DOI 10.36622/1682-7813.2025.28.2.005
УДК 004.056.5

МЕТОДИКА АВТОМАТИЗИРОВАННОГО ОБЕЗЛИЧИВАНИЯ ДАННЫХ ДЛЯ ПЛАТФОРМЫ «ДОВЕРЕННАЯ СРЕДА ОБМЕНА ИНФОРМАЦИЕЙ»

О.К. Альсова, Р.А. Пермяков, А.А. Якименко, А.В. Иванов

В статье представлена методика автоматизированного обезличивания данных, которая описывает полный цикл их преобразования для обеспечения конфиденциальности персональной информации о субъекте. Методика представлена в виде укрупненного алгоритма, включающего последовательность взаимосвязанных этапов, реализующих процедуру обезличивания данных и оценку ее эффективности на основе расчета набора количественных показателей и мер. Методика положена в основу разработки программной системы «Обезличивание данных», входящей в состав платформы «Доверенная среда обмена информацией». Платформа разрабатывается совместно Центром компетенций Национальной технологической инициативы ТУСУР, НГТУ и ООО «СИБ», в рамках ключевого проекта центра «Технология интеллектуального управления данными для платформы «Доверенная среда обмена информацией», включающая в себя автоматизированные системы обезличивания и обогащения данных», по договору № 70-2021-00246 от 14.12.2021 г. Методика реализована с учетом требований российского законодательства и международных рекомендаций и стандартов в сфере защиты информации. Применение методики обеспечивает математически гарантированное обезличивания данных на основе применения теории риска.

Ключевые слова: методика обезличивания данных, персональные данные, метод обезличивания, риск раскрытия информации, информационные потери.

Введение

Обезличивание персональных данных – один из ключевых этапов всего процесса управления и защиты информации, который необходим для обеспечения конфиденциальности информации о субъекте персональных данных (ПД). Основная идея состоит в преобразовании данных таким образом, чтобы снизить риск раскрытия информации, но при этом сохранить полезность данных на приемлемом уровне для их дальнейшего использования. Решение этой задачи требует рассмотрения целого ряда вопросов, связанных с подготовкой данных, выбором и реализацией методов обезличивания, оценкой эффективности процедуры обезличивания, созданием программно-инструментальных средств для обезличивания данных.

К настоящему времени накоплен большой мировой опыт в области защиты ПД субъекта на основе применения технологий обезличивания. Однако в России фактически проведено очень мало исследований в этом направлении. Существующие зарубежные методики и программные инструменты обезличивания данных [1-8] не подходят или слабо адаптированы для решения конкретных задач, так как не учитывают специфику российского законодательства в вопросах информационной безопасности [9-13]. Поэтому на настоящий момент крайне актуальна разработка и исследование методов и методик, а также технологий и программных платформ, реализующих процедуры обезличивания ПД субъекта и обеспечивающих их надежную защиту в российских условиях.

Цель исследования заключалась в разработке методики автоматизированного обезличивания данных, которая включает весь процесс преобразования данных: от постановки задачи, выбора и применения методов обезличивания до оценки эффективности процедуры обезличивания в целом. Методика положена в основу работы системы «Обезличивание данных» в рамках платформы «Доверенная среда обмена информацией».

1. Общая характеристика методики автоматизированного обезличивания данных

Приведем общую характеристику разработанной методики.

1. Методика автоматизированного обезличивания данных предназначена для реализации процедуры обезличивания данных с целью расширения возможностей их использования и предоставления широкому кругу потребителей (внешних пользователей) при обеспечении конфиденциальности ПД субъектов.

2. Методика является автоматизированной, что предполагает использование программно-технических средств при реализации этапов методики, но не является автоматической и не может быть реализована без участия специалиста.

3. Методика разработана с учетом требований российского законодательства в сфере защиты информации [9-13] и международных рекомендаций и стандартов в области обезличивания, использования и предоставления обезличенных данных (ОД) внешним пользователям [1-8].

4. Методика представлена в виде укрупненного алгоритма, включающего последовательность взаимосвязанных этапов, направленных на реализацию процедуры обезличивания данных и оценку ее эффективности на основе расчета набора количественных показателей и мер.

5. Методика включает, как один из этапов, оценку риска раскрытия информации на ОД. В рамках этапа математически оценивается эффективность процедуры обезличивания на основе расчета набора количественных показателей, значения которых сравниваются с установленными пороговыми значениями. Таким образом, выполняется математически гарантированное обезличивание данных на базе теории риска в соответствии с мировой практикой и международными стандартами в сфере защиты информации.

2. Укрупненный алгоритм автоматизированного обезличивания данных

Алгоритм, реализующий методику автоматизированного обезличивания данных, представлен в виде блок-схемы на рис. 1. Ниже рассмотрены и описаны основные этапы алгоритма.

1. Постановка и формализация задачи обезличивания данных

На этапе постановки задачи должны быть четко определены и сформулированы ключевые положения (пункты), связанные с процессом обезличивания данных. На рис. 2 представлены основные пункты постановки задачи обезличивания данных в виде структурированного графа, к ним относятся:

– цели и задачи обезличивания данных,

– исходный набор или исходная БД (ИБД), подлежащая обезличиванию,

– подходы, методы, модели, алгоритмы, применяемые на каждом из этапов обезличивания информации,

– условия, ограничения и предполагаемый вид публикации обезличенной БД (ОБД),

– программное обеспечение и т. д.

Ниже рассмотрены и описаны основные пункты постановки задачи.

1.1. Цели и задачи обезличивания данных

Основная цель обезличивания персональных данных – обеспечение конфиденциальности информации о субъекте (субъектах) ПД и гарантирование безопасности любых операций, связанных с ПД (сбор, хранение, обработка, передача) в отношении угроз и атак злоумышленников, направленных на раскрытие личных данных.

В методических рекомендациях Роскомнадзора [11] представлены типовые классы задач, состоящие из наиболее часто встречающихся задач обработки ПД в государственных и муниципальных органах. Согласно [11] к ним относят: статистическую обработку и статистические исследования ПД; сбор и хранение ПД; обработку поисковых запросов (поиск данных о субъектах и поиск субъектов по известным данным); актуализацию ПД, интеграцию данных различных Операторов; ведение учета субъектов ПД. Обезличивание ПД направлено на безопасное выполнение задач обработки ПД с точки зрения обеспечения информационной безопасности и защиты субъекта ПД от раскрытия его личности и/или персональной информации о нем.

