ПРОТОКОЛ АНОНИМИЗАЦИИ НАБОРОВ ДАННЫХ ДЛЯ ПУБЛИКАЦИИ В ОТКРЫТЫХ ИСТОЧНИКАХ

Борисов Р.С. [1]***, Ефименко А.А.***[2]

Ключевые слова: *открытые данные, структура набора данных, метаданные, протокол, методы анонимизации наборов данных, конфиденциальность.*

Аннотация

Цель работы: разработка программно-функциональных моделей анонимизации для совершенствования научно-методической базы регулирования опубликования в открытых источниках наборов данных, потенциально содержащих сведения, составляющие конфиденциальную информацию.

Методы: системно-правовой анализ внешних и атрибутивных свойств и организационных приемов обработки наборов данных, позволяющих обеспечить анонимизацию информации при публикации в открытых источниках.

Результаты: определены требования к машиночитаемым форматам наборов данных; проанализирован понятийный аппарат в области анонимизации; приведена классификация методов анонимизации наборов данных; предложен организационно-технический протокол анонимизации наборов данных для их публикации в открытых источниках, протокол обеспечивает итерационную процедуру полной анонимизации в зависимости от уровня конфиденциальности обрабатываемых данных.

DOI: 10.21681/1994-1404-2023-2-54-66

Введение

Развитие направления поиска, сбора и анализа информации, полученной из общедоступных источников (*Open Source Intelligence, OSINT*), приводит к необходимости совершенствования механизмов противодействия возможным атакам злоумышленников. Одним из источников информации для *OSINT* могут служить публикуемые различными государственными органами и коммерческими предприятиями сведения, обязательные к публичному размещению в соответствии с законодательством РФ [2]. Кроме этого, публикации подлежит информация о всевозможных социально-экономических процессах, протекающих в стране. Эти сведения являются общедоступными и предназначены для исследователей, разрабатывающих эффективные и обоснованные способы совершенствования законодательства, государственного управления, экономики, социальной и других сфер деятельности. Публикуемая таким образом информация может содержать *конфиденциальные* сведения, раскрытие которых может нанести определённый ущерб интересам государства, общества и граждан. Таким образом, возникает противоречие между возможностью публикации информации в открытом доступе и защитой содержащихся в ней конфиденциальных сведений.

К *конфиденциальным* сведениям относятся как персональные данные, так и другая информация (в частности, корпоративные или инсайдерские данные), раскрытие которой может нанести ущерб и которую можно перерабатывать только с разрешения обладателя. Необходимо заметить, что наборы данных и результаты исследований могут обладать следующим *свойством*: они становятся конфиденциальными только при определенном объеме. Например, сведения о цене одного контракта могут быть не конфиденциальными, а цена всех контрактов за год (оборот компании) уже будет конфиденциальной. При обработке данных, обладающих таким свойством, необходимо заранее установить, при каких преобразованиях может возникнуть ситуация появления конфиденциальных сведений. Для этого при заключении соглашения об информационном взаимодействии от обладателя набора данных необходимо получить перечень инсайдерской (коммерческой, служебной и др.) информации организации и проанализировать его на предмет того, какая информация из этого перечня обладает рассмотренным свойством.

[1] ***Борисов Роман Сергеевич,*** *кандидат технических наук, доцент кафедры информационного права, информатики и математики Российского государственного университета правосудия, Российская Федерация, г. Москва.*

E-mail: bestseller@bk.ru

[2] ***Ефименко Алексей Анатольевич,*** *кандидат технических наук, доцент кафедры информационного права, информатики и математики Российского государственного университета правосудия, Российская Федерация, г. Москва.*

E-mail: alex192@mail.ru

Преобразования, которые могут привести к получению данных более строгого с точки зрения конфиденциальности класса, следующие.

