценке. При привлечении внешних специалистов назначаются несколько экспертов по правовым вопросам и вопросам деидентификации из круга экспертов, управляемого специализированными агентствами в каждой области.

Группа по оценке состоит из членов, не имеющих прямой заинтересованности в целях использования данных.

A.3 Проведение оценки

Группа по оценке оценивает адекватность деидентификации на основе базовых документов и с использованием модели k-анонимности:

– предварительный обзор – обзор подготовленных диспетчером данных базовых документов и проверки путем собеседования наличия в наборе данных элементов, идентифицирующих личность, и соответствия методов деидентификации цели использования данных;

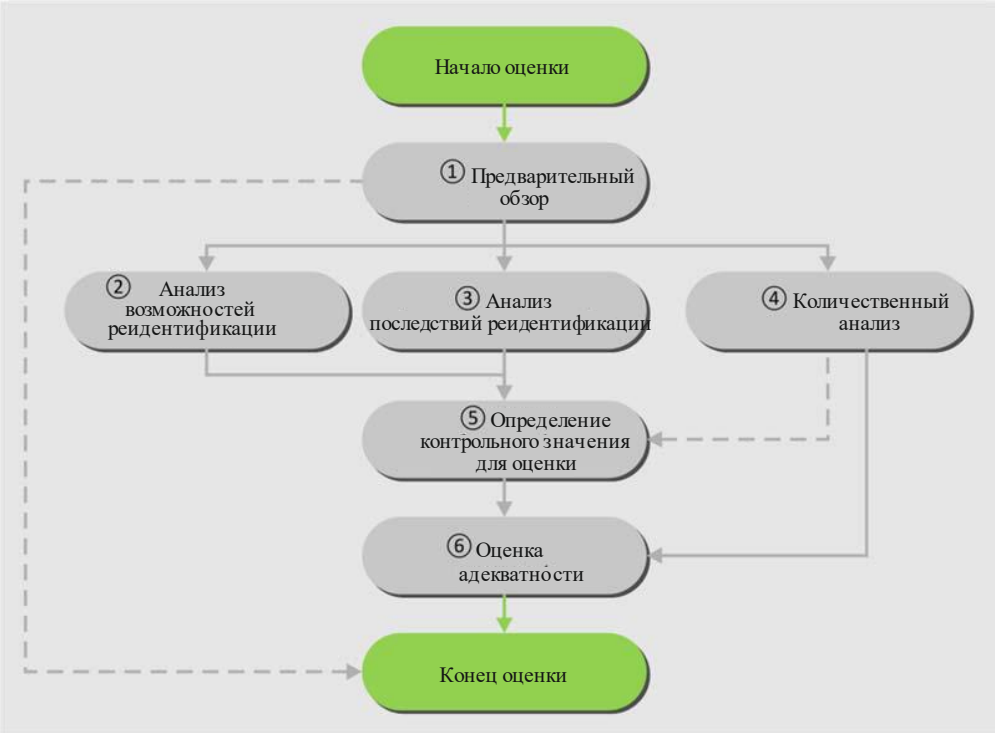
– анализ возможностей попыток реидентификации – анализ возможностей попыток реидентификации, включая намерение, уровень защиты PII и возможности диспетчера данных, использующего или получающего данные;

– анализ последствий реидентификации – оценка возможных последствий для субъекта данных преднамеренной или непреднамеренной реидентификации;

– количественный анализ – проверка точности значения *K*, представленного диспетчером данных;

– определение контрольного значения для оценки – группа по оценке тщательно определяет контрольное значение для оценки, принимая во внимание возможности реидентификации, ее последствия, результаты количественного анализа и назначение данных;

– оценка адекватности – определение адекватности деидентификации путем сравнения расчетных значений, полученных на основе среднего контрольного значения и количественного анализа.



X.1148(20)_FA.2

Рисунок A.2 – Процедура оценки адекватности

A.4 Дополнительные меры по деидентификации

– Если по результатам оценки деидентификация признана неадекватной, то диспетчер данных реализует дополнительные меры по деидентификации на основе мнения группы по оценке.

