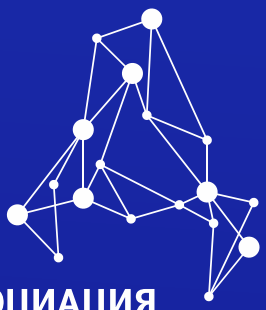


АССОЦИАЦИЯ
БОЛЬШИХ ДАННЫХ

ТЕХНОЛОГИИ ЗАЩИЩЕННОЙ ОБРАБОТКИ ДАННЫХ. ОБЕЗЛИЧИВАНИЕ

Москва
2025

СОДЕРЖАНИЕ

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Повсеместное использование сквозной аналитики больших объемов данных и широкое применение систем на базе искусственного интеллекта, повлекшее необходимость массового обучения ИИ-моделей, привели к увеличению потребностей во все больших объемах статистически и семантически качественных данных.

Совместная работа организаций по обмену данными, передача обработки данных во внешние экспертные системы на основе ИИ, привлечение внештатных специалистов и подрядчиков — все это усиливает риски утечек и повышает требования к защите данных и степени их приватности.

Существенно увеличить аналитическую ценность данных удается благодаря возможности многократного транзитного использования наборов данных несколькими их операторами, связанными единым сценарием. Например, анализ пересечения данных рекламной площадки и набора данных банковских транзакций позволяет провести объективную оценку эффективности рекламных кампаний. Приведенный пример демонстрирует вполне обоснованное применение конфиденциальных вычислений с учетом контекста использования и с сохранением ожидаемого уровня конверсии аналитических выводов на основе данных из двух источников. Также можно привести множество примеров совместной работы с данными внутри корпоративного контура.

Как при использовании наборов данных несколькими операторами, так и в сценариях внутрикорпоративной обработки данных доступ к ним необходимо строго регламентировать, например, с применением технологий защищенной обработки данных (ТЗОД). Выбор наиболее подходящей технологии играет при этом ключевую роль: он влияет как на уровень риска компрометации данных в конкретном сценарии, так и на решение более приземленных задач бизнеса, таких, например, как ускорение сдачи нового функционала в продуктивную эксплуатацию, достигаемое благодаря упрощению процессов.

Наиболее актуальные и популярные на сегодня ТЗОД можно условно разделить на три группы:

1. Методы обфускации (скрытия) данных.
Использование искаженных данных считается относительно безопасным — их обработка может происходить в открытом виде, если применяются следующие методы:

· обезличивание;
· генерация синтетических данных;
· обеспечение дифференцированной приватности;
· доказательства с нулевым разглашением (Zero-Knowledge Proofs, ZKP).

2. Обработка шифрованных данных.
Эти методы строятся на том, что данные шифруются или преобразуются с использованием математических методов или аппаратной защиты. В их числе:

· гомоморфное шифрование (Homomorphic Encryption, HE);
· защищенные многосторонние вычисления (Secure Multi-Party Computation, SMPC);
· конфиденциальные вычисления, или (другое название этого термина) доверенные среды исполнения (Trusted Execution Environments, TEE).

3. Федеративное обучение.
Его идея в том, что обработка данных происходит на базе источников, при этом партнеры по обработке обмениваются не самими данными, а обученными на их основе моделями.

ТЕХНОЛОГИИ ЗАЩИЩЕННОЙ ОБРАБОТКИ ДАННЫХ. ОБЕЗЛИЧИВАНИЕ

Цель нашего аналитического доклада — помочь сделать осознанный выбор методик защищенной обработки данных и, в частности, сформировать правильное понимание применимости технологии обезличивания. Доклад охватывает обзор основных подходов, примеры применения методов обезличивания, анализ рисков, их юридическую интерпретацию, а также рекомендации для тех, кто начинает или пересматривает процессы обезличивания данных.

Важно отметить, что область обращения персональных данных регламентируется динамически, при этом поправки и новые рекомендации Роскомнадзора не всегда могут быть восприняты операторами данных однозначно. На практике в ходе валидации выбранного метода и выстраивании процесса обезличивания важно не только учесть требования регулятора, но и наложить их на внутренние процессы организации. Метод обезличивания следует выбирать таким образом, чтобы не терялся смысл данных и сохранялась полезность информации в ходе аналитики и обучения моделей.

Правильно выстроенный процесс обезличивания позволяет повысить качество тестирования, упростить работу с внешними подрядчиками, ускорить вывод новых разработок на рынок, снизить сопутствующие расходы, избежать репутационных издержек и существенно уменьшить вероятность штрафов.

ПОДХОДЫ К ОБЕЗЛИЧИВАНИЮ

Классификация атрибутов

Согласно тексту Федерального закона N 152-ФЗ «О персональных данных» от 27.07.2006, под понятие персональных данных подпадает любая информация, прямо или косвенно относящаяся к определенному или определяемому физическому лицу (субъекту персональных данных) [1].

В зависимости от степени полноты, сочетания и контекста такую информацию могут составлять различные наборы атрибутов, которые могут быть классифицированы по уровням идентификации.

Рассмотрим, как разные виды персональных данных (ПДн) позволяют идентифицировать личность, какие риски это влечет и какие требования к их обработке из этого следуют.

Прямые идентификаторы

Прямые идентификаторы — это атрибуты, позволяющие однозначно и напрямую установить личность субъекта данных без использования дополнительной информации.

Примеры:
- ФИО (в определённом контексте);
- номер телефона;
- адрес электронной почты;
- номер паспорта;
- СНИЛС, ИНН;
- номер банковской карты;
- уникальный ID клиента, пациента или сотрудника;
- порядковый номер записи в базе (если он неизменяемый).

Риски, связанные с прямыми идентификаторами: прямая идентификация, злоупотребление, компрометация личности.

Требования: обязательное удаление или криптографическое преобразование атрибутов (маскирование, шифрование, токенизация).

Косвенные идентификаторы

Косвенные идентификаторы — это атрибуты, которые по отдельности не идентифицируют субъекта, но могут использоваться для восстановления личности в сочетании с внешними данными, статистикой или информацией из дополнительных источников.

Примеры:
- IP-адрес (особенно статический)
- тип устройства, User-Agent
- точные координаты или адрес местоположения
- уникальные параметры браузера
- геометки и пути движения
- данные о времени посещения, активности
- идентификаторы cookies или трекеров
- уникальные комбинации редко встречающихся атрибутов
- длинные описания в составе текстовых полей (например, описания опыта работы, информация о товаре и т. п.).

Риски, связанные с косвенными идентификаторами: атакующий может сопоставить поведение или данные с информацией, полученной в результате утечек.

Требования: применение обобщения, округления, маскирования и механизмов дифференциальной приватности (DP).

Квазиидентификаторы

Квазиидентификаторы представляют собой атрибуты, которые в совместной комбинации могут однозначно идентифицировать субъект, даже если каждый атрибут по отдельности выглядит как безвредный. Классический пример: сочетание даты рождения, данных пола и почтового индекса — в 87% случаев эти три значения однозначно идентифицируют человека.

Примеры:
- возраст или дата рождения
- пол
- регион, город, почтовый индекс
- профессия
- образование
- семейное положение
- количество детей
- тип жилья
- доход (категориальный)

Риски в разрезе квазиидентификаторов: редко встречающиеся комбинации используются для выделения записей и совмещения с внешними данными.

Примечание : квазиидентификаторы представ-

ляют собой центральную точку обезличивания с использованием обобщения, подавления, группировки, перемешивания, применения моделей k-анонимности, l-разнообразия, t-близости.

Чувствительные атрибуты

К **чувствительным** относят атрибуты, раскрытие которых способно причинить существенный ущерб субъекту ПДн, включая дискриминацию, финансовый вред или утрату репутации.

Ключевое свойство: даже если личность не установлена, сам факт раскрытия значения опасен.

Примеры:
- диагнозы, результаты анализов, истории болезней
- этническая принадлежность
- политические взгляды
- религиозные убеждения
- сексуальная ориентация
- заработная плата
- состояние кредита, просрочки
- судимости
- геолокация в реальном времени
- биометрические данные

Риски, связанные с чувствительными атрибутами: вывод (inference), дискриминация, утечка конфиденциальной информации, восстановление профиля.

Требования: для чувствительных атрибутов необходимо применять методы l-разнообразия, t-близости, β -сходства, DP-ограничений.