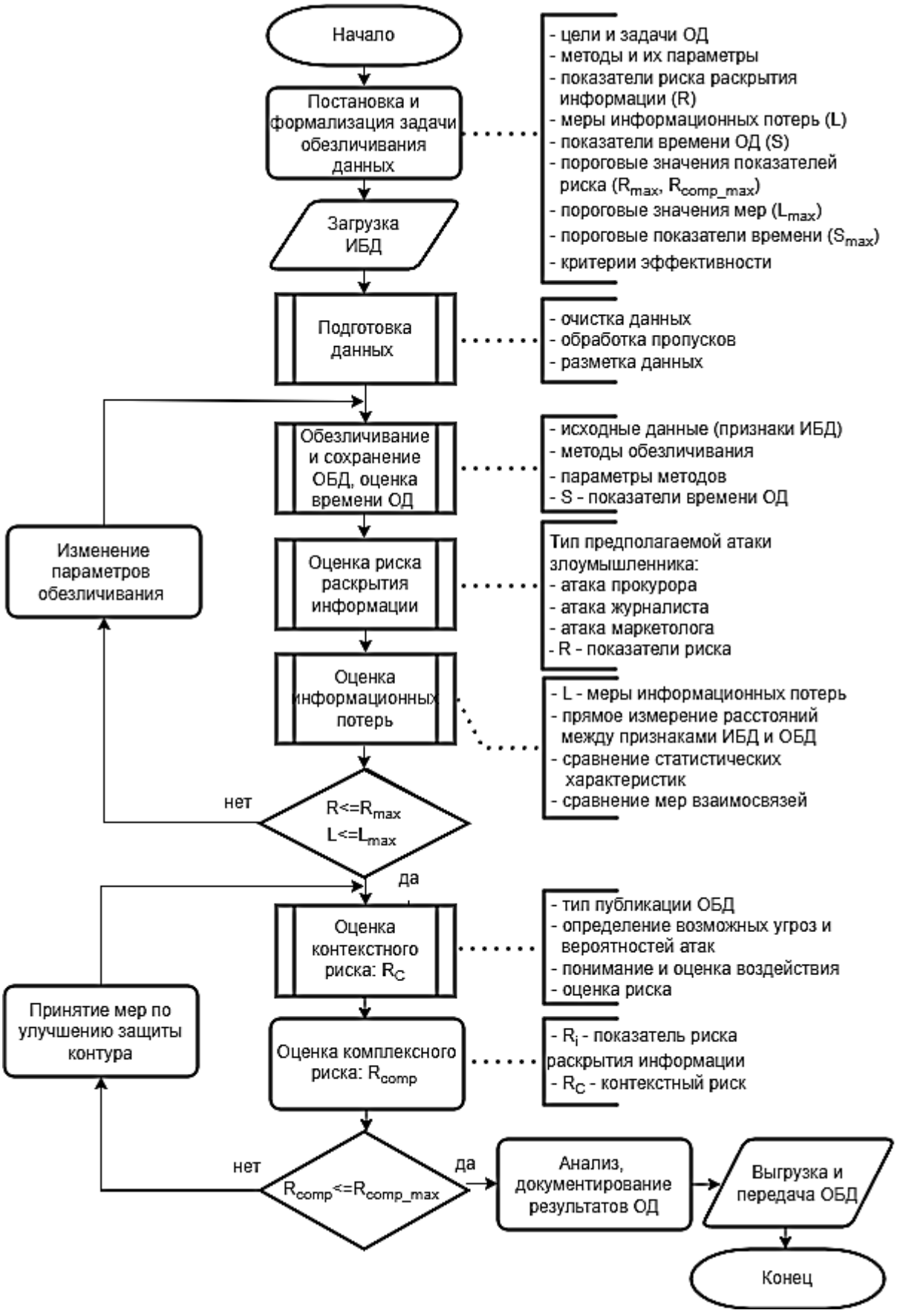

Рис. 1. Блок-схема укрупненного алгоритма автоматизированного обезличивания данных