– По завершении дополнительной деидентификации диспетчером данных группа по оценке приступает к повторной оценке.

A.5 Использование данных

– Если деидентификация оценена (возможно, после повторной оценки) как адекватная, то деидентифицированные данные можно использовать для анализа больших данных или разрешить их передачу третьей стороне.

– В принципе, предоставление или раскрытие данных для открытого доступа или для пользователей данных, с которыми не существует договора, в отсутствие надлежащей стратегии снижения риска для моделей публикации данных запрещено ввиду высокого риска реидентификации.

– Данные уничтожаются, как только цель использования данных достигнута или они больше не нужны.

– В процессе использования данных для их эффективного применения в форме деидентифицированных данных должны соблюдаться последующие этапы управления.

Приложение B

Подходы к деидентификации неструктурированных данных

(Данное Приложение является неотъемлемой частью настоящей Рекомендации.)

В отличие от деидентификации структурированных данных механизмы деидентификации неструктурированных данных применяются не к полям структурированных данных, а к исходным данным. В случае фотографий, показанных на рисунке B.1, ниже, деидентификация означает удаление лиц или замену их другими.

Рисунок B.1 − Пример деидентификации лиц людей на фотографиях

Существуют неструктурированные данные четырех типов:

1) неструктурированные текстовые данные – веб-данные, отчетные документы, блоги, новости и т. д.;
2) неструктурированные видеоданные – все видеоданные не структурированы, и некоторая информация меток содержит упорядоченные данные;
3) неструктурированные аудиоданные – все аудиоданные не структурированы, и некоторая информация меток или результаты распознавания голоса преобразуются в текстовые данные;
4) неструктурированные данные журналов событий – сгенерированные машиной данные журналов событий не структурированы, но обычно имеют некоторую структуру и могут быть *преобразованы в структурированную форму.*

Для представления синтаксической информации неструктурированных данных, включающих текст, голос, изображение и видео, система деидентификации должна содержать следующие три блока:

1) блок обнаружения мультимедийной информации для выявления текстовых метаданных во входных мультимедийных данных, содержащий:
   - детектор речи, который преобразует входной голосовой сигнал в текст для отслеживания объектов или действий в этом голосовом сигнале;
   - оптический детектор распознавания символов, который извлекает символы из входного изображения;
   - визуальный детектор, который извлекает объекты или действия из входного изображения или удаляет изображения из входного фото- или видеоизображения;
   - визуальный детектор предложений, который извлекает текстовые предложения из входного фото- или видеоизображения;
2) формирователь, подключенный к базе знаний, который разделяет текстовые метаданные и контекстную информацию на синтаксическую информацию, отражающую внешнюю конфигурацию, и семантическую, отражающую внутреннюю информацию:

– синтаксическая информация включает информацию об источнике мультимедийных данных, информацию о мультимедийных данных, сгенерированных этим источником, и информацию об обнаруженных объектах, извлеченных из значимой области;

– семантическая информация включает информацию о событиях, происходящих в значимой области, соединяя мультимедийные данные с контекстной информацией;

3) блок деидентификации удаляет идентифицируемую PII из базы знаний и текстовых метаданных.

Для механизма деидентификации неструктурированных данных следует определить соответствующие требования и уровень безопасности следующим образом.

– Цель деидентификации: определить целевой объект, подлежащий защите в приложении или онлайновых услугах.

– Способ выполнения деидентификации: определить механизм, который следует применять для деидентификации. Каков должен быть уровень деидентификации (черный прямоугольник, пикселизация, размытие)?

– Деидентификация и реидентификация: следует определить необходимость восстановления или реидентификации. Возможно ли будет восстановить оригинальную фотографию, если она потребуется полиции для расследования преступления?

Дополнение I

Примеры типичных методов деидентификации

(Данное Дополнение не является неотъемлемой частью настоящей Рекомендации.)