Нечувствительные атрибуты

Нечувствительными называют атрибуты, которые не раскрывают личность сами по себе и не наносят вред даже в раскрытом виде. Примерами таких атрибутов могут служить тип устройства, год обслуживания, категория товара или параметры транзакций без указания связи с конкретным лицом. Степень безопасности при этом зависит от контекста: в одном наборе эти данные могут быть безопасны, а в другом способны превратиться в квазиидентификаторы.

Примеры:
- технические параметры устройства
- общие категории операций
- агрегированные статистические показатели
- неуникальные признаки транзакций
- демографические атрибуты с низкой комбинационной мощностью

Риски, связанные с нечувствительными атрибутами: образование уникальных комбинаций.

Примечание: необходим контроль взаимосвязей.

Класс эквивалентности (equivalence class)

Классом эквивалентности называется группа записей, которые совпадают по набору квазиидентификаторов после применения методов обобщения или подавления.

Формальное определение класса эквивалентности:

$$EC = \{x \in D: QI(x) = v\}$$

где QI — набор квази-идентификаторов, v — их значения.

Классы эквивалентности EC представляют собой основу всех синтаксических моделей:

Модель «Требование к классу эквивалентности EC»

k-анонимность > $(k - 1)$ уникальных значений

l-разнообразие ≥ l уникальных чувствительных атрибутов в EC

t-близость $D\,(PEC\,, Pglobal\,) \leq t$

β-сходство $PEC(x) \leq (1 + \beta)\, Pglobal\,(x)$

Таким образом, для каждого типа данных или атрибута в зависимости от решаемой задачи и контекста использования данных может быть выбрана своя актуальная методика из списка рекомендуемых.

Характер данных также может влиять на выбор методики. Например, для обезличивания набора фиксированных свойств, характеризующих физическое лицо в какой-то момент времени, и хронологической последовательности его действий или истории изменения свойств этого субъекта (таких, в частности, как транзакции в банке) могут применяться разные подходы и сценарии.

Сами данные, подлежащие обезличиванию, могут относиться к различным форматам: тексты, фото, сканы, аудио, видеопотоки, плоские

таблицы в базах данных и пр. Подходы к обработке при этом остаются единообразными.

При выборе метода обезличивания важно учитывать в конкретных бизнес-сценариях наложение рекомендуемых видов модификаций рассматриваемого типа атрибутов на результаты подавления части информации.

Для оценки степени сохранения смысловых характеристик данных следует опираться как на базовые форматно-логические проверки, выполняемые модулями интеграции и контроля качества, так и на анализ соответствия данных требованиям конкретных аналитических задач и задач обучения моделей.

Выбор методики и сценария обезличивания конкретного набора данных

Подбирая набор методов для конкретного сценария обезличивания, следует заранее понимать, как и для чего эти данные будут использоваться. Исходя из этого, нужно сохранить необходимый уровень связанности и смысловых характеристик либо, напротив, погасить и размыть большую их часть.

Основные этапы формирования сценария обработки данных:

- определиться с целями, исходя из необходимости соблюдения требований регулятора, а также путем выделения свойств данных, принципиально важных для конкретных сценариев;
- сформулировать ожидания от баланса полезности и приватности;
- провести анализ особенностей систем и интеграционных связей между ними;
- скоординировать требования внешних и внутренних регуляторов применительно к сценариям пользователей данных;
- выбрать сценарий создания безопасного стенда или набора данных с учетом понимания влияния этого сценария на общий уровень безопасности.

На этапе формирования модели обезличивания важным шагом становится определение баланса между максимальной безопасностью данных и сохранением пользы от них в конкретном сценарии.

Значимую роль при выборе баланса играет контекст использования самих данных. Так, данные, передаваемые на обработку за пределы компании (например, для внешней аналитики), должны быть скорее безопасными, чем полезными. Но если данные находятся внутри защищенного контура компании и с ними работают только внутренние сотрудники, то баланс можно сместить в сторону сохранения полезности.

Результатом оптимально выстроенного процесса обезличивания станут безопасные данные, полезные для достаточно широкого круга сценариев:

- маркетинг и аналитика — сохраняют сегменты и распределения;
- анализ рисков в банках;
- контроль качества — дают возможность работать с данными, сохраняющими и исходные форматы, и качество;
- для обучения моделей — позволяют обучать модели на данных, отражающих смысловые характеристики исследуемой области.

Перечислим примеры характеристик, которые могут влиять на качество обучения моделей и сбора аналитики:

- гендерный баланс;
- социально-демографическая структура;
- родственные связи;
- страна и оператор в телефоне;
- валидность паспортов, ИНН, СНИЛС, адресов;
- очевидные ошибки в данных.

Например, бессмысленный набор цифр в поле «ИНН», возможно, важно оставить бессмысленным. Спецсимволы, на которые могут при тестировании среагировать интеграционные модули, есть смысл оставить без изменений, чтобы сохранить возможность воспроизведения ошибок.

Следуя описанному пути выбора методики, важно иметь возможность не только настроить баланс между пользой и безопасностью, но и каким-то образом его измерить. Важно понимать, в какой конфигурации данные можно передавать наружу, а в какой — следует оставить их внутри контура. Определение допустимого соотношения безопасности и пользы

обеспечивает аргументы при ответе на вопрос о том, можно ли пойти навстречу аналитикам и сохранить больше качественных показателей и свойств или нужно усилить информационную безопасность и уничтожить большую часть полезных признаков.

ПРИМЕНИМОСТЬ КАЖДОГО ИЗ МЕТОДОВ

Технически существует множество методов обработки, которые при грамотном использовании и сочетании способны минимизировать риск компрометации чувствительных данных. В этой статье мы в первую очередь рассмотрим методики из списка рекомендаций, приведенных в Приказе Роскомнадзора от 19.06.2025 N140 «Об утверждении требований к обезличиванию персональных данных и методов обезличивания персональных данных, за исключением случаев, указанных в пункте 9.1 части 1 статьи 6 Федерального закона от 27 июля 2006 г. N 152-ФЗ "О персональных данных"» [2]

Логичный вопрос: следует ли применять (или хотя бы иметь обязательную возможность применить) все указанные методики или достаточно выбрать некоторые из них, ориентируясь на применимость к типу и способам использования данных?

В зависимости от свойств и состава данных могут возникать сценарии, в которых выбор конкретного метода или их комбинации вполне оправдан и обоснован. Пытаться применять все разрешенные методы к одному набору данных в этом случае необязательно. Выбирать и комбинировать следует самостоятельно, ориентируясь на применимость метода для конкретного вида данных, целевое использование, а также суммарную оценку безопасности формируемой общей модели процесса с учетом контекста использования.

Рассмотрим рекомендуемые Роскомнадзором методы и прокомментируем их применимость в конкретных случаях.

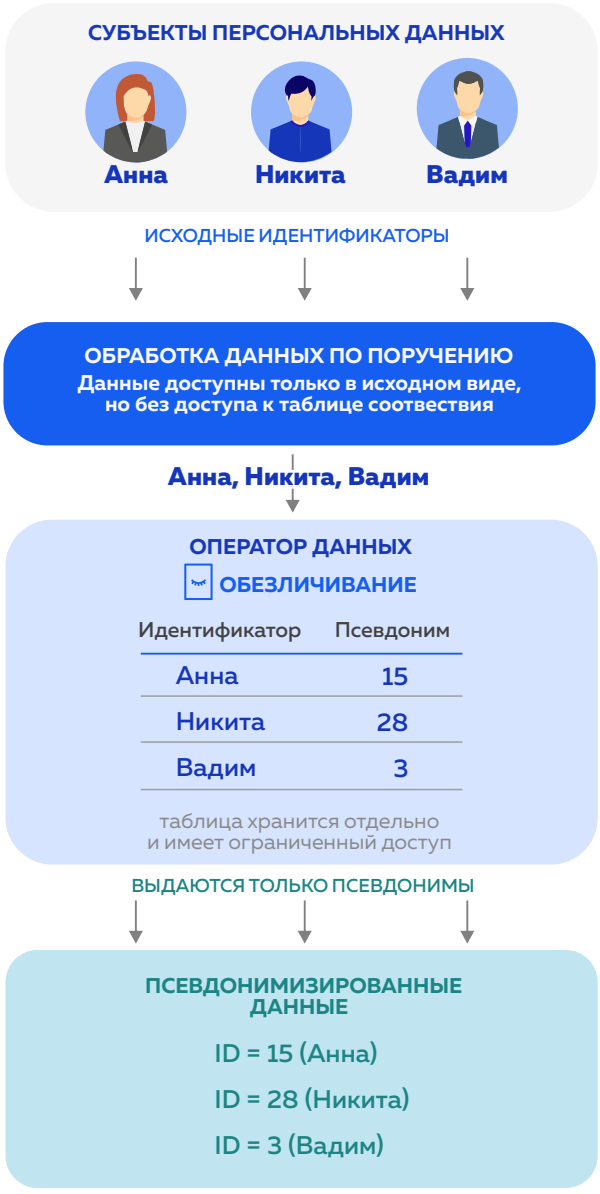
Метод введения идентификаторов

Метод может быть реализован, например, путем сохранения исходных значений ПДн в отдельном справочнике и генерации их идентификаторов. В исходной таблице конкретные значения заменяются на соответствующие им идентификаторы.