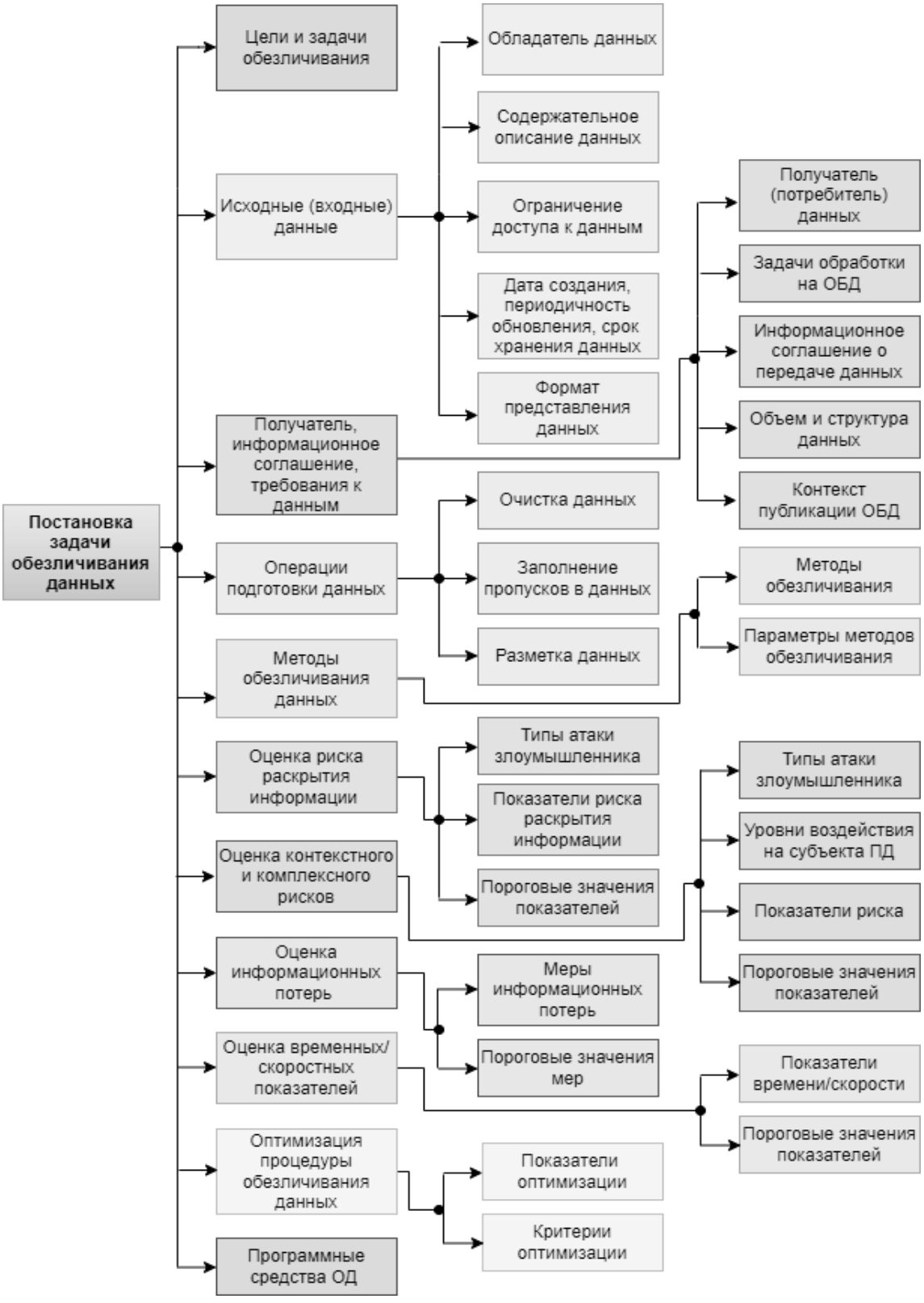

Рис. 2. Структура постановки задачи обезличивания данных

Процесс обезличивания включает решение ряда задач, связных как непосредственно с реализацией обезличивания, так и с оценкой его эффективности. Можно выделить следующие основные задачи:

– выделение и описание признаков (атрибутов) субъектов ПД в ИБД, содержащих конфиденциальную информацию,

– выбор методов ОД и программно-инструментальных средств ОД,

– предварительная обработка ИБД,

– обезличивание ИБД на основе выбранных методов и средств, формирование ОБД,

– оценка эффективности ОД (оценка риска раскрытия информации, оценка информационных потерь),

– документирование основных этапов ОД, полученных результатов, описание ОБД.

1.2. Исходные (входные) данные

Одним из пунктов постановки задачи является описание исходного набора данных, или ИБД. При описании ИБД указываются следующие ключевые позиции:

– обладатель (хранитель) ИБД (полное наименование, общие сведения, контактные данные, ответственное лицо),

– содержательное описание ИБД (предметная область, смысловое словесное описание; цель сбора данных, структура данных, название и смысловое содержание признаков),

– ограничение доступа (наличие признаков, содержащих конфиденциальную информацию о субъекте ПД, описание этих признаков),

– дата создания/последнего изменения ИБД,

– периодичность обновления, срок актуализации и хранения ИБД,

– формат представления данных (текстовый, графический, табличный, иной).

Обладатель (хранитель) данных – лицо, которому принадлежат данные. Обладатель данных несет ответственность за безопасность обработки и хранения данных, обеспечивает контроль процессов передачи и обмена данными путем их обезличивания, а также отвечает за внедрение других средств контроля, которые предотвращают неправомерное (незаконное) использование данных и/или их повторную идентификацию.

Обладателями ИБД могут быть государственные органы и структуры, коммерческие организации, частные компании и физические лица. Согласно Федеральному закону от 27 июля 2006 г. №149-ФЗ (ред. от. 12.12.2023) «Об информации, информационных технологиях и о защите информации» [9] обладателем информации может быть гражданин (физическое лицо), юридическое лицо, Российская Федерация, субъект Российской Федерации, муниципальное образование.

Содержательное описание ИБД включает описание предметной области, к которой принадлежат данные (медицина, демография, экономика, экология и т. п.), что представляют собой данные (например, результаты исследований в определенной сфере), для чего собирались данные, структура данных (набор признаков), название и смысловое содержание признаков ИБД. Выделяются и описываются признаки ИБД, которые содержат ПД субъектов. Именно эти признаки ограничивают доступ к данным и подлежат обезличиванию.

Данные могут обновляться с некоторой периодичностью (частотой), обусловленной требованиями пользователей и зависящей от сути данных. Для оперативных данных периодичность обновления устанавливается ежедневная или еженедельная; для долговременных данных – ежемесячная; ежеквартальная; ежегодная; по мере поступления и т. п. Также может быть установлен срок актуализации и хранения данных, хотя и рекомендуют бессрочное хранение всех версий наборов данных для полного обеспечения информационных потребностей пользователей.

Исходные данные могут быть представлены в разных форматах, в том числе не являющимися машиночитаемыми (например, форматы презентаций – *ppt, pptx*; форматы текстовых документов – *doc, docx, pdf*; форматы изображений – *jpeg, gif, tiff, png* и т. п.). Для последующей обработки данные должны быть преобразованы в один из машиночитаемых форматов (*csv, xls, xlsx, sql, son, xml*) и представлены в структурированном табличном виде.

В [14] предложено ввести паспорт набора данных, который содержит классификационные признаки, позволяющие отнести данные к определенным классам. Паспорт описывает ключевые параметры данных в унифицированном виде, позволяет однозначно идентифицировать набор данных при его публикации в открытых источниках. В [15] описан протокол обработки

(анонимизации) наборов данных для их публикации в открытых источниках.

1.3. Получатель (потребитель) данных, информационное соглашение о передаче данных, требования к данным со стороны получателя

Наборы данных могут обрабатываться как их обладателем, так и передаваться для использования внешнему получателю (потребителю), в том числе возможна открытая публикация данных на общедоступных Интернет-ресурсах. В случае внешнего потребителя поступает запрос на предоставление набора данных, в котором указывается: для чего нужны данные и как будут использоваться (какие задачи предполагается решать на ОБД); требования к данным со стороны получателя (необходимый объем данных; допустимые информационные потери и степень обезличивания). Между обладателем и получателем заключается информационное (пользовательское) соглашение, в котором оговариваются права и обязанности сторон. В частности, режим использования данных пользователем, например: возможность загрузки; использование только в научных и образовательных целях; предоставление результатов обработки данных третьим лицам или их открытая публикация; создание программных продуктов с использованием полученных данных и т. п. Также описываются меры и средства контроля безопасности и конфиденциальности данных, защиты от раскрытия персональной информации, которые должны обеспечить стороны информационного соглашения.

Потребитель (получатель) данных – физическое лицо или организация (компания), которая получает данные от обладателя данных в соответствии с условиями соглашения о предоставлении данных. В частном случае информационный обмен может осуществляться внутри компании или организации, владеющей данными, а не только с внешними потребителями.

Важно определить и содержательно описать задачи, которые в дальнейшем предполагается решать на ОБД. От этого зависит выбор признаков (атрибутов) ИБД, подлежащих обезличиванию, необходимый объем обезличенных данных, выбор методов обезличивания и их параметров, а также пороговых значений рисков раскрытия информации и информационных потерь. Наиболее часто обезличивание выполняется для последующего решения на ОБД исследовательских задач, связанных со статистической обработкой и анализом данных (осуществление выборки по заявленным параметрам и проведение исследований по заданным параметрам субъектов). Отметим, что этот класс задач наиболее интересен с позиции широты применяемых методов обезличивания, так как только для этого класса задач рекомендованы к использованию методы изменения состава и/или семантики, наравне с методами других классов (декомпозиция, перемешивание). Задачи статистического исследования данных условно классифицируются на следующие основные типы: первичный (предварительный) анализ данных; изучение структуры и взаимосвязей между показателями субъекта ПД; кластеризация субъектов ПД; классификация субъектов ПД; прогнозирование показателей субъекта ПД.