В данном Дополнении приведены некоторые примеры и описания типовых методов деидентификации.

I.1 Статистические инструменты деидентификации

– Выборка – процесс, при котором вместо всего набора данных публикуется выборка из этого набора. Если опубликована подвыборка, то вероятность реидентификации может уменьшиться.

– Агрегирование – набор статистических функций, приводящих к значению, характеризующему весь набор данных.

I.2 Криптографические инструменты деидентификации

– Детерминированное шифрование [b-ISO/IEC 11770] – схема шифрования, которая при отдельных прогонах алгоритма шифрования для данного открытого текста и ключа всегда создает один и тот же зашифрованный текст.

– Шифрование с сохранением порядка [b-AGRAWAL] – схема шифрования, при которой сохраняется последовательность фрагментов открытого текста.

– Гомоморфное шифрование [b-ISO/IEC 18033-6] – схема шифрования, которая позволяет выполнять вычисления в зашифрованном тексте, генерируя таким образом зашифрованный результат, который после расшифровки совпадает с результатом операций, выполненных над открытым текстом.

– Шифрование с сохранением формата [b-NIST 800-38G] – схема шифрования, при которой зашифрованный текст имеет тот же формат, что и открытый текст.

– Гомоморфный обмен секретными ключами [b-ISO/IEC 18033-6] – алгоритм обмена секретными ключами, при котором секретный ключ шифруется с использованием гомоморфного шифрования.

I.3 Методы скрытия

– Маскировка – процесс замены поля значением или его удаления. Примером метода скрытия может служить замена номера телефона звездочками или случайно сгенерированным псевдонимом.

– Локальное скрытие – процесс, который скрывает или удаляет из выбранных записей определенные значения атрибутов. Удаление данных повышает степень защиты конфиденциальности, но может понизить полезность набора данных.

– Скрытие записей – процесс, заключающийся в удалении из набора данных целой записи или записей.

I.4 Методы псевдонимизации

Это процесс, который удаляет ассоциацию с субъектом данных и добавляет ассоциацию между конкретным набором характеристик, относящихся к субъекту данных, и одним или несколькими псевдонимами. Как правило, псевдонимизация осуществляется путем замены прямых идентификаторов псевдонимом, таким как случайно сгенерированное значение. К примерам прямых идентификаторов относятся имена, адреса электронной почты и государственные номера. Псевдонимом заменяются все прямые идентификаторы и, возможно, дополнительные или все остальные идентифицирующие атрибуты.

I.5 Методы обобщения

– Округление – процесс замены числового значения другим значением, приблизительно равным ему, но с более коротким, простым или наглядным представлением.

– Кодирование сверху и снизу – процесс, для которого атрибут со значениями выше верхней границы (или ниже нижней границы) устанавливается в качестве наибольшего (или наименьшего) возможного порогового значения.

I.6 Методы рандомизации

– Добавление шума – процесс, при котором к выбранному атрибуту набора данных добавляется случайное значение, которое невозможно предсказать.

– Перестановка – процесс перестановки значений выбранного атрибута между записями в наборе данных без их изменения.

– Микроагрегация – процесс, при котором все значения непрерывных атрибутов заменяются их средними значениями, рассчитанными определенным образом.

I.7 Синтетические данные

Синтетические данные – это подход, при котором искусственно генерируются микроданные, соответствующие заранее определенной статистической модели данных. По определению, синтетические наборы данных не содержат данных, собранных от существующих субъектов данных, но выглядят реальными с точки зрения их предполагаемого назначения.

Дополнение II

Подходы к процессу деидентификации

(Данное Дополнение не является неотъемлемой частью настоящей Рекомендации.)

В данном Дополнении приведены некоторые примеры и подробные описания подходов к процессу деидентификации.

II.1 Информационно-ориентированный подход к деидентификации

Учитывая, что во избежание раскрытия PII методы деидентификации изменяют исходные данные, возникает явное противоречие между их полезностью и конфиденциальностью. Задача состоит в том, чтобы защитить конфиденциальность с минимальной потерей точности; в идеале пользователи должны проводить свой анализ на деидентифицированных данных без потери точности результатов этого анализа по сравнению с работой с исходными данными.