Варианты: замены на UID или хэши, допустимые к использованию на уровне рекомендаций регулятора.

Примеры базовых сценариев псевдонимизации через введение идентификаторов [3]

ТЕХНОЛОГИИ ЗАЩИЩЕННОЙ ОБРАБОТКИ ДАННЫХ. ОБЕЗЛИЧИВАНИЕ



Идентификаторы могут быть сформированы с применением счетчика, генератора случайных номеров или симметричного шифрования.

Использование введения идентификаторов ограничивается бизнес-сценариями, где неважен семантический смысл исходных ПДн и их формат. Например, при замене ФИО на UID возможно сохранение соответствия между остальными атрибутами физического лица, однако невозможен учет гендерного баланса, родственных связей и агрегации по национальному признаку, если есть такая необходимость.

Помимо базовых сценариев замены на идентификаторы, существует набор расширенных сценариев:

- на основе асимметричного шифрования;
- с использованием цепочечного шифрования; сложносоставные псевдонимы;
- на основе доказательства с нулевым разглашением;
- по протоколу конфиденциальных вычислений (включая схемы разделения секретов).

Метод изменения состава и семантики

Метод построен на искажении значений атрибутов ПДн, а также, возможно, частичном обобщении или удалении части сведений.

В целом это достаточно гибкий метод, позволяющий в необходимом объеме сохранить качество и смысл данных. Чтобы соблюсти баланс с безопасностью, в сценариях, где применяется этот метод, важно следить за объемами изменений и уровнем размытия данных.

Метод декомпозиции

При использовании декомпозиции персональные данные разбиваются на отдельные элементы так, что некоторая идентифицирующая их часть может быть удалена, а другая часть перенесена в другое место. Оптимальным в этом случае является физическое разделение баз данных, сохраняющих элементы разных категорий.

Чтобы сохранить пользу от данных, этот метод нужно сочетать с псевдонимизацией — созданием сопоставления через карты соответствий (маппинга). Например, для сквозной аналитики имеет смысл учесть возможность или необходимость обратного воссоединения набора по исходным идентификаторам и связям после внешней обработки.

Особое внимание следует уделить раздельному хранению декомпозированных элементов и карт соответствия, учитывая возможность увеличения поверхности атак и утечек.

Метод перемешивания

Метод основан на перестановке отдельных значений между собой. Его можно применять в случаях, когда нужно сохранить валидность общего набора данных — например, обобщенных сумм за период или внешних ссылок на фиксированные справочники.

Стоит обратить внимание на соответствие исходному значению и оценить последствия его исключения в конкретном случае: снизится ли в результате вариативность в малых группах или, наоборот, исключение приведет к тому, что безопасность данных повысится? При выборе данного метода следует обратить внимание на возможность искажения детальной статистики с сохранением групповой.

Метод обобщения или агрегации ПД

Обобщение представляет собой замену точных значений на более общие категории или диапазоны, затрудняющую идентификацию субъекта. Агрегация (частный случай обобщения) — это объединение отдельных записей в групповые показатели (сумма, среднее, количество и т. д.): вместо конкретных данных на обработку передаются обобщенные метрики.

Для обоих этих методов важно, чтобы в созданных группах находилось достаточное количество элементов — это необходимо, чтобы исключить возможность однозначного сопоставления элементов групп с исходными значениями.

Для преобразования точных значений в менее детализированные применяются различные механизмы:

- обобщение числовых данных: округление, категоризация (переход к диапазонам), агрегирование до групповых статистик;
- обобщение категориальных данных: иерархическое обобщение (движение вверх по таксономии);
- укрупнение категорий: объединение детальных категорий в более общие, сокращение множества значений.

Варианты подходов к обобщению для числовых типов:

- биннинг — каждый элемент данных заменяется на значение ближайшей границы того диапазона, в который попадает исходное значение;
- замена средними значениями;
- перевод числовых в категориальные значения с указанием диапазона;
- дифференциальная приватность (DP).

Примеры:

Категориальное значение возрастных характеристик с учетом диапазона

Исходное значение	Обобщенное значение
50	50–60 лет
54	50–60 лет
51	50–60 лет

Варианты подходов к обобщению текстовых и других типов данных:

- маппинг по справочнику с соотнесением на уровни группы;
- усечение;
- замена с использованием генерации псевдослучайных чисел с учетом групп;
- дифференциально-экспоненциальная приватность.

Примеры:

Маппинг по справочнику на уровни на примере профессий

Исходное значение	Обобщенное значение
Врач общей практики	Медицинский работник
Неонатолог	Медицинский работник
ЛОР	Медицинский работник

Усечение даты до уровня года с сохранением формата и использованием константы значения элементов

Исходное значение	Обобщенное значение
02.03.2016	01.01.2016
15.03.2016	01.01.2016

МЕТОДЫ СОХРАНЕНИЯ КОНСИСТЕНТНОСТИ И ИХ ПРИМЕНИМОСТЬ

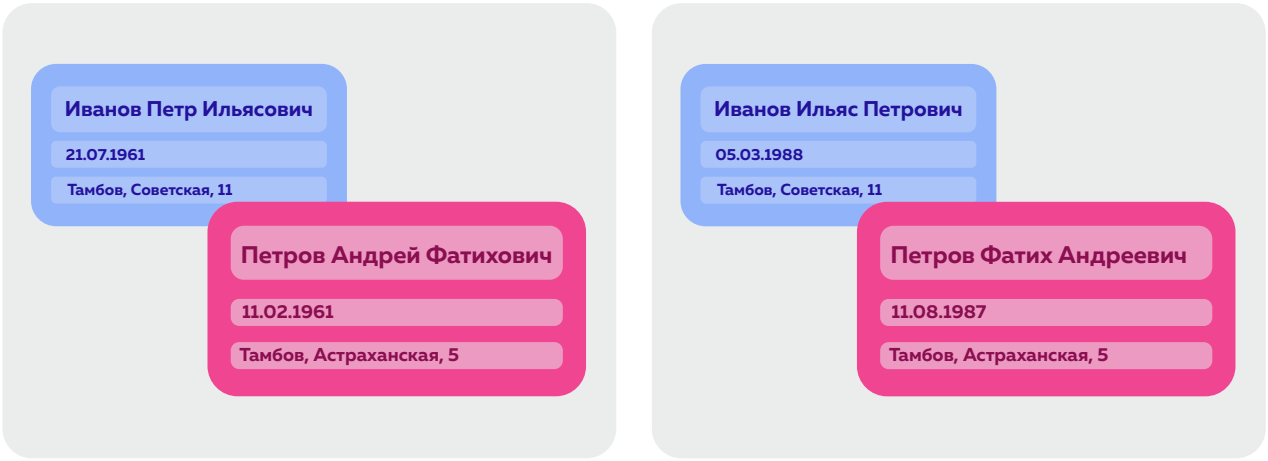
Описание методов сохранения консистентности

В ряде случаев при выборе подходов к обезличиванию возникает необходимость сохранить консистентность исходного набора данных. Консистентностью в данном контексте принято считать единообразие замен для одинаковых исходных данных.

Консистентность замен бывает важно соблюсти при обезличивании повторяющихся значений — в частности, при маскировании нескольких интеграционно связанных систем, а также данных, требующих сквозной аналитики.