В рамках одного статистического исследования возможно решение задач разных классов. Например, первичный анализ данных выполняется, как правило, в начале любого исследования, и включает расчет числовых характеристик признаков, графический анализ, исследование законов распределения признаков. Взаимосвязи между показателями оцениваются для определения предикторов, которые оказывают статистически значимое влияние на результирующий признак и которые следует учитывать при решении задач классификации, регрессии. После решения задачи кластеризации, как правило, решается задача классификации, заключающаяся в отнесении объекта к одному из выделенных классов (кластеров) и т. д.

Определение объема и структуры передаваемых данных (объем выборки из ИБД, набор признаков из исходной совокупности признаков) – важное решение, которое зависит от задач, которые планируется решать на ОБД. Если подробно

рассмотрены и учтены возможности будущего использования ОД, сужена область их применения, то может потребоваться меньший объем обезличивания данных при соблюдении заданного уровня требований к защите информации.

Еще один важный вопрос, который требует рассмотрения – контекст публикации (выпуска) или использования данных, всего возможно три контекста: открытая публикация; закрытая публикация; полуоткрытая публикация. Каждый из них представляет разные отношения и, следовательно, уровень доверия между поставщиком данных и получателем данных. В свою очередь, они будут представлять разный уровень контроля и риска раскрытия конфиденциальной информации. Рассмотрим подробнее возможные контексты публикации (выпуска) данных.

Открытая публикация – публикация данных в открытом доступе в сети Интернет, данные доступны для скачивания и использования без каких-либо условий и ограничений. В случае открытого доступа данные становятся общедоступными, нет никакого контроля над тем, как они будут использоваться. Поэтому требуются самые жесткие предположения об угрозах конфиденциальности, поскольку невозможно оценить мотивы злоумышленника и уровень знаний и инструментов, которыми он может обладать, и которые будет использовать.

Закрытая публикация – публикация данных в ограниченном доступе для определенной категории лиц/организаций. В случае закрытой публикации возможны два варианта: либо проводится внутреннее первичное или вторичное исследование, либо данные передаются для проведения внешнего исследования.

Внутреннее первичное или вторичное исследование (повторное использование данных). Предполагается, что исследование данных проводится внутри организации, владеющей данными. Например, медицинские учреждения и организации ведут истории болезней пациентов, в которых содержится персональная информация, спонсоры клинических испытаний хранят и поддерживают огромное количество данных, собранных в ходе клинических исследований. Использование данных в исследовательских целях требует их обезличивания для сохранения конфиденциальной информации. В ходе клинических испытаний данные собираются для достижения целей отдельных исследований, но совокупная информация часто может оказаться бесценной при выявлении закономерностей, которые не были в центре внимания первоначального анализа. Чтобы иметь возможность использовать данные для целей, не указанных в первоначальном протоколе, спонсоры должны получить согласие на такое использование данных пациентов. Это не всегда может быть возможно или эффективно, альтернативой является обезличивание данных. Любой тип вторичного исследования, где данные используются для целей, отличных от тех, которые указаны в первоначальном протоколе и не охватываются информированным согласием, может потребовать определенного уровня обезличивания.

Внешнее исследование предполагает передачу данных внешним исследователям на договорной основе с соблюдением заданных ограничений и требований к безопасности процесса передачи и использования ПД. В качестве примеров ограничений можно привести: запрет на попытки повторной идентификации; запрет на попытки связаться с любым из субъектов в наборе данных; требование аудита, позволяющее проводить выборочные проверки для обеспечения соответствия соглашению, или требование о регулярных проверках третьей стороной и т. п. Обмен данными с известными исследователями, в соответствии со строгими договорами, с помощью безопасных средств, гарантирует, что процесс безопасен, а связанные с ним риски очень низки.

Полуоткрытая публикация – публикация данных, сочетающая варианты как открытого, так и закрытого доступа к данным. Набор данных доступен любому пользователю для открытого скачивания, однако условием получения данных является необходимость регистрации в организации, предоставляющей данные, и подтверждение согласия на условия использования, обработки и обмена данными (соглашение об условиях использования). В этом случае

дополнительные меры защиты и конфиденциальности данных предусмотрены соглашением об использовании данных, но их трудно обеспечить в силу предоставления открытого доступа к данным.

1.4. Определение операций подготовки данных перед проведением процедуры обезличивания

Определяются операции подготовки данных, которые требуется выполнить перед проведением процедуры обезличивания данных, а также способы и методы их реализации. К основным операциям предобработки данных относятся: очистка; заполнение пропущенных значений; разметка (определение типов признаков ИБД и типов значений признаков ИБД).

1.5. Определение методов обезличивания данных

В этом пункте для каждого признака ИБД, подлежащего обезличиванию, указывается метод (методы) обезличивания и его (их) параметры.

В России определены четыре основных метода обезличивания данных приказом Роскомнадзора [10]:

– метод введения идентификаторов (замена части сведений (значений персональных данных) идентификаторами с созданием таблицы (справочника) соответствия идентификаторов исходным данным),

– метод изменения состава и/или семантики (изменение состава или семантики персональных данных путем их замены результатами статистической обработки, обобщения или удаления части сведений),

– метод декомпозиции (разбиение множества (массива) персональных данных на несколько подмножеств (частей) с последующим раздельным хранением подмножеств),

– метод перемешивания (перестановка отдельных записей, а также групп записей в массиве персональных данных).

Каждый из вышеперечисленных методов фактически представляет собой целую группу (класс) методов, объединенных общим подходом к обезличиванию данных. В рамках каждой группы представлены разнообразные варианты реализации методов обезличивания и возможные алгоритмы их работы. К основным методам изменения состава и/или семантики относятся: глобальное и локальное обобщение, кодирование сверху и/или снизу; локальное подавление, микроагрегирование, добавление шума, округление, маскирование по шаблону, случайная выборка, удаление. Система «Обезличивание данных» обладает полным функционалом, в ней реализованы все классы методов обезличивания данных.