На практике трудно добиться идеальной деидентификации без ущерба для полезности набора данных. В области больших данных эта проблема усугубляется из-за количества и разнообразия данных. С одной стороны, низкоуровневой деидентификации (например, когда деидентификация заключается только в скрытии прямых идентификаторов) обычно недостаточно для обеспечения неидентифицируемости. С другой стороны, слишком строгая деидентификация может помешать связыванию данных об одном и том же человеке (или о похожих людях), поступающих из разных источников, и таким образом уменьшить многие потенциальные преимущества больших данных.

В этом разделе описываются два подхода к информационно-ориентированной деидентификации для преодоления противоречия между полезностью и конфиденциальностью. Чтобы измерить полезность опубликованного набора деидентифицированных данных, можно использовать как общие меры полезности, так и меры, зависящие от назначения данных.

II.1.1 Подход к деидентификации с предпочтением полезности

В области больших данных информация о человеке часто собирается из нескольких независимых источников. Поэтому при создании больших данных главной задачей является возможность связывать между собой записи, относящиеся к одному и тому же человеку (или к одной и той же/подобной категории людей).

При подходе к деидентификации с предпочтением полезности к микронабору данных применяется метод деидентификации с эвристическим выбором параметров и с подходящими свойствами сохранения полезности, после чего измеряется риск раскрытия. В результате подход к деидентификации с предпочтением полезности работает медленно и не дает формальных гарантий конфиденциальности. Например, риск реидентификации можно оценить эмпирически, пытаясь установить взаимосвязь между записями исходного и деидентифицированного наборов данных. Если риск считается слишком высоким, метод деидентификации следует повторить с более строгими параметрами конфиденциальности и, возможно, с большей потерей полезности, итеративно изменяя параметры, пока эмпирически определенный риск раскрытия не станет достаточно низким, таким как в официальной статистике.

Безусловно, хотя с точки зрения полезности ассоциируемость желательна, она представляет угрозу конфиденциальности: в деидентифицированных наборах данных точность ассоциаций должна быть значительно ниже, чем в исходных наборах. Ассоциируемость, совместимая с методом деидентификации или с моделью конфиденциальности деидентификации, определяет, может ли аналитик связать между собой независимо деидентифицированные данные (в соответствии с этим методом/моделью), относящиеся к одному и тому же человеку.

II.1.2 Подход к деидентификации с предпочтением конфиденциальности

Модель конфиденциальности усиливается параметром, обеспечивающим верхнюю границу риска реидентификации и, возможно, риска раскрытия атрибутов. Усиление модели достигается с помощью метода деидентификации, зависящего от модели, с параметрами, полученными из параметров модели.

В число хорошо известных моделей конфиденциальности входят k-анонимность и ее расширения, а также ε-дифференциальная конфиденциальность, которая часто приводит к низкой полезности/ ассоциируемости данных.

При подходе к деидентификации с предпочтением конфиденциальности если полезность результирующих деидентифицированных данных слишком мала, то используемую модель конфиденциальности следует либо усилить с помощью альтернативного метода деидентификации, наносящего меньший ущерб полезности, либо выбрать менее строгий параметр конфиденциальности или даже прибегнуть к другой модели конфиденциальности при деидентификации.

II.2 Ролевой подход к деидентификации

В этом разделе описаны три типа подходов, выполняющие роли и обязанности друг друга в процессе деидентификации. Ролевой подход можно в целом охарактеризовать как отвечающий на вопросы "кто", "какой" и "где и каким образом":

– Кто имеет доступ к данным?

– Какой анализ может или не может проводиться?

– Где должен осуществляться доступ к данным и их анализ и как получить такой доступ?

II.2.1 Централизованная деидентификация

В процессе контроля статистического раскрытия данных основной упор делается на централизованной деидентификации, выполняемой диспетчером данных, имеющим доступ ко всему исходному набору данных. Этому централизованному подходу присущи некоторые преимущества и недостатки, как показано в таблице II.1.