Интеграция данных — объединение данных, полученных из одного или нескольких источников (поставщиков данных), и гармонизация этих данных для предоставления пользователю. Если интегрируются данные, обладающие вышеописанным свойством, то пользователю должно быть направлено предупреждение о возможности получения данных, содержащих информацию, составляющую конфиденциальные сведения, а также должен быть предусмотрен механизм, не позволяющий пользователю за несколько запросов данных получить такую информацию.

Агрегирование данных — сведение данных вместе по тому или иному признаку, непосредственно или с использованием определенного алгоритма обработки. Например, если есть реестр контрактов, где каждая запись содержит сведения о сумме контракта, заказчике и поставщике. Из такого реестра можно получить, например, список всех заказчиков, которые встречаются в контрактах, и для каждого заказчика посчитать, сколько контрактов он заключил и какова сумма заключенных им контрактов. Полученные цифры будут агрегированными данными.

Обеспечение сокрытия конфиденциальных сведений, содержащихся в публикуемых данных, может осуществляться посредством реализации специального *протокола обработки набора данных* [5]. При этом отнесение сведений к конфиденциальным осуществляется на основе унифицированного идентификатора (паспорта набора данных) [4] с помощью специализированного классификатора [3]. При реализации такого подхода одной из значимых *задач* является выбор и применение определённых процедур преобразований с целью защиты конфиденциальной информации [5].

В процессе преобразований форматы наборов данных, подлежащих публикации, должны быть приведены к виду, пригодному для машинной обработки, а сокрытие конфиденциальных сведений должно быть надёжным.

Преобразования наборов данных в машиночитаемые форматы

Документы, предлагаемые для публикации в открытых источниках, могут быть представлены в форматах, предназначенных для восприятия человеком, но не являющимися машиночитаемыми (пригодными для машинной обработки). К таким форматам относятся, например, форматы презентаций *PPT* и *PPTX*, форматы приложения *Word* пакета *Microsoft Office*: *DOC*, *DOCX*, популярные форматы документооборота *PDF* (включая отсканированные копии бумажных документов), формат *web*-страниц *HTML*, изображений *JPEG*, *PNG*, *TIFF*, *GIF* и др. Основная цель представления данных в таких форматах состоит в упрощении восприятия информации человеком.

Для обработки машинными методами данные необходимо преобразовать в машиночитаемые форматы. К таким форматам представления и публикации данных относятся: *CSV* (*TSV*), *JSON*, *XML*, *XLSX*, *XLS*, *API*, а также любые из открытых форматов, реализующих модель *RDF*. Задачи преобразования форматов относятся к разряду вычислительно-ёмких и требуют применения специализированных вычислительных средств [1, 7].

На *рис. 1* показаны популярные форматы входных наборов данных, подлежащих машинной обработке.

Формат *CSV* предназначен для хранения данных в плоской табличной форме. В том случае, если используемые данные образуют сложные иерархические структуры, более удобными форматами являются *JSON* и *XML*. Для связывания наборов данных различных форматов используют модель *RDF*. Данные, имеющие значительный объём, можно заархивировать, используя популярную спецификацию открытого стандарта, например, *ZIP*, *GZ*, *7z*, *RAR* и др.

Машиночитаемые форматы данных должны обладать следующими свойствами:

- простота машинной обработки;
- распространённость и унифицированность формата;
- возможность структурирования данных в формате;
- поддержка большого объёма данных в файле формата.

Список наиболее популярных машиночитаемых форматов данных с их краткими характеристиками представлен в *табл.* 1.

Приведение данных к машиночитаемому формату необходимо в первую очередь для предоставления исследователям возможности проведения основных операций логической обработки информации, повышающих удобство работы с данными. К таким операциям относятся *сортировка, округление, выборка* и *группировка* данных [14].

Сортировка предусматривает упорядочение данных по заданному признаку, операции *округления* позволяют привести наборы данных к более удобному для восприятия формату, *выборка* позволяет отобрать записи набора данных в соответствии с некоторым правилом, а *группировка* — объединить данные по заданному признаку.