Например, в две разные системы банка может быть заведен один и тот же клиент. Чтобы собрать все данные о нем в аналитическую витрину, важно сделать так, чтобы при обезличивании атрибуты клиента в разных системах изменились одинаково.

Другой пример: в исходной системе четко прослеживаются родственные связи между родителями и детьми, проживающими по одному адресу. Если при обезличивании фамилий отчества и адреса будут заменены на разные, неконсистентные значения, то такого рода связи потеряются.



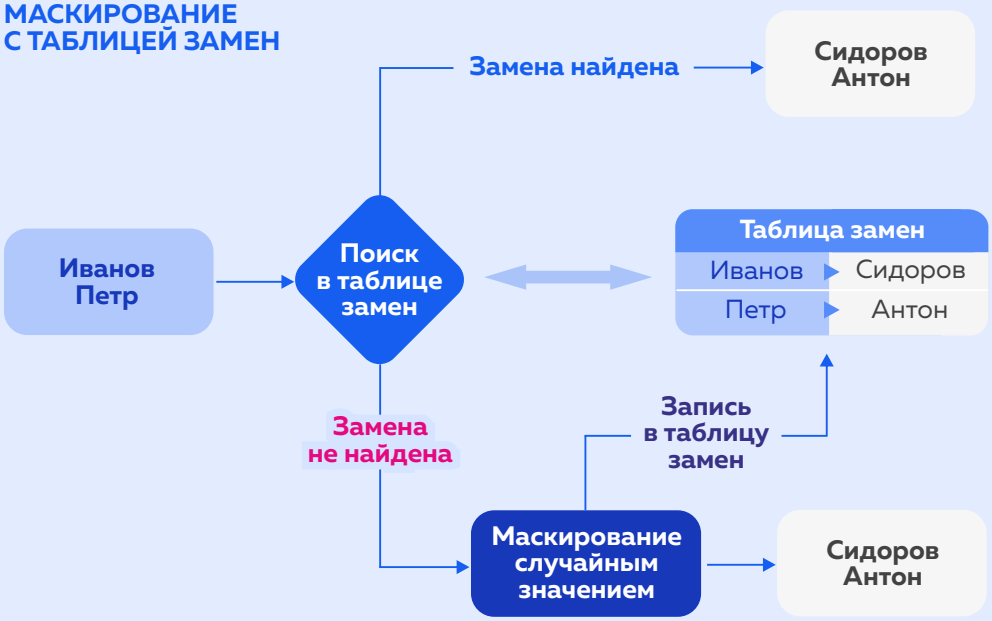
Существует несколько подходов к сохранению консистентности. Каждый из них применим в той или иной мере в зависимости от контекста и целей конкретной обработки данных. Чаще всего выбор сводится к вариантам использования таблицы замен или единой «формулы» замен.

При использовании таблицы замен подобранные пары замен сохраняются: если исходное значение встречается повторно, в качестве результата возвращается ранее подобранная замена.

В подходе единой «формулы» замен используется детерминированный генератор псевдослучайных чисел (ГПСЧ), инициализируемый значением, вычисленным на основе исходных данных, контекстных параметров и секретов. Семейство ГПСЧ-алгоритмов позволяет отказаться от хранения пар замен, сохраняя при этом консистентность маскирования.

ТЕХНОЛОГИИ ЗАЩИЩЕННОЙ ОБРАБОТКИ ДАННЫХ. ОБЕЗЛИЧИВАНИЕ

МАСКИРОВАНИЕ С ТАБЛИЦЕЙ ЗАМЕН



МАСКИРОВАНИЕ С ГСПЧ

Оценка применимости алгоритмов сохранения консистентности

ПРЕИМУЩЕСТВА	ОГРАНИЧЕНИЯ И РИСКИ
· Отсутствие зависимости от хранилищ пар замен · Высокая масштабируемость · Возможность разделения контекстов (сессии, типы данных и т. п.) · Детерминированность результата, обусловленная данными на входе	· При компрометации генератора сидов (seeds) возможно восстановление маскированных данных · Для малых доменов входных данных (короткие строки, номера) возможны атаки полного перебора · При известной функции генерации маскирование остается обратимым, если не обеспечены криптографические меры

Сравнение методов хранения пар замен и метода ГПСЧ

Критерий	ГПСЧ-обезличивание	Хранение пар замен
Согласованность	Высокая (при одинаковом входе)	Средняя (зависит от стратегии хранения)
Скорость	Высокая, зависит преимущественно от CPU	Средняя, в наибольшей степени зависит от скорости хранилища пар замен
Масштабируемость	Лучше масштабируется	Ограничена объемом хранилища
Риски при компрометации	Высокие при утечке параметров	Высокие при утечке хранилища
Управление ключами или секретами	Критически важно	Не требуется
Зависимость от внешних компонентов	Меньше (если ГПСЧ локальный)	Больше (если есть централизованное хранилище)

Рекомендации по усилению безопасности при выборе методики сохранения консистентности замен

Обе описанные методики сохранения консистентности замен применимы в большинстве сценариев. Однако необходимо тщательно продумать архитектуру защиты ключевых элементов, в отдельных случаях проработать разделение контекста. Также важно провести оценку устойчивости к атакам.

Рекомендации по усилению безопасности при использовании ГПСЧ

- Криптографически стойкие хэш-функции.
- Внешний секрет — обязательная часть сид-генерации.
- Контекстуальная изоляция — разные сиды для разных атрибутов и сессий.
- Пороговые или дифференциальные методы постобработки, если требуется устойчивость к статистическим атакам.
- Ограничение детерминизма там, где это допустимо, — например, использование стохастических методов при генерации текстов.
- Ротация и контроль секретов — использование Vault-сервисов.

Рекомендации по усилению безопасности при использовании таблиц пар замен

- Не хранить исходные значения при формировании пар замен.
- Применять криптографически стойкие хэш-функции.
- Шифровать соль, если не используются внешние секреты.

ФОРМАЛЬНЫЕ МОДЕЛИ ПРИВАТНОСТИ (ДЛЯ ОЦЕНКИ ОБОИХ МЕТОДОВ)

Формальные модели приватности определяют математические критерии, позволяющие количественно оценивать уровень обезличивания и лучше понимать, насколько эффективно данные защищены от повторной идентификации, вывода чувствительных атрибутов и связывания записей. Формальные модели служат основой для объективного управления рисками и выбора корректных методов обезличивания, помогая обеспечить соответствие требованиям законодательства и стандартов в области защиты данных.

Что представляют собой формальные модели приватности

Формальная модель задает измеримый уровень приватности, указывая, какие угрозы считаются снятыми и какие метрики используются для проверки качества анонимизации. Благодаря этому приватность данных можно не только описывать, но и проверять, сравнивать и контролировать на всех этапах обработки данных.

Основные категории моделей приватности

1. Синтаксические модели: оценивают структуру и свойства обезличенных данных. К ним относятся следующие модели:

- *k-анонимность* — защита от выделения записи;
- *l-разнообразие* — защита чувствительных атрибутов;
- *t-близость* и *β-сходство* — контроль отклонений распределений.

Синтаксические модели позволяют ограничить вероятность идентификации при анализе структуры набора данных.

2. Семантические модели: обеспечивают защиту независимо от внешних данных и любых знаний нарушителя. И вот основные модели этой категории:

- *дифференциальная приватность (DP);*
- *модель приватности Pufferfish* — представляет собой обобщение DP, отличается устойчивостью к атакам с использованием сторонней информации [4].

В обеих этих моделях обеспечивается защита не столько набора данных, сколько работы алгоритма, что обеспечивает строгие гарантии «неразличимости» участия субъектов.

3. Криптографические модели: обеспечивают возможность работы с данными без их раскрытия. К этим моделям относятся:

- *гомоморфное шифрование (HE);*
- *многосторонние вычисления (SMPC);*
- *доверенные среды исполнения (TEE)* — особенно полезны для совместной обработки данных несколькими организациями.

4. Контекстуальные модели: учитывают доступные злоумышленнику внешние источники и вероятности утечек:

- карта Карно (k-map);
- δ -присутствие;
- приватность на основе δ-разглашения;
- приватность на основе моделей Blowfish или Pufferfish.

Такие модели важны в сценариях, где значительную роль играет совмещение с внешними утечками.