Таблица II.1 − Характеристики централизованной деидентификации

	Описание
Преимущества	• Физическим лицам не нужно деидентифицировать предоставляемые ими записи данных. Можно ожидать, что диспетчер данных, у которого больше вычислительных ресурсов и, возможно, больше опыта в области деидентификации, адекватно выполнит деидентификацию всего набора данных. • Диспетчер данных имеет полное представление об исходном наборе данных и, следовательно, находится в лучшем положении для оптимизации компромисса между полезностью данных и сохранившимся риском раскрытия.
Недостатки	• Все стороны, предоставляющие исходные данные, должны доверять диспетчеру данных (поскольку он имеет доступ ко всем исходным данным). Это не проблема в официальной статистике, где в роли диспетчера данных выступает государственный институт статистики, но в типичном сценарии анализа больших данных может оказаться серьезным препятствием, например когда диспетчер данных, собирающий данные из нескольких источников, является просто частной компанией (брокер данных). • Деидентификация может представлять собой слишком тяжелую вычислительную нагрузку для одного диспетчера данных, особенно в случае больших данных. • В одном сценарии обработки больших данных участвуют много диспетчеров, что делает централизованный подход неуправляемым.

Подходы локальной деидентификации и коллективной деидентификации дополняют вышеперечисленные преимущества и недостатки.

II.2.2 Локальная деидентификация

Локальная деидентификация – это альтернативный подход ограничения раскрытия персональной информации, подходящий для сценариев (включая работу с большими данными), когда отдельные люди (субъекты данных) не доверяют (или доверяют лишь частично) диспетчеру данных. Каждый субъект деидентифицирует свои данные, прежде чем передать их диспетчеру данных.

С учетом защиты конфиденциальности данные, собранные определенным источником, должны быть деидентифицированы в источнике перед предоставлением. Однако независимая деидентификация каждым источником приводит к большей потере информации, чем при централизованной деидентификации, поскольку субъекты деидентифицируют свои данные, не видя данных других субъектов. Другими словами, у субъектов нет общего представления о наборе данных, что затрудняет им поиск хорошего компромисса между достигнутым ограничением риска раскрытия и потерей информации.

II.2.3 Коллективная деидентификация

Процесс коллективной деидентификации сочетает в себе низкую потерю полезности, характерную для централизованной деидентификации, с высоким уровнем конфиденциальности, достигаемым при локальной деидентификации. Проблема централизованной деидентификации заключается в том, что если субъект данных не верит, что диспетчер данных правильно использует и/или деидентифицирует его данные, он может предоставить ложные данные (вызывая таким образом систематическую ошибку в ответах) или вообще их не предоставить (вызывая ошибку пропущенных данных). Поэтому субъекты могут сотрудничать друг с другом, чтобы определить риск раскрытия их данных, а затем локально применить нужный уровень защиты распределенным и коллективным способом, который отличается двумя основными свойствами:

– потери информации не больше, чем при получении набора данных в рамках централизованного подхода, при том же уровне конфиденциальности. Этот метод превосходит локальный подход тем, что он гарантирует меньшие потери информации;

– как субъекты, так и диспетчер данных получают не больше сведений о конфиденциальных атрибутах любого другого конкретного субъекта данных, чем сведения, содержащиеся в окончательном деидентифицированном наборе данных. Этот метод превосходит централизованный подход, предлагая также конфиденциальность по отношению к сборщику данных.

Кроме того, коллективный подход может привести к более гладкой работе протоколов без каких-либо внешних принудительных механизмов. При деидентификации микроданных защита конфиденциальности, получаемая одним субъектом, влияет на защиту конфиденциальности, получаемую другими. Для повышения общей полезности при коллективном подходе требуется безопасное многостороннее преобразование электронных протоколов, позволяющее двум или более участникам выполнять преобразование наборов данных каждого из них таким образом, чтобы ни одной из сторон не пришлось явно передавать набор данных другой стороне. Поскольку безопасное многостороннее преобразование позволяет преобразовывать запросы без необходимости централизации хранения всех данных, это уменьшает вред от утечки данных и обеспечивает возможность вычислений между сторонами, которые не полностью доверяют друг другу. В определенных контекстах многосторонние вычисления могут улучшить как конфиденциальность, так и полезность данных.