При этом важной задачей с точки зрения исследователей является гармонизация данных — приведение их к общему виду. Из-за разницы в методологиях расчетов, сбора, происхождения и структуры исходных данных и единицы измерения бывают разные. В результате гармонизации данные приводятся к единым единицам измерения.

Информация о преобразованиях единиц измерения должна заноситься в метаданные набора данных, а именно в *Паспорт набора данных* [4]. Желательно, чтобы форматы, к которым преобразуются данные, соответствовали международным или межгосударственным форматам [5]. Например, на коллегиях Евразийской экономической комиссии на регулярной основе

Рис. 1. Форматы входных наборов данных

Таблица 1

Машиночитаемые форматы данных

Наименование	Краткое описание	Машинная обработка	Унифицированность	Возможность структурирования (схема)	Поддержка большого объёма данных
XLS, XLSX	Форматы используются в табличных процессорах, ограничены по объёму, частично машиночитаемые	Ограничена	Ограничена	Нет	Нет
CSV	Один из наиболее популярных форматов публикации таблиц с поддержкой условно неограниченного объёма данных	Есть	Есть	Частично (Frictionless Data)	Есть
JSON	Популярный текстовый формат обмена данными в веб-приложениях, основанный на JavaScript	Есть	Есть	Есть (JSON Schema)	Ограничена
XML	Популярный в государственных и корпоративных системах текстовый формат для структурированных наборов данных. Представляет собой хорошо стандартизированный расширенный язык разметки	Есть	Есть	Есть (*XSD*)	Ограничена
SQL, BSON, JSON lines	Популярные машиночитаемые форматы баз данных	Есть	Есть	Есть	Есть
ZIP, GZ, 7z, RAR	Форматы архивирования, предназначенные для уменьшения объёма публикуемых данных	Есть	Частично	Частично, зависит от формата	Есть

Рис. 2. Персональные данные

публикуются сведения о единой структуре и форматах предоставления различных статистических документов государствами-членами Евразийского экономического союза[3]. Такой подход значительно упрощает как процедуры публикации данных, так и их дальнейшее исследование.

Методы преобразования данных для возможности их отнесения к менее строгому классу с точки зрения конфиденциальности

Чтобы опубликовать наборы данных или результаты исследований, содержащие конфиденциальную информацию, необходимо их преобразовать таким способом, чтобы выявление конфиденциальных сведений было затруднено или невозможно. Сложность этой задачи обусловлена многообразием видов персональных данных *(рис. 2)* и других конфиденциальных сведений.

При публикации данных необходимо учитывать, в каком виде разрешает публикацию обладатель этого набора данных: в исходном, обезличенном (персональные данные), деперсонифицированном (коммерческая, служебная, профессиональная тайна) или анонимизированном виде, что должно быть отражено в Паспорте набора данных (результатов исследований).

Из всего объёма сведений, имеющих ограничения на возможную публикацию, наиболее хорошо в законодательном плане проработаны вопросы защиты персональных данных. Обработка персональных данных регулируется Федеральным законом «О персональных данных»[4] и множеством других нормативных актов.

Очевидно, что применяемые для обработки персональных данных способы и средства можно распространить и на другие сферы ограничения публикации открытых данных [6, 8, 10].

В соответствии с российским законодательством [2] обработка персональных данных с целью их последующего изучения исследователями должна проводиться только после их обязательного обезличивания. При этом определение принадлежности этих персональных данных конкретному объекту без дополнительной информации должна быть исключена.

В результате обезличивания наборы данных не должны терять такие свои свойства, как *полнота, структурированность, релевантность, семантическая целостность, применимость* и *анонимность* [9, 11].

Под *полнотой* в данном случае понимается сохранение всей информации о персональных данных, которая существовала до обезличивания.

Структурированность предполагает сохранение структуры взаимосвязи между обезличенными данными.