При выборе методов обезличивания учитываются типы признаков ИБД (прямой идентификатор, косвенный идентификатор, чувствительный, нечувствительный) и типы значений признаков ИБД (количественный, порядковый, классификационный), а также требования к свойствам обезличенных данных, которые обеспечивает применение метода [10]:

– полнота (сохранение всей информации о конкретных субъектах или группах субъектов, которая имелась до обезличивания),

– структурированность (сохранение структурных связей между ОД конкретного субъекта или группы субъектов, соответствующих связям, имеющимся до обезличивания),

– релевантность (возможность обработки запросов по обработке ПД и получения ответов в одинаковой семантической форме),

– семантическая целостность (сохранение семантики персональных данных при их обезличивании),

– применимость (возможность решения задач обработки ПД, стоящих перед Оператором, осуществляющим обезличивание ПД, обрабатываемых в информационных системах ПД, без предварительного деобезличивания всего объема записей о субъектах),

– анонимность (невозможность однозначной идентификации субъектов данных, полученных в результате обезличивания, без применения дополнительной информации).

Для обезличивания ПД могут быть выбраны разные методы обезличивания и их параметры, определяющие степень (уровень) анонимизации данных. Для каждого признака ИБД используется определенный метод обезличивания или их совокупность в зависимости от типа признака и типа значений признака. Таким образом, для ИБД можно определить множество моделей (вариантов) преобразования (обезличивания) данных, каждая из которых соответствует определенной комбинации методов ОД с заданными параметрами, примененных для признаков, описывающих субъекта ПД.

1.6. Определение показателей риска раскрытия информации

Определяется набор (множество) показателей риска раскрытия информации R и их пороговые значения $R_{\max}$, которые оценивают эффективность реализации процедуры обезличивания данных:

$$R = \{r_1, r_2, ..., r_p\} = \{r_i\}, \quad (1)$$

где r_i – i-ый показатель риска раскрытия информации,

p – количество показателей риска раскрытия информации,

$$R_{\max} = \{r_{\max 1}, r_{\max 2}, ..., r_{\max p}\} = \{r_{\max i}\}, \quad (2)$$

где $r_{\max i}$ – пороговое значение i-го показателя риска раскрытия информации.

Выбор расчетных формул для оценки показателей риска зависит от предполагаемого типа атаки злоумышленника на уровне данных (прокурора, журналиста, маркетолога). К основным показателям риска в условиях разных типов атак относятся [5, 8, 12]: вероятность повторной идентификации i-ой записи; доля записей, вероятность повторной идентификации которых выше заданного порога; максимальный риск раскрытия информации; средний риск раскрытия информации (глобальный риск). Также должны быть установлены пороговые значения показателей риска раскрытия информации: $R_{\max}$. Пороговые значения показателей определяются в зависимости от ряда факторов, в том числе от специфики предметной области, к которой принадлежат данные, типа публикации (выпуска) ОБД, условий информационного соглашения об обмене данными.

1.7. Определение показателей контекстного и комплексного риска повторной идентификации ОБД

В этом пункте указываются подходы и методы оценки контекстного (степень защиты контура обработки данных) и комплексного риска (R_c, R_{comp}) повторной идентификации ОБД. Расчет показателей риска зависит от ряда факторов, а именно [3, 13]: предполагаемая атака злоумышленника (внутренняя, случайная, внешняя), уровень воздействия, которое оказывает на субъекта ПД раскрытие личной информации о нем в контексте потери конфиденциальности, целостности, доступности; вид публикации (выпуска) ОБД. Также устанавливается пороговое значение комплексного риска повторной идентификации ОБД R_{comp_max} в зависимости от возможных угроз информационной безопасности и требований к уровню защиты ПД.

1.8. Определение мер информационных потерь

Определяется набор (множество) мер информационных потерь L и их пороговые значения L_{max}, которые оценивают эффективность обезличенной БД для ее дальнейшего использования и решения поставленных на ОБД задач:

$$L = \{l_1, l_2, ..., l_m\} = \{l_i\}, \quad (3)$$

где l_i – i-ая мера информационных потерь,

m – количество мер информационных потерь,

$$L_{\max} = \{l_{\max 1}, l_{\max 2}, ..., l_{\max m}\} = \{l_{\max i}\}, \quad (4)$$

где $l_{\max i}$ – пороговое значение i-ой меры информационных потерь.

Используются разные типы мер информационных потерь для обеспечения комплексной оценки дальнейшей применимости ОБД для решения поставленных на ОБД задач [5, 8, 12]:

– прямое измерение расстояний между значениями признаков в ИБД и ОБД,

– сравнение статистических характеристик признаков в ИБД и ОБД (расчет разностей статистических характеристик),

– сравнение мер взаимосвязей между признаками в ИБД и ОБД (расчет разностей мер взаимосвязей).

1.9. Определение временных (скоростных) показателей процедуры ОД

Определяется набор (множество) временных (скоростных) показателей S для оценки быстроты реализации процедуры ОД:

$$S = \{s_1, s_2, ..., s_v\} = \{s_i\}, \quad (5)$$

где s_i – i-ый временной (скоростной) показатель,

v – количество показателей.

В частном случае также задаются пороговые значения показателей $S_{\max}$:

$$S_{\max} = \{s_{\max 1}, s_{\max 2}, ..., s_{\max v}\} = \{s_{\max i}\}, \quad (6)$$

где $s_{\max i}$ – пороговое значение i-ого временного (скоростного) показателя.

Задание пороговых значений имеет смысл в условиях обработки ИБД большого объема, либо при условии, что операции обезличивания ИБД будут выполняться постоянно или с некоторой периодичностью, например, при добавлении (обновлении) информации в ИБД и при жестких временных ограничениях. В любом случае, оценка показателей времени (скорости) обезличивания апостериорна и выполняется после завершения процедуры обезличивания.

В ряде случаев скорость и время, необходимое для выполнения операций обезличивания, также могут стать серьезным ограничением при выборе методов обезличивания. Например, применение методов микроагрегирования к группе признаков связано со значительными затратами вычислительных ресурсов, и время обезличивания этими методами зависит квадратично от объема (количества записей) ИБД. Поэтому методы микроагрегирования ограниченно применяются при работе с большими данными.

1.10. Определение параметров оптимизация процедуры обезличивания данных

Указывается критерий оптимизации, который позволяет выбрать один из вариантов (моделей) обезличивания данных. Для статистической оценки эффективности и выбора варианта ОД используется набор показателей риска раскрытия информации, мер информационных потерь, временных (скоростных) показателей с установленными пороговыми значениями. Фактически решается оптимизационная задача выбора модели преобразования данных в соответствии с установленными критериями при заданных ограничениях в одной из постановок.

Первый вариант постановки оптимизационной задачи: минимизация информационных потерь при выполнении ограничений на пороговые значения показателей риска раскрытия информации. В качестве критерия оптимизации выступает показатель (показатели) информационных потерь, целевая функция минимизирует значение (значения) показателя, в качестве ограничений используется показатель (показатели) риска раскрытия информации.