Почему формальные модели критически важны для оценки рисков и качества анонимизации

Столь важное значение формальным моделям придается по следующим причинам:

- Они вводят строгие и проверяемые условия необратимости, снижая риск повторной идентификации.
- Описывают приватность количественно, позволяя сравнивать и оптимизировать методы обезличивания.
- Являются основой для рискориентированного подхода, соблюдения которого требуют современные стандарты и закон N152-ФЗ.
- Обеспечивают воспроизводимость, очень важную для аудита, сертификации и передачи данных третьим лицам.
- Позволяют выбирать корректные методы для конкретных сценариев, в том числе предусматривающих публикацию данных, машинное обучение, обмен с партнерами, использование в медицине и др.

РИСКИ

Понятие риска и необходимость его учета

Обезличивание данных представляет собой обработку данных в целях снижения вероятности раскрытия информации о субъекте при сохранении полезности набора данных. Однако, даже после применения методов обезличивания, остается остаточный риск, связанный с возможностью повторной идентификации, вывода (восстановления) чувствительных атрибутов или связывания информации с внешними источниками.

Риск (R) в контексте защиты данных определяется как сочетание двух компонентов:

1) вероятность реализации негативного события (P) — вероятность успешности атаки нарушителя;
2) тяжесть последствий (C) — ущерб или последствия для субъекта и организации.

Общая оценка риска традиционно (см. ГОСТ Р 58771-2019 «Менеджмент риска. Технологии оценки риска») формализуется таким образом:

$$R = P \cdot C$$

Где:
- P — это вероятность того, что атака приведет к раскрытию данных или информации;
- C — показатель значимости последствий, определяемой законодательством, внутренними нормативными документами и отраслевыми требованиями.

В рамках данной статьи мы концентрируемся именно на оценке вероятности негативных событий, поскольку последствия (C) регулируются локальными и федеративными актами (законом N152-ФЗ, внутренними политиками обработки данных, отраслевыми стандартами) и их количественная оценка представляет собой отдельную правовую задачу.

Баланс между полезностью и риском

Эффективная система обезличивания требует поиска равновесия между двумя ключевыми свойствами данных:
- полезностью (U) — пригодностью набора данных для аналитики, машинного обучения, статистики;
- приватностью (P) — устойчивостью набора данных к атакам.

Их взаимосвязь часто представляется через функции

$$U = f(\varepsilon) \text{ и } P = g(\varepsilon),$$

где ε - параметр приватности (например, в дифференциальной приватности), или через аналогичные параметры в других моделях.

Типичные взаимозависимости между функциями и их параметром таковы:

- чем ниже ε, тем выше приватность, но ниже полезность;
- чем выше ε, тем выше полезность, но ниже приватность.

Баланс между полезностью и риском определяют следующие ключевые факторы:
- задача аналитики;
- требования к качеству данных;
- контекст угроз;
- объем внешних данных у нарушителя;
- уровень риска, приемлемый для организации.

Оценка риска на основе вероятности негативных событий

Вероятность негативного события (то есть успешной атаки) определяется таким образом:

$$P = \sum_{i=1}^{n} w_i \cdot P(Attack_i)$$

Где:

- $Attack_1, Attack_2, \ldots$ — отдельные типы атак;
- $P(Attack)$ — вероятность успешности данной атаки;
- w — весовой коэффициент, отражающий значимость данной атаки в контексте модели угроз.

Такой подход реализован, например, в модели, описанной в работе Маттео Гиоми [5], где риск рассчитывается как сумма взвешенных вероятностей трех основных атак:

1) атака выделения (singling out),
2) атака вывода (inference),
3) атака связывания (linkage).

Модель угроз и поверхность атак

Для количественной оценки вероятности атак необходимо определить две модели:

- модель угроз — описание мотивов, ресурсов и возможностей нарушителей;
- поверхность атак — описание атрибутов и механизмов, доступных злоумышленнику.

Вероятность успешности атаки можно описать функцией, зависящей от контекстных условий:

$$P(Attack) = F_{\text{(модель угроз, поверхность атак, внешние данные)}}$$

Приведем примеры факторов:

- наличие внешних утечек или открытых источников;
- уникальность комбинаций квазиидентификаторов;
- редкие значения чувствительных атрибутов;
- наличие кросс-наборов данных для атак связывания;
- возможность использования машинного обучения для восстановления недостающих атрибутов.

Подходы к количественной оценке рисков

Оценка на основе формальных моделей приватности

Формальные модели задают вероятность успеха определенных классов атак [6]. Подход удобен тем, что риски выражаются в виде аналитически вычислимых формул. Сложность подхода в том, что он требует глубокого учета ограничений каждой модели.

Пример: k-анонимность

Вероятность выделения записи можно оценить по следующей простой формуле:

$$P_{singling\_out} = \frac{1}{k}$$

Если класс эквивалентности содержит k записей, злоумышленник не может выделить конкретного субъекта лучше чем с вероятностью 1/k (в соответствии с требованиями закона N152-ФЗ параметр $k \geq 2$).

Ограничения: k-анонимность не учитывает:

- актуальности внешних источников данных;
- возможности связывания по другим атрибутам;
- атак вывода чувствительных атрибутов;
- корреляции между атрибутами.

Другие примеры:

- l-разнообразие — ограничивает вероятность угадывания чувствительного атрибута;
- t -близость — ограничивает отклонение распределений, но не защищает от атак связывания;
- β-сходство — ограничивает доминирование значений внутри группы;
- δ-присутствие — оценивает вероятность присутствия субъекта в наборе.

Оценка на основе метрик различия и статистических дивергенций

Этот подход предполагает измерение различий между двумя наборами данных — исходным и обезличенным.

Наиболее распространенные метрики следующие:

- расстояние до ближайшей записи (DCR);
- нормализованное расстояние до ближайшего соседа (NDDR);

- дивергенция Кульбака — Лейблера (KDL) и дивергенция Йенсена — Шеннона (JS) для сравнения распределений;
- метрика Вассерштейна (EMD), применяемая для оценки t-близости.

Общая идея метода изложена в следующей формуле:

$$P_{risk} = f(d(original, anonymized))$$

Где d — мера расстояния.

Ограничения этого подхода изложены в работе Георгия Ганева и Эмилиано де Христофаро [7]:

1. Метрики различий плохо учитывают семантику. Например, две записи могут сильно отличаться в абсолютных значениях, но легко связываться по вероятностным структурам.

2. Корреляции между атрибутами могут сохраняться, даже если присутствуют значимые различия.

3. Этот подход конфликтует с полезностью: большие различия снижают риски, но ухудшают качество.

Пример проблемы

Пусть запись в исходном наборе имеет значение возраста, равное 38 лет, а в обезличенном — 40 лет. Различие есть, но если набор содержит только одного человека 35–45 лет в регионе X, то риски связывания сохраняются.

Оценка на основе моделирования атак (attack simulation)

Подход, основанный на моделировании атак (attack simulation), является на сегодняшний день самым точным.

Общая идея метода описывается формулой:

$$P = \sum_i w_i \cdot P_i$$

Где каждая P_i — это вероятность успеха конкретной атаки, рассчитанная по самому эффективному алгоритму.

Ниже перечислим основные атаки.

Атака выделения (Singling Out)

Цель атаки — выявление одной записи среди множества.
Вероятность успеха атаки можно оценить на основе следующих подходов:

- k-анонимность: $P_{singlig\ out} = \frac{1}{k}$;
- анализ редких сочетаний квазиидентификаторов;
- оценка распределений.

Атака вывода (Inference Attack)

Цель атаки — вывод чувствительного атрибута (SA) по значениям квазиидентификаторов.
Формально описать атаку можно через условную энтропию:

$$P_{inference} = 1 - \frac{H(SA|QI)}{H(SA)}$$

Где:
- $H(SA|QI)$ — условная энтропия чувствительного атрибута;
- $H(SA)$ — энтропия исходного распределения.
- QI - квазиидентификатор, SA - чувствительный атрибут.

Чем ниже условная энтропия, тем ниже неопределённость чувствительного атрибута при заданных квазиидентификаторах, и тем выше риск его вывода.

Модель оценки логически объясняет применение l-разнообразия, (c,l)-разнообразия, k-анонимности с учётом чувствительных атрибутов.