Релевантность обеспечивает семантическое однообразие форм запросов и ответов при обработке обезличенных данных.

Семантическая целостность предполагает соответствие семантики отдельных атрибутов данных до обезличивания и после.

Под *применимостью* понимается возможность применения операций обработки данных без предварительного проведения «деобезличивания» всего их объёма.

[3] Решение коллегии Евразийской экономической комиссии от 12 ноября 2013 г. № 254 (ред. от 29.05.18) «О структурах и форматах электронных копий таможенных документов». URL: https://docs.eaeunion.org

[4] Федеральный закон от 27 июля 2006 г. № 152-ФЗ «О персональных данных» (с изменениями от 06.02.23).

Таблица 2

Соотношение методов обезличивания предъявляемым требованиям

Требования	Метод идентификаторов	Изменение состава и семантики	Декомпозиция	Перемешивание
Полнота	Да	Частично	Да	Да
Структурированность	Да	Да	Да	Да
Релевантность	Частично	Да	Да	Да
Семантическая целостность	Да	Частично	Да	Да
Применимость	Да	Да	Да	Да
Анонимность	Частично	Да	Частично	Да

Анонимность предполагает наличие допустимой пороговой величины для возможного сопоставления записей обезличенных и исходных данных.

Дополнительными требованиями к обезличиванию относятся: возможность обратимости обезличенных данных, увеличение стойкости к деобезличиванию при увеличении объёма и возможность обеспечить заданный уровень анонимности.

Основные термины, касающиеся методов обезличивания персональных данных, введены Приказом Федеральной службы государственной статистики от 19 апреля 2013 г. № 165[5] и включают следующие:

1. *Введение идентификаторов*. Часть сведений, подлежащих обезличиванию, заменяется идентификаторами с составлением справочных таблиц, обеспечивающих обратимость обезличенных данных.

2. *Изменение состава и семантики*. Предполагает изменение исходных данных посредством статистической обработки, путём обобщения или удаления отдельных записей.

3. *Декомпозиция*. Множество персональных данных разбивается на подмножества с последующим их раздельным хранением.

4. *Перемешивание*. Предполагает использование методов перестановки записей массива персональных данных.

Соответствие перечисленных методов и предъявляемым к ним требованиям приведены в *табл. 2*.

В соответствии с российским законодательством [2] обработка обезличенных персональных данных также является обработкой персональных данных. Этот *вывод* можно сделать на основе анализа термина «*обезличивание*», предполагающего, что идентификация субъекта персональных данных в результате обезличивания возможна с использованием дополнительной информации.

Следовательно, при подготовке публикации сведений в открытом доступе необходимо получить согласие субъектов персональных данных на их обработку (что в данном случае является трудноосуществимым) или обеспечить анонимизацию — такой вид обезличивания, в результате применения которого идентификация субъекта персональных данных с использованием дополнительной информации практически невозможна [10].

Понятие анонимизации не используется в законодательстве РФ, но широко применяется в Европейском законодательстве, при этом используются термины «*анонимизированность*» и «*псевдоанонимизация*».

Под *анонимизированностью* понимается такая степень обезличенности, при которой данные не могут быть отнесены к определяемому субъекту. Такие данные уже не считаются персональными. Прямой аналог понятия «*анонимизация*» в законодатеьстве РФ в настоящее время отсутствует [13].

Под *псевдоанонимизацией* понимают обезличивание, при котором восстановление исходных данных возможно при использовании дополнительной информации. То есть данный термин фактически соответствует понятию «*обезличивание*», применяемому в Российском законодательстве.

При этом к обработке анонимизированных данных не выдвигается никаких требований, а к псевдоанонимизированным требуется относиться так же, как и к персональным данным.

Методы псевдоанонимизации персональных данных, используемые в Европейском Союзе, определяются документом *GDPR* (*General Data Protection Regulation* — Общий регламент по защите персональных данных)[6], принятом 27 апреля 2016 г. и вступившем в силу 25 мая 2018 г.