Второй вариант постановки оптимизационной задачи: минимизация риска раскрытия информации при выполнении ограничений на пороговые значения показателей информационных потерь. В качестве критерия оптимизации выступает показатель (показатели) риска раскрытия информации, целевая функция минимизирует значение (значения) показателя, в качестве ограничений используется показатель (показатели) информационных потерь.

Третий вариант постановки оптимизационной задачи: минимизация информационных потерь и риска раскрытия

информации; введение вектора предпочтений.

Четвертый вариант постановки оптимизационной задачи: минимизация времени (скорости) обезличивания при соблюдении ограничений на пороговые значения риска раскрытия информации. Этот вариант актуален только в условиях обработки больших данных и многократного решения задачи ОД, например, по мере поступления новой информации.

Важно подчеркнуть, что конечной целью обезличивания является снижение риска повторной идентификации субъекта ПД до приемлемого уровня при сохранении максимально возможной полезности данных.

1.11. Определение программных средств обезличивания данных (программные решения, среды, пакеты программ, библиотеки)

Выбор программного обеспечения зависит от того, какие методы и модели обезличивания данных, какие показатели их эффективности планируется использовать. В идеале программное обеспечение должно поддерживать различные методы и алгоритмы преобразования данных, статистической оценки показателей риска раскрытия информации и информационных потерь. Разработано достаточно много зарубежных открытых программных систем обезличивания данных, среди которых можно особо выделить: *ARX - Data Anonymization Tool*, *sdcMicro*, *μ-ARGUS* [16]. Однако эти программные решения реализуют спектр подходов, методов и моделей, которые рекомендованы к применению органами по защите конфиденциальности ПД в зарубежных странах, что не всегда соответствует требованиям российского законодательства в сфере защиты информации.

Поэтому разрабатывается собственное программное обеспечение обезличивания данных в рамках платформы «Доверенная среда обмена информацией», реализующее широкий набор методов ОД, оценки рисков раскрытия информации и информационных потерь с учетом требований Роскомнадзора в области защиты данных и специфики российского законодательства.

2. Загрузка исходных данных (ИБД)

На втором этапе загружаются исходные данные для обезличивания – ИБД, представляющая собой реляционную БД, основанную на реляционной табличной модели представления данных. В контексте задачи обезличивания строки таблицы соответствуют субъектам ПД, а столбцы – атрибутам (признакам) субъекта данных, которые могут принимать как количественные, так и качественные значения.

Выбор реляционной модели представления данных обусловлен преимуществами реляционного подхода в организации данных, а именно: обеспечение безопасного и надежного управления данными на основе правил целостности; обеспечение стандартного способа представления данных и отправки запросов; обеспечение табличного способа хранения структурированной информации и доступа к ней, который интуитивно понятен, гибок и эффективен; обеспечение высокой производительности выполнения операций чтения/записи, поиска данных.

В качестве СУБД выбрана *PostgreSQL*, учитывая ее основные достоинства и преимущества в сравнении с другими СУБД. К основным достоинствам *PostgreSQL* относятся: использование объектно-реляционных БД; высокая расширяемость; поддержка широкого набора типов данных; поддержка БД неограниченных размеров по количеству записей и возможность хранения таблиц размером в 32 Тбайта; многоверсионный контроль параллелизма; открытый и бесплатный исходный код.

Также на вход модуля (подсистемы) обезличивания данных могут подаваться данные в виде файла формата *csv*.

Математически ИБД представляется в виде матрицы:

$$X = \left\{ x_{ij} \right\}_{i,j=1}^{N,P},$$

где x_{ij} – значение j-го атрибута (признака) i-го субъекта ПД в ИБД;

N – количество строк, записей (субъектов персональных данных);

P – количество признаков.

Обезличенная БД (ОБД) также представляется в виде матрицы:

$$X_{об.} = \left\{ x_{ijоб} \right\}_{i,j=1}^{n,p},$$

где $x_{ijоб.}$ – значение j-го атрибута (признака) i-го субъекта ПД в ОБД;

n – количество строк, записей (субъектов персональных данных);

p – количество признаков.

В общем случае не все признаки и/или записи ИБД включаются в ОБД в обезличенном виде, т. е. может быть $P \neq p$ и $N \neq n$.

3. Подготовка данных перед проведением процедуры обезличивания

Выполняются операции, связанные с подготовкой исходных данных перед проведением процедуры обезличивания: очистка данных; обработка пропусков; разметка данных.

Очистка данных предполагает обнаружение и удаление ошибок и несоответствий в данных в целях повышения их качества. В ходе очистки данных выявляются информационные дубли, ошибки регистрации, несоответствия в значениях атрибутов (признаков).

В ходе обработки пропусков выполняется либо разметка пропусков (пропуски сохраняются, и обезличенная БД сдержит пропуски в тех же позициях, что и исходная БД), либо пропуски заполняются с использованием одного из статистических методов заполнения пропусков.

Операция разметки данных предполагает определение типов признаков и типов значений признаков, подлежащих обезличиванию. Признаки соответствуют одному из четырех типов: прямые идентификаторы; квазиидентификаторы (косвенные идентификаторы); чувствительные и нечувствительные признаки.

Прямые идентификаторы – признаки (атрибуты), которые однозначно идентифицируют субъекта ПД. В качестве примера прямых идентификаторов можно привести следующие: ФИО, СНИЛС; номер медицинского полиса и т. п.

Квазиидентификаторы (*косвенные идентификаторы*) – признаки (атрибуты), которые идентифицируют субъекта ПД с той или иной степенью неопределенности. В ряде случаев комбинация косвенных идентификаторов может дать однозначную идентификацию субъекта или обеспечить высокий риск раскрытия информации. Например, сочетание косвенных признаков: дата заболевания, пол, возраст, район местожительства при определенных условиях может привести к идентификации личности и увеличивает риск раскрытия персональной информации о субъекте.

Чувствительны признаки – признаки (атрибуты), которые содержат «деликатную» информацию о субъекте ПД. Например, медицинская информация, особенности течения заболевания, индивидуальные показатели и т. п. Именно эта информация представляет наибольший интерес для решения задач статистического анализа БД и не может быть изменена.

Нечувствительные признаки – признаки (атрибуты), которые не относятся ни к одной из вышеперечисленных категорий и не представляют интереса для дальнейшей обработки. Нечувствительные признаки могут не включаться в ОБД.

Значения признаков измеряются либо в количественной, либо в качественной шкалах, соответственно признаки относятся либо к количественному, либо к качественному типам. Качественные признаки подразделяются, в свою очередь, на классификационные (номинальные) и порядковые (ранговые). Принадлежность признака к одной из шкал измерений зависит от вида операций, которые допустимо выполнять со значениями признака (арифметические операции, операции сравнения). Отдельно выделяется тип признака – дата/время.