Атака связывания (Linkage / Record Linkage Attack)

Цель атаки — сопоставление исходных и обезличенных записей.
Для оценки успеха атаки применяются следующие методы оценки.
Модель Феллеги — Сунтера (классическая вероятностная модель) описывается формулой:

$$Match\ Score(x, y) = \sum_{i=1} log \frac{m_i}{u_i}$$

Где m_i — вероятность совпадения атрибута i у одной и той же записи, u_i — вероятность совпадения у разных записей.

Кластерно-векторная атака (CVPL) использует многомерные векторные представления:

$$P_{linkage} = \frac{1}{[X]} \sum_{x \in X} \mho[closest(x) = x^{+}]$$

Где x^{+}— соответствующая реальная запись.

CVPL указывает на верхнюю границу риска: если записи кластеризуются в одном пространстве, то они потенциально сопоставимы.

Преимущества моделирования атак

Подход на основе моделирования атак обладает целым рядом полезных свойств:

- он охватывает все типы атак;
- учитывает реальные структуры данных;
- позволяет оценить риски при использовании синтетических данных;
- дает возможность получить верхние оценки вероятности, что важно для аудита.

Комбинированная модель риска

На практике оптимальным является подход, основанный на комбинации всех трех методов:

$$R = \sum_{i} w_i \cdot (P_i^{formal} + P_i^{metrics} + P_i^{attack})$$

Этот подход предпочтителен по ряду причин:

- он опирается на формальные гарантии;
- проверяется метриками;
- валидируется моделированием атак.

Подход обеспечивает максимально точную оценку риска, согласованную с современной практикой использования технологий для повышения конфиденциальности, а также с регуляторными требованиями и потребностями бизнеса.

Преимущества применения модели рисков

Модель рисков является обязательным компонентом качественного обезличивания. Она позволяет провести широкий круг мероприятий:

- определить уровни остаточного риска;
- сопоставить эффективность методов;
- выбрать параметры моделей приватности;
- обеспечить юридическую и техническую защиту;
- сбалансировать полезность и приватность данных;
- обеспечить предсказуемую и проверяемую безопасность данных.

Модель угроз составляется в соответствии с рекомендациями ФСТЭК [8].

КЕЙСЫ ПРИМЕНЕНИЯ МЕТОДИК ОБЕЗЛИЧИВАНИЯ

Финансовый сектор

Для финансового сектора обезличивание может применяться в рамках аналитики по основным направлениям бизнеса, а также с целью качественного улучшения фактуры для используемых инструментов тестирования и обучения моделей.

Ниже перечислены направления, в которых обосновано применение обезличенных наборов данных с учетом сохранения их качественных и смысловых характеристик:

- кредитный конвейер: расчет потенциальных рисков;
- профиль клиента: churn-анализ, прогнозирование оттока, маркетинговые таргетированные рассылки;
- противодействие мошенничеству (anti-fraud): тестирование гипотез по выявлению подозрительных операций;
- управление инвестиционными портфелями: выявление устойчивых тенденций и прогностика.

Приведем примеры, отражающие специфику сохранения социально-демографических характеристик при их обезличивании в сценариях финансового сектора.

Сохранение диапазонов дат рождения

Используемая в банке модель машинного обучения вычисляет продукты, которые уместно предложить подписчикам рассылки: маркетолог передает на вход модели обезличенные сведения о человеке, модель их анализирует

и подбирает варианты. Результат учитывает, в частности, возраст клиента. Если очень упрощать, взять ипотеку 20-летнему юноше модель посоветует, а пенсионеру — вряд ли.

Если замаскировать даты рождения клиентов случайным образом, возраст изменится и модель будет советовать неподходящие варианты — например, пенсионные вклады вместо кредитов на жилье. Чтобы не дезинформировать маркетологов, важно сохранить возраст клиентов в обезличенной базе.

Сохранение родственных связей

Для сохранения родственных связей важна консистентность замен ФИО, адресов проживания.

Представим, что в базе хранятся данные Ивана Петровича Худина и его дочери Ольги Ивановны Худиной. Эту родственную связь легко потерять при маскировании — например, если обезличить отца, заменив его имя на «Сергей», а дочь сделать «Васильевной», либо замаскировать фамилию отца на «Бабин», а дочь сделать «Травиной».

Потеря консистентности может стать проблемой, если, например, ИИ-модель учитывает домохозяйства. И когда маркетолог отправит ей сведения о клиентах — супругах с детьми, модель проанализирует информацию, но из-за разницы в отчествах и фамилиях не распознает в этих людях родственников, поэтому предложит продукты, которые обычно выбирают бездетные холостяки.

Медицинские исследования и здравоохранение

В контексте медицинских исследований уровни размытия социально-демографических характеристик приобретают высокую значимость. Кроме непосредственного ограничения идентификации субъекта, на потенциальный уровень расширения диапазона данных и возможности их сохранения влияет медицинская тайна.

Примеры сценариев применимости обезличивания для наборов данных:

- индивидуальные карточки пациентов и истории болезни — валидация их ведения и анализ;
- разработка и тестирование лекарственных препаратов;
- редкие заболевания и эпидемии — обмен статистикой, верификация диагноза, второе мнение;
- агрегированная аналитика заболеваний по регионам с учетом сохранения социальных и возрастных групп;
- исследования и клинические испытания;
- телемедицина: сбор анамнеза, второе мнение, массовая валидация поставленных диагнозов;
- анализ использования ДМС по диапазонам лет и гендеру;
- обезличивание данных клиентов по истечении срока согласия обработки ПДн.

Рекламные технологии и розничная торговля

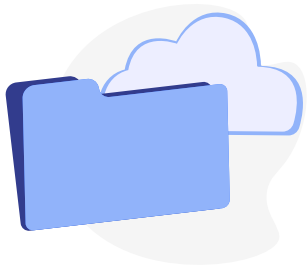
В этой сфере можно выделить следующие примеры сценариев применимости:

- анализ потребительского спроса и таргетирования: маркетинговая атрибуция (касание), обмен данными с целью выявления паттернов и зависимостей транзакций от таргетированных предложений;
- образование экосистем и апсейл-схем, объединяющих дочерние компании;
- рекламные хабы и площадки (агентства): изучение и борьба за спрос, таргетирование и изучение социально-демографических привычек.

Государственные учреждения и органы власти

Приведем примеры известных сценариев применимости в госсекторе:

- ведение реестров с возможностью агрегации: сбор статистики;
- обращения граждан: анализ, выявление паттернов, выделение сегментов.



ЮРИДИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ

Основные нормативные требования по вопросам обезличивания персональных данных в России изложены в Федеральном законе N152-ФЗ от 27.07.2006 «О персональных данных» (ст. 3 п. 9, ст. 5 ч. 7, ст. 6 ч. 1 п. 9, ст. 23 ч. 12) и Приказе Роскомнадзора N140 от 19.06.2025 «Об утверждении требований к обезличиванию персональных данных и методов обезличивания персональных данных».

Особняком стоит регулирование вопросов обезличивания ПДн, предоставляемых операторами в Национальную систему управления данными (НСУД) по требованию уполномоченного органа. Ниже мы коснемся отдельных вопросов, с которыми чаще всего сталкиваются организации, интересующиеся вопросами обезличивания ПДн.

НСУД — национальное озеро данных

Национальная система управления данными (НСУД) создана в целях повышения эффективности обмена и использования государственных данных для предоставления государственных и муниципальных услуг и исполнения государственных и муниципальных функций в электронной форме и иных данных, а также в целях повышения эффективности государственного и муниципального управления (п. 3 Положения, утвержденного Постановлением Правительства РФ N733 от 14.05.2021) [9].

НСУД может наполняться как государственными органами, так и коммерческими организациями. В августе 2024 г. в закон N152-ФЗ была добавлена норма, обязывающая организации предоставлять в НСУД обезличенные персональные данные по требованию уполномоченного органа. Список случаев, когда может потребоваться формирование составов обезличенных ПДн (то есть когда уполномоченный орган может направить организации такое требование), содержится в Постановлении Правительства РФ N538 от 24.04.2025 [10].

В нем перечислены ситуации — такие, например, как введение чрезвычайного положения или проведение статистических исследований, и ничего не говорится ни об отраслевой принадлежности организации, ни о ее размере или масштабах бизнеса.

Правительство также издало ряд нормативных актов, устанавливающих порядок взаимодействия уполномоченного органа и организаций в рамках передачи обезличенных данных в НСУД, требования к обезличиванию ПДн в случае их передачи, правила и методы обезличивания.