[5] Приказ Федеральной службы государственной статистики от 19 апреля 2013 г. № 165 «Об утверждении Методологических положений 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» // СПС «КонсультантПлюс».

[6] Регламент Европейского Парламента и Совета Европейского Союза 2016/679 от 27 апреля 2016 г. о защите физических лиц при обработке персональных данных и о свободном обращении таких данных, а также об отмене Директивы 95/46/ЕС (Общий Регламент о защите персональных данных / General Data Protection Regulation / GDPR). URL: https://base.garant.ru/71936226

Эти методы несколько отличаются от используемых в РФ и включают: добавление шума (*noise addition*); подмена (*substitution*); *K*-анонимность (агрегация); *L*-разнообразие (*L-diversity*); *T*-близость (*T-closeness*); дифференциальная приватность (*differential privacy*).

В настоящее время в Федеральный закон «О персональных данных» № 152-ФЗ готовятся поправки, которые определяют такой порядок обезличивания информации, который не позволит определить ее принадлежность субъекту персональных данных (анонимизация персональных данных).

На данный момент в России нет отдельного правового режима для обезличенных данных, т. е. фактически они регулируются как персональные данные. В поправках предлагается разработать методологию обезличивания персональных данных и использовать *риск-ориентированный подход* к обезличиванию данных, введение коэффициента обезличивания информации для каждого метода. Предполагается также установка для каждой категории персональной информации своего коэффициента обезличивания.

Дополнительно предлагается оценивать методы обезличивания по специальной модели, которая позволяет оценить риск (*вероятность*) определения принадлежности данных к субъекту персональных данных после применения того или иного метода обезличивания. Такой подход соответствует мировой практике. В том случае, когда после обезличивания субъект персональных данных может быть установлен только после определенных действий с использованием дополнительной информации, то коэффициент оценки риска (вероятность) будет установлен на уровне до 0,8. В случае, когда установление субъекта персональных данных практически невозможно или возможно только после трудоемких и дорогих процедур, данные признаются обезличенными (анонимизированными, по европейской терминологии), а коэффициент оценки риска (вероятность) составляет от 0,8 до 1.

Для возможности применения предложенного подхода уполномоченному федеральному органу, которым является Роскомнадзор, необходимо будет разработать и утвердить методики обезличивания с коэффициентом анонимности от 0,8.

Таким образом, при работе с конфиденциальной информацией, содержащей персональные данные, необходимо применять методы обезличивания, не позволяющие установить субъект персональных данных (анонимизация) или обеспечивающие максимальный коэффициент обезличивания.

При подготовке к публикации наборов данных или результатов исследований, содержащих коммерческую или служебную тайну, предлагается за основу взять методы, описанные в Методологических положениях по формированию массивов деперсонифицированных микроданных (Приказ Росстата от 19 апреля 2013 г. № 165). В документе описаны методы, позволяющие получать и предоставлять исследователям обезличенные статистические данные респондентов — юридических лиц для проведения научных и аналитических исследований, построения экономических моделей, принятия управленческих решений. В соответствии с документом каждая из записей массива данных содержит множество переменных, относящихся к некоторому субъекту. Каждая из переменных может относиться к прямым или косвенным идентификаторам.

К *прямым* идентификаторам относятся переменные, однозначно идентифицирующие субъект персональных данных. *Косвенные* идентификаторы позволяют идентифицировать субъект персональных данных с некоторой вероятностью. При этом определённая совокупность косвенных идентификаторов может обеспечить однозначную идентификацию субъекта.

Переменные могут относиться к классу конфиденциальных или неконфиденциальных. К *конфиденциальным* относятся идентификаторы, содержащие сведения, составляющие коммерческую, служебную тайну, и др. сведения конфиденциального характера. Все остальные идентификаторы относятся к *неконфиденциальным*.