Типы признаков ИБД во многом определяют выбор методов и процедур обезличивания, которые будут применяться в ходе обработки ИБД, а также входные данные для этапа оценки риска раскрытии информации. Риск раскрытия информации считается равным 100% для прямых

идентификаторов и оценивается по совокупности показателей в случае работы с косвенными идентификаторами.

4. Обезличивание и сохранение ОБД, оценка времени ОД

Выполняется обезличивание ИБД и последующее сохранение обезличенных данных. Обезличивание реализуется с помощью выбранных на этапе постановки задачи (этап 1) методов обезличивания. Выполняется также оценка временных (скоростных) показателей реализации процедуры ОД. В частном случае, если предполагается многократное выполнение процедуры ОД в условиях обработки больших данных и/или жестких временных ограничений, и скоростные показатели превышают заданные пороговые значения $S_{\max}$, выполняется возврат на этап постановки задачи (этап 1) и выбираются более эффективные методы обезличивания с точки зрения временных затрат. На блок-схеме (рис. 1) частный случай не рассмотрен.

5. Оценка риска раскрытия информации

Выполняется оценка риска раскрытия информации путем расчета показателей риска раскрытия информации R, определенных на этапе постановки задачи.

Формулы расчета показателей риска раскрытия информации зависят от предполагаемой модели атаки злоумышленника (прокурора, журналиста, маркетолога). Как правило, невозможно предположить какому именно типу атаки будет подвергаться обезличенная БД, поэтому имеет смысл рассчитать весь набор показателей риска раскрытия информации для обеспечения комплексной оценки эффективности процедуры обезличивания данных с учетом разных внешних условий и угроз.

6. Оценка информационных потерь

Выполняется оценка информационных потерь, возникающих вследствие обезличивания данных. Выполняется расчет мер информационных потерь, определенных на этапе постановки задачи.

7. Сравнение расчетных и пороговых значений показателей риска раскрытия информации, мер информационных потерь

Выполняется сравнение расчетных значений показателей риска раскрытия информации с установленными пороговыми значениями показателей: $R \le R_{\max}$. Выполняется сравнение расчетных значений мер информационных потерь с установленными пороговыми значениями мер: $L \le L_{\max}$. Если неравенства выполняются, то переход к следующему этапу, иначе изменяются параметры обезличивания и реализуется возврат на этап обезличивания (этап 4). Изменение параметров обезличивания сводится к выбору других параметров методов обезличивания или даже к выбору других методов ОД, либо к установлению менее жестких пороговых (предельных) значений для показателей риска раскрытия информации и/или мер информационных потерь.

8. Оценка контекстного риска

Выполняется оценка контекстного риска, т. е. оценивается насколько защищен контур, в котором обрабатываются ИБД и ОБД. Оценка выполняется с использованием методов экспертного оценивания, т. е. поставщик данных отвечает на ряд вопросов, касающихся организации среды обработки данных (внешней или внутренней), особенностей организации процесса обработки данных, угроз информационной безопасности, возникающих на разных уровнях (сетевые и технические ресурсы, аппаратное и программное обеспечение, стороны и лица, участвующие в обработке, масштабы обработки и т. п.) [3, 13]. В зависимости от ответа на вопрос выставляется оценка в баллах, отражающая степень защиты контура, затем формируется комплексная оценка по совокупности ответов на вопросы, и на заключительном этапе рассчитывается оценка контекстного риска с учетом вероятностей возможных угроз – R_c.

9. Оценка комплексного риска

Выполняется расчет показателя комплексного риска R_{comp}, представляющего

собой произведение риска раскрытия информации (данных) и контекстного риска:

$$R_{comp} = R_{comp} R_i, \quad (7)$$

где R_i – показатель риска раскрытия информации, в качестве которого может использоваться один из показателей в наборе R.

10. Сравнение расчетного и порогового значения показателя комплексного риска

Выполняется сравнение расчетного значения комплексного риска R_{comp} с установленным пороговым значением $R_{comp} \le R_{comp\_max}$. Если неравенство выполняется, то переход к заключительным этапам (этап 11, 12), иначе – принимаются технические и организационные меры, направленные на улучшение защиты контура и снижение контекстного риска при обработке информации. Далее происходит возврат на этап оценки контекстного риска (этап 8) с учетом изменившихся условий.

11. Анализ, документирование результатов ОД

Проводится анализ полученных результатов и их интерпретация, описываются и документируются ход и результаты выполнения процедуры автоматизированного обезличивания данных.

В описании результатов обезличивания данных представляется следующая информация:

– признаки (атрибуты) субъектов ПД из ИБД, значения которых обезличивались;

– методы обезличивания данных, показатели риска раскрытия информации, меры информационных потерь, которые использовались в процессе обработки данных;

– пороговые (предельные) значения показателей риска раскрытия информации и комплексного риска, мер информационных потерь, которые были заданы;

– расчетные значения показателей риска раскрытия информации для ИБД и ОБД;

– расчетные значения показателей контекстного и комплексного риска;

– расчетные значения мер информационных потерь на ОБД;

– расчетные значения временных показателей обезличивания данных;

– критерии оценки эффективности процедуры обезличивания данных и выбора оптимального варианта обезличивания;

– выводы о соответствии расчетных показателей риска раскрытия информации и комплексного риска установленным пороговым значениям, о соответствии расчетных мер информационных потерь установленным пороговым значениям, о степени обезличивания данных и степени защищенности конфиденциальной информации о субъектах ПД, хранящейся в ОБД.

12. Выгрузка и передача ОБД

На последнем этапе реализуется выгрузка и передача ОБД получателям для дальнейшего использования (первичное или вторичное внутреннее исследование, внешнее исследование) или открытая публикация (публичный доступ к данным).

Заключение

В статье приведена общая характеристика и рассмотрены основные положения методики автоматизированного обезличивания данных.

Предложен и описан укрупненный алгоритм, реализующий методику автоматизированного обезличивания данных. Алгоритм представляет собой последовательность взаимосвязанных этапов, итерационно взаимодействующих друг с другом: на любом шаге возможен возврат на предыдущие этапы для их корректировки с учетом анализа промежуточных результатов. Также возможно многократное повторение этапов алгоритма для выполнения установленных требований к процедуре автоматизированного ОД и достижения конечных целевых показателей.

Алгоритм положен в основу работы программной системы «Обезличивание данных», входящей в состав платформы «Доверенная среда обмена информацией», которая реализуется в рамках совместного проекта Центра компетенций Национальной

технологической инициативы по направлению технологии доверенного взаимодействия ТУСУР, НГТУ и ООО «СИБ».