Библиография

[b-ISO/IEC 11770] ISO/IEC 11770 (all parts), *Information technology – Security techniques – Key management*.

[b-ISO/IEC 18033-6] ISO/IEC 18033-6, *Information technology security techniques – Encryption algorithms – Part 6: Homomorphic encryption.*

[b-ISO/IEC 20889] ISO/IEC 20889 (2018), *Privacy enhancing data de-identification terminology and classification of techniques.*

[b-ISO/IEC 27001] ISO/IEC 27001 (2018), *Information technology – Security technique – Information security management systems.*

[b-ISO/IEC 29100] ISO/IEC 29100 (2011), *Information technology – Security technique – Privacy framework.*

[b-NIST 800-38G] NIST Special Publication 800-38G (2016), *Recommendation for Block Cipher Modes of Operation: Methods for Format-Preserving Encryption.*

[b-NISTIR 8053] NISTIR 8053 (2015), *De-Identification of Personal Information.*

[b-AGRAWAL] Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y. (2004), *Order preserving encryption for numeric data, SIGMOD '04 Proceedings of the 2004 ACM SIGMOD international conference on Management of data, Paris, France, June, pp. 563-574.*

[b-KOREA] Korean Ministry of the Interior, *Guidelines on De-identification Measures, June 2016.*
<http://www.privacy.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000821178&fileSn=2&nttId=7187&toolVer=&toolCntKey_>
(Дата последнего обращения: 26 июля 2019 г.)
<https://www.privacy.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000827161&fileSn=0>
(На английском языке, дата последнего обращения: 12 декабря 2020 г.)

[b-UKAN] UK Anonymization Network, *The anonymisation decision-making framework, 2016*
<https://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf>

СЕРИИ РЕКОМЕНДАЦИЙ МСЭ-Т

Серия A	Организация работы МСЭ-Т
Серия D	Принципы тарификации и учета и экономические и стратегические вопросы международной электросвязи/ИКТ
Серия E	Общая эксплуатация сети, телефонная служба, функционирование служб и человеческие факторы
Серия F	Нетелефонные службы электросвязи
Серия G	Системы и среда передачи, цифровые системы и сети
Серия H	Аудиовизуальные и мультимедийные системы
Серия I	Цифровая сеть с интеграцией служб
Серия J	Кабельные сети и передача сигналов телевизионных и звуковых программ и других мультимедийных сигналов
Серия K	Защита от помех
Серия L	Окружающая среда и ИКТ, изменение климата, электронные отходы, энергоэффективность; конструкция, прокладка и защита кабелей и других элементов линейно-кабельных сооружений
Серия M	Управление электросвязью, включая СУЭ и техническое обслуживание сетей
Серия N	Техническое обслуживание: международные каналы передачи звуковых и телевизионных программ
Серия O	Требования к измерительной аппаратуре
Серия P	Качество телефонной передачи, телефонные установки, сети местных линий
Серия Q	Коммутация и сигнализация, а также соответствующие измерения и испытания
Серия R	Телеграфная передача
Серия S	Оконечное оборудование для телеграфных служб
Серия T	Оконечное оборудование для телематических служб
Серия U	Телеграфная коммутация
Серия V	Передача данных по телефонной сети
Серия X	**Сети передачи данных, взаимосвязь открытых систем и безопасность**
Серия Y	Глобальная информационная инфраструктура, аспекты протокола Интернет, сети последующих поколений, интернет вещей и "умные" города
Серия Z	Языки и общие аспекты программного обеспечения для систем электросвязи

Отпечатано в Швейцарии
Женева, 2021 г.