С учетом возможных нюансов применения технологий обезличивания в конкретной ситуации, целесообразно еще до внедрения программного продукта или решения для обезличивания оценить существующее регулирование (с учетом возможных изменений в нем уже после выхода данного аналитического доклада) и при необходимости получить необходимые разъяснения от уполномоченного органа.

Обезличивание в деятельности операторов

В рамках деятельности операторы вправе осуществлять обезличивание имеющихся ПДн. Но необходимо учитывать, что в результате обезличивания эти данные не утратят статус персональных, поэтому операторы будут по-прежнему обязаны соблюдать в отношении них все требования, установленные законом N152-ФЗ. На текущий момент обезличивание — это именно мера технической защиты ПДн, а не способ прекращения их обработки (уничтожения).

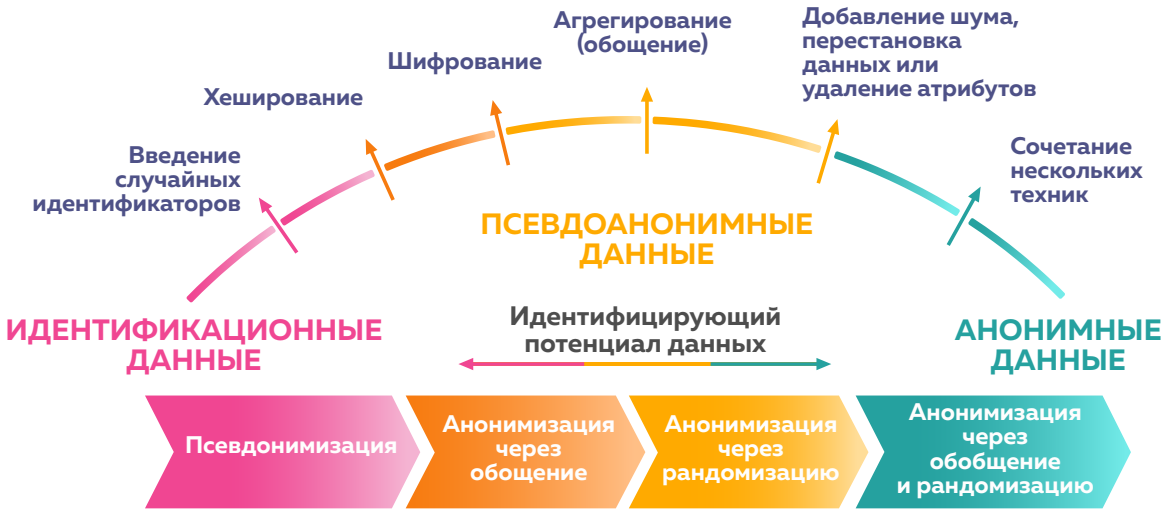
Заинтересованность во внедрении обезличивания может появиться, например, если организация планирует обрабатывать ПДн, не спрашивая согласия субъекта ПДн, в статистических или иных исследовательских целях (кроме прямого маркетинга, согласно п. 9 ч. 1 ст. 6 закона N152-ФЗ) либо в рамках экспериментальных правовых режимов (п. 9.1 ч. 1 ст. 6 того же закона). В этом случае важно соблюдать требования, установленные Роскомнадзором (приказ N140), включая, например, обязанность раздельного хранения ПДн, подлежащих обезличиванию, и обезличенных ПДн (в соответствии с пп. 6 п. 1 Приложения 1 к приказу N140).

Терминология: обезличивание, псевдонимизация, анонимизация

И в литературе, и на практике для описания процесса снижения идентифицирующего потенциала ПДн используются различные термины. Российское законодательство содержит лишь термин «обезличивание». Слово «псевдонимизация» (pseudonimization) сходно с ним

по смыслу, если трактовать его так, как оно понимается в европейском законодательстве. Свой нюанс есть у термина «анонимизация»: он обычно используется для описания ситуаций, когда данные перестают быть персональными в силу того, что более не могут быть соотнесены с каким-либо конкретным человеком.

Учитывая, что в России к анонимизации нормативные требования не установлены, однозначно говорить о том, какие именно действия оператора (кроме полного уничтожения) влекут прекращение статуса ПДн у определенной информации, не приходится.



	ИДЕНТИФИКАЦИОННЫЕ ДАННЫЕ Прямые идентификаторы не тронуты или частично маскированы, косвенные идентификаторы не тронуты			**ПСЕВДОНИМНЫЕ ДАННЫЕ** Прямые идентификаторы удалены или изменены, косвенные идентификаторы не тронуты			**АНОНИМНЫЕ ДАННЫЕ** Все идентификаторы удалены, чтобы разорвать связь данных с человеком
	Беспрепятственная идентификация	Деперсонификация	Усиленная деперсонификация	Псевдонимизация	Усиленная псевдонимизация	Деидентификация	Анонимизация
Прямые идентификаторы Позволяют определить человека без дополнительных сведений или путем ссылки на иные сведения	ФИО: Мунтян Алексей Витальевич СНИЛС: 1234567890	ФИО: **Мунтян А. В.** СНИЛС: 1234567890	ФИО: Мунтян А. В. СНИЛС: *******67890**	ФИО: **5L7T-LX619Z** СНИЛС: *****67890	ФИО: 5L7T-LX619Z СНИЛС: **d5efe83e**	**Прямые идентификаторы удалены**	Прямые идентификаторы удалены
Косвенные идентификаторы Позволяют сопоставлять и объединять данные для определения человека	Возраст: 39 лет IP-адрес: 192.168.0.1	Возраст: 39 лет IP-адрес: 192.168.0.1	Возраст: 39 лет IP-адрес: 192.168.0.1	Возраст: 39 лет IP-адрес: 192.168.0.1	Возраст: 39 лет IP-адрес: 192.168.0.1	Возраст: **30-40 лет** IP-адрес: **252.112.1.90**	**Косвенные идентификаторы удалены**
Примеры техник модификации и защиты данных, снижающих или исключающих возможность определения человека	Техники не используются	Редуцирование ФИО	Маскирование СНИЛС	Использование вместо ФИО искусственных идентификаторов (псевдонимов)	Защита через хеширование СНИЛС	Удаление прямых идентификаторов, обобщение возраста и замена адреса	Удаление косвенных идентификаторов и добавление шума для сокрытия связи данных с человеком

С учетом вышесказанного обезличивание в настоящее время является одной из самых урегулированных технологий защищенной обработки данных в России. Это дает как преимущества в ее использовании в предусмотренных законом случаях, так и ограничения в части технической реализации различных методов. Необходимо учитывать актуальную (на момент применения) редакцию нормативных актов, касающихся обезличивания, и разъяснения уполномоченного органа.

Передача обезличенных данных

Одной из серых зон в законодательстве остается вопрос о статусе данных, полученных в результате обезличивания ПДн, при их передаче организации, не имеющей возможности восстановить их связь с прямо или косвенно определенным или определяемым человеком.

Для оператора, применяющего обезличивание, эти данные продолжают оставаться персональными — на этот счет есть множественные разъяснения от регуляторов, а также прямое указание закона в самом термине «персональные данные, полученные в результате обезличивания» (см., например, ст. 13.1 закона N152-ФЗ). Следовательно, их передача любому третьему лицу должна осуществляться с учетом требований этого закона, включая наличие правового основания на такую передачу: закон, договор, согласие и т. п. (см. подробнее ч. 1 ст. 6, ч. 2, 3 ст. 10, ч. 1, 2 ст. 11 закона N152-ФЗ и др.).

Компания-получатель, в свою очередь, может не иметь ни возможности, ни цели в определении принадлежности таких данных конкретному человеку. Это открывает вопрос о возможном признании таких данных для компании-получателя не имеющими статуса персональных. На текущий момент однозначный ответ на этот вопрос со стороны регулятора отсутствует, в том числе из-за возможных злоупотреблений со стороны недобросовестных участников оборота данных. Заинтересованным лицам целесообразно прорабатывать этот вопрос с учетом конкретных обстоятельств ситуации и при направлении запросов регулятору (если в этом возникнет необходимость) описывать эти обстоятельства детально.

ОСОБЕННОСТИ ВЫСТРАИВАНИЯ ПРОЦЕССА ОБЕЗЛИЧИВАНИЯ ОПЕРАТОРОМ

При формировании процесса обезличивания есть несколько этапов, которые выстраиваются вокруг основного сценария и, как правило, требуют проработки наравне с непосредственным обезличиванием данных.