Система «Обезличивание данных» поддерживает основные этапы методики автоматизированного обезличивания данных, требующие программной реализации. В рамках системы реализован широкий набор методов обезличивания данных, оценки рисков раскрытия информации и информационных потерь для обеспечения процедуры обезличивания.

Методика реализована с учетом требований российского законодательства и международных рекомендаций и стандартов в сфере защиты информации. Применение методики обеспечивает математически гарантированное обезличивания данных на основе использования теории риска.

Список литературы

1. ISO/IEC 27018:2019 – Information technology – Security techniques. URL: https://www.iso.org/standard/76559.html (дата обращения 05.04.2025).

2. ISO/IEC 20889:2018 – Privacy enhancing data de-identification terminology and classification of techniques. URL: https://www.iso.org/standard/69373.html (дата обращения 05.04.2025).

3. ENISA Data Pseudonymization: Advanced Techniques & Use Case / Athena Bourka (ENISA): ENISA, 2021. 53 p.

4. Flexible Data Anonymization Using ARX – Current Status and Challenges Ahead/J Software Pract Exper 50, 2020. Vol. 7, P. 1277-1304.

5. Templ M. Statistical Disclosure Control for Microdata: Methods and Applications in R / M. Templ. Cham, Switzerland: Springer, 2017.

6. Benschop T. Statistical Disclosure Control for Microdata: A Practice Guide / T. Benschop, M. Welch., 2019. URL: https://sdcpractice.readthedocs.io/en/latest/ (дата обращения 05.04.2025).

7. De-identification Guidelines for Structured Data / Information and Privacy Commissioner of Ontario. Toronto, Ontario, 2016.

8. Emam K. El. Guide to the De-Identification of Personal Health Information / K. El Emam. Boca Raton, FL: CRC Press. 2013.

9. Федеральный закон от 27 июля 2006 года, №152-ФЗ «О персональных данных». URL: https://base.garant.ru/12148567/ (дата обращения 05.04.2025).

10. Приказ Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций от 05.09.2013 № 996 «Об утверждении требований и методов по обезличиванию персональных данных». URL: https://base.garant.ru/70451476/?ysclid=m917mdewb964376449 (дата обращения 05.04.2025).

11. Методические рекомендации по применению приказа Роскомнадзора от 5 сентября 2013 г. N 996 «Об утверждении требований и методов по обезличиванию персональных данных». URL: https://base.garant.ru/70541864/ (дата обращения 05.04.2025).

12. Приказ Росстата от 19.04.2013, №165 «Методологические положения по формированию массивов деперсонифицированных микроданных годового структурного обследования по форме федерального статистического наблюдения №1 – предприятие «Основные сведения о деятельности организации» общего пользования для представления пользователям в аналитических целях». URL: https://www.garant.ru/products/ipo/prime/doc/70270390/ (дата обращения 05.04.2025).

13. Методический документ. «Методика оценки угроз безопасности информации» (утв. ФСТЭК России 05.02.2021). URL: https://www.consultant.ru/document/cons_doc_LAW_378330/?ysclid=m9jph91le6687045123 (дата обращения 05.04.2025).

14. Борисов Р.С. Паспорт наборов данных и результатов исследований для публикации в открытых источниках / Р.С. Борисов, А.А. Ефименко // Правовая информатика. 2022, №2. С. 66-79.

15. Борисов Р.С. Протокол анонимизации наборов данных для публикации в открытых источниках / Р.С. Борисов, А.А. Ефименко // Правовая информатика. 2023, №2. С. 54-66.

16. 10 Best Data Anonymization Tools and Techniques to Protect Sensitive Information.

URL: https://blog.gramener.com/10-best-data-anonymization-tools-and-techniques-to-protect-sensitive-information/ (дата обращения 05.04.2025).

Новосибирский государственный технический университет
Novosibirsk State Technical University

Центр компетенций НТИ «Технологии доверенного взаимодействия» ТУСУР
NTI Competence Center «Trusted Interaction Technologies» TUSUR

Поступила в редакцию 7.04.25

Информация об авторах

Альсова Ольга Константиновна – канд. техн. наук, доцент, Новосибирский государственный технический университет, e-mail: alsova@corp.nstu.ru

Пермяков Руслан Анатольевич – заместитель директора, Центр компетенций НТИ по направлению «Технологии доверенного взаимодействия» ТУСУР, e-mail: pra@nti.tusur.ru

Якименко Александр Александрович – канд. техн. наук, доцент, заведующий кафедрой вычислительной техники, Новосибирский государственный технический университет, e-mail: yakimenko@corp.nstu.ru

Иванов Андрей Валерьевич – канд. техн. наук, доцент, заведующий кафедрой защиты информации, Новосибирский государственный технический университет, e-mail: andrej.ivanov@corp.nstu.ru

THE METHODOLOGY OF AUTOMATED DATA DEPERSONALIZATION FOR THE PLATFORM "TRUSTED INFORMATION EXCHANGE ENVIRONMENT"

O.K. Alsova, R.A. Permyakov, A.A. Yakimenko, A.V. Ivanov

The article presents a methodology of automated data depersonalization, which describes the full cycle of their transformation to ensure the confidentiality of personal information about the person. The methodology is presented as a consolidated algorithm, including a sequence of interrelated stages implementing the procedure of data depersonalization and assessing its effectiveness based on the calculation of a set of quantitative indicators and measures. The methodology is the basis for the development of the software system "Data Depersonalization", which is part of the platform "Trusted Information Exchange Environment". The platform is being developed jointly by the Competence Center of the National Technology Initiative of TUSUR, NSTU and SIB LLC, as part of the key project of the center 'Technology of intelligent data management for the platform "Trusted information exchange environment", which includes automated systems for depersonalization and data enrichment', under contract No. 70-2021-00246 dated December 14, 2021. The methodology is implemented taking into account the requirements of Russian legislation and international recommendations and standards in the field of information security. The use of the methodology ensures mathematically guaranteed data depersonalization based on the application of risk theory.

Keywords: automated data depersonalization methodology, personal data, depersonalization method, risk of information disclosure, information losses.

Submitted 7.4.25

Information about the authors

Olga K. Alsova – Cand. Sc. (Technical), Associate Professor, Novosibirsk State Technical University, e-mail: alsova@corp.nstu.ru

Ruslan A. Permyakov – Deputy Director of the NTI Competence Center for Trusted Interaction Technologies at TUSUR, e-mail: pra@nti.tusur.ru

Alexander A. Yakimenko – Cand. Sc. (Technical), Associate Professor, Head of the Department of Computer Engineering, Novosibirsk State Technical University, e-mail: yakimenko@corp.nstu.ru

Andrej V. Ivanov – Cand. Sc. (Technical), Associate Professor, Head of the Department of Information Security, Novosibirsk State Technical University, e-mail: andrej.ivanov@corp.nstu.ru