Нормативная база

При выборе подходящих методик маскирования важно обеспечить преемственность правил между всеми командами, использующими обезличенные данные внутри компании. Это полезно и для ускорения последующих реализаций маскирования данных из разных источников, и для консистентности данных, передаваемых через интеграционные модули.

Имеет смысл начать с проработки и обсуждения возможных вариантов уровня удаления или размытия и сохранения характеристик данных с учетом основных бизнес-сценариев и тест-кейсов.

Определяя оптимальный для конкретного случая баланс между пользой и безопасностью, разумно сделать несколько итераций обезличивания и анализа его результата. При проверке результата следует оценивать как возможность работы с данными в контуре и решения бизнес-задач, так и изменение уровня безопасности.

Инкрементальная обработка

После завершения реализации обезличивания данных необходимо продумать процесс выявления и обработки дельт, появляющихся новых атрибутов и их значений.

Для отслеживания изменений в физической модели и составе данных следует выстроить процесс их транслирования инициаторами изменений, либо обеспечить их автоматическое отслеживание, либо использовать комбинацию этих подходов.

Синтетические данные

Иногда при обновлении копий данных, используемых на тестовых стендах, возникает необходимость дополнительной генерации синтетических значений. На практике это чаще всего

связано с обезличиванием реальных данных перед их передачей в тестовый контур.

Если у команды тестирования имеется типовой набор синтетических данных, его использование может быть недостаточным для воспроизведения отдельных сценариев, особенно в случаях, когда соответствующая система уже находится в продуктивной эксплуатации.

В таких ситуациях допускается формирование дополнительных наборов синтетических тестовых данных, ориентированных на воспроизведение конкретных сценариев. При этом необходимо обеспечить отсутствие конфликтов между обновляемыми реальными данными и ранее сгенерированными синтетическими значениями.

Точки контроля и аудита

В ходе маскирования и по его завершении необходимо предусмотреть возможность отслеживания результатов маскирования, а также влияния изменений, внесенных в модель процесса, на результат. Оценивать можно как сам факт успешного получения результата маскирования с внесением дополнительных метрик в системные журналы (логи), так и контрольные выборки из исходных и маскированных строк.

Здесь же полезно отслеживать хронологию применения методик, включая сведения о том, кто и когда вносил изменения и как они повлияли на результаты.

ПЕРСПЕКТИВЫ

Действующее в России регулирование обезличивания детально описывает методы и техники, однако не затрагивает вопросов оценки достаточности их применения для защиты субъектов ПДн.

Полагаем, что для дальнейшего повышения правовой определенности и создания регуляторных стимулов для более широкого использования обезличивания в различных сценариях, связанных с обработкой чувствительных данных, потребуются новые шаги, нацеленные на создание и закрепление конкретных метрик приватности — защищенности информации о частной жизни граждан. К таким метрикам можно отнести рассмотренные в настоящем докладе численные оценки рисков повторной идентификации, позволяющие определить вероятность отнесения информации в обезличенном наборе данных к конкретному физическому лицу. Подобные оценки позволяют правильно подобрать методы обезличивания для конкретного сценария использования данных, а также определить глубину их применения.

При этом сама модель оценки рисков повторной идентификации может быть закреплена в рамках технического стандарта [11], а правовое регулирование может быть дополнено соответствующими отсылочными нормами для использования операторами ПДн в рамках установленных процедур.

Такой подход позволит ускорить внедрение инновационных методов и подходов к обезличиванию данных в процессы участников рынка и государственных органов, сохраняя должный уровень защиты приватности.

ОБ АВТОРАХ ДОКЛАДА

АЛЕКСЕЙ НЕЙМАН
Исполнительный директор
Ассоциации больших данных,
руководитель FIT Academy of Russia,
Master of Data Science, CDMP, PMP

ВАЛЕРИЙ ХВАТОВ
Специалист
по кибербезопасности
и распределенным вычислениям,
технический директор
DGT Network

ОЛЬГА СЕРДОБИНЦЕВА
Владелец продукта
«Маскировщик»,
IT-компания HFLabs

АЛЕКСАНДР ПАРТИН
Адвокат и партнер
Privacy Advocates,
соучредитель «РППА.Офис»,
сопредседатель
Privacy & Legal Innovation
кластера РАЭК, CIPP/E, CIPM

АЛЕКСЕЙ МУНТЯН
Генеральный директор Privacy Advocates,
внешний менеджер по защите данных
нескольких транснациональных холдингов,
соучредитель Regional Privacy
Professionals Association (RPPA.pro),
сопредседатель Privacy & Legal Innovation и кластера РАЭК

ИСТОЧНИКИ

1. Федеральный закон «О персональных данных» от 27.07.2006 152-ФЗ (последняя редакция) https://www.consultant.ru/document/cons_doc_LAW_61801/d44bdb356e6a691d0c72fef05ed16f68af0af9eb/

2. Приложение 1 к приказу Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций от 19.06.2025 N140 «Требования к обезличиванию персональных данных, за исключением случаев, указанных в пункте 9.1 части 1 статьи 6 Федерального Закона от 27 июля 2006 г. N152-ФЗ "о персональных данных"» https://www.consultant.ru/document/cons_doc_LAW_511184/40cb9b325643edc065344a567cca4ec41c38ace9/

3. DATA PSEUDONYMISATION: ADVANCED TECHNIQUES & USE CASES Technical analysis of cybersecurity measures in data protection and privacy, JANUARY 2021. The European Union Agency for Cybersecurity, ENISA https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20Data%20Pseudonymisation%20-%20Advanced%20Techniques%20and%20Use%20Cases.pdf

4. Hardt, Moritz, Katrina Ligett, and Frank McSherry. 2012. "A Simple and Practical Algorithm for Differentially Private Data Release." arXiv:1012.4763. Preprint, arXiv, March 15. https://doi.org/10.48550/arXiv.1012.4763.

5. Giomi, Matteo, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. 2022. "A Unified Framework for Quantifying Privacy Risk in Synthetic Data." arXiv:2211.10459. Preprint, arXiv, November 18. https://doi.org/10.48550/arXiv.2211.10459.

6. Iwaya, Leonardo, Ala Alaqra, Marit Hansen, and Simone Fischer-Hübner. 2024. Privacy Impact Assessments in the Wild: A Scoping Review. https://doi.org/10.48550/arXiv.2402.11193.

7. Ganev, Georgi, and Emiliano De Cristofaro. 2023. "On the Inadequacy of Similarity-Based Privacy Metrics: Reconstruction Attacks against "Truly Anonymous Synthetic Data"." arXiv:2312.05114. Preprint, arXiv, December 8. https://doi.org/10.48550/arXiv.2312.05114

8. Методический документ. Методика оценки угроз безопасности информации (утв. ФСТЭК России 05.02.2021) https://www.consultant.ru/document/cons_doc_LAW_378330/

9. Постановление Правительства РФ от 14.05.2021 N733 (ред. от 28.05.2025) «Об утверждении Положения о федеральной государственной информационной системе "Единая информационная платформа национальной системы управления данными" и о внесении изменений в некоторые акты Правительства Российской Федерации» https://www.consultant.ru/document/cons_doc_LAW_384215/

10. Постановление Правительства РФ от 24.04.2025 N538 «Об утверждении перечня случаев формирования составов персональных данных, полученных в результате обезличивания персональных данных, сгруппированных по определенному признаку, при условии, что последующая обработка таких данных не позволит определить принадлежность таких данных конкретному субъекту персональных данных» https://www.consultant.ru/document/cons_doc_LAW_504095/92d969e26a4326c5d02fa79b8f9cf4994ee5633b/

11. Проект предварительного национального стандарта «Информационные технологии. КИБЕРБЕЗОПАСНОСТЬ И ЗАЩИТА КОНФИДЕНЦИАЛЬНОСТИ. Методы и технологии анонимизации данных»
https://fstec.ru/tk-362/deyatelnost-tk362/rassmotrenie-dokumentov-smezhnymi-tk/tk-22-informatsionnye-tekhnologii

АССОЦИАЦИЯ БОЛЬШИХ ДАННЫХ

www.rubda.ru

Адрес: Москва,
Пресненская набережная, 10с2

+7 (495) 252-72-60
info@rubda.ru