Если переменные относятся к определённому набору неисчисляемых признаков (категорий), то такие переменные называются *категориальными*.

Под *деперсонификацией* (анонимизацией) понимаются процедуры защиты данных от раскрытия (маскировка) с помощью определённых методов. К таким методам относится *модификация* данных — искажение массива данных перед тем, как предоставить к нему доступ, и сокращение — *фильтрация* наборов данных для уменьшения их детализации.

После применения описанных методов деперсонификации итоговый массив данных может принадлежать к одному из множеств: абсолютно анонимные данные — обработанные методами контроля раскрытия информации путем удаления отдельных переменных и модификации данных до такой степени, что идентификация является невозможной, и фактически анонимные данные — такие, для которых невозможно полностью исключить раскрытие конфиденциальных данных.

К формально обезличенным данным относятся такие, у которых удалены прямые идентификаторы, при этом косвенные идентификаторы и наблюдаемые переменные в основном сохраняются.

Общая классификация методов деперсонификации приведена на *рис. 3*.

Непертурбативная маскировка обеспечивает сокращение объёма данных без их модификации. Деперсонификация обеспечивается за счёт снижения уровня детализации или посредством применения к исходному массиву определённых фильтров.

К группе этих методов относятся следующие способы деперсонификации.

Формальная анонимизация представляет собой процесс обезличивания данных и достигается посредством удаления из исходных данных формальных идентификаторов, указывающих на объект. После удаления

Рис. 3. Методы деперсонификации

этих идентификаторов однозначное опознавание объекта может быть проведено только по косвенным идентификаторам. Несмотря на то, что данная процедура не является абсолютно надёжным способом деперсонификации, она является обязательной для публикуемых микроданных.

Выборка является одним из популярных способов предоставления доступа к данным, при котором публикуется отобранный случайным образом диапазон (как правило, небольшой) из массива данных. Для обеспечения деперсонификации выбранного диапазона следует использовать методы пертурбативной маскировки.

Сокращение детализации представляет собой метод деперсонификации наборов данных посредством уменьшения масштаба шкалы измерения или сокращением числа категорий косвенного идентификатора объекта. Популярным приемом для данного подхода является использование интервальной шкалы для *численных показателей* [12, 15].

Кодирование сверху и снизу. Метод представляет собой пороговый способ укрупнения данных. При использовании этого метода в отдельную самостоятельную категорию кодируются переменные, имеющие значения как выше, так и ниже определённого порога.

Локальное подавление. Часто в наборах данных присутствуют экстремальные значения, величины которых значительно отличаются от средних значений. Это особенно актуально для косвенных переменных, которые могут служить маркерами для идентификации объекта. Например, отчество частного лица, состоящее из двух слов, значительно снизит качество деперсонификации. Локальное подавление обеспечивает изъятие таких экстремальных данных из набора.

При удалении экстремальных данных используется два основных подхода:

1. *Удаление экстремальных данных с фиксацией пропущенных значений.* Такой подход позволяет повысить качество наборов данных, при этом сведения о пропущенных данных не будут содержать величину экстремума и его полярность.

2. *Полное удаление экстремальных данных*. Это приводит к незначительному ухудшению *статистических* [12] и *информационных* [15] *показателей* набора данных, однако повышает уровень конфиденциальности сведений. Используется в случаях, когда число экстремумов в наборах данных очень мало.

Описанные варианты метода локального подавления приводят к отклонению данных и изменению оценочных величин, определённых на их основе, поскольку метод локального подавления и другие методы, основанные на сокращении данных, приводят к ухудшению качества данных.

Пертурбативные методы основаны на модификации данных. Массив наборов данных искажается перед тем, как предоставить к нему доступ. Использовать пертурбативные методы следует таким образом, чтобы статистические характеристики, рассчитанные на базе модифицированного массива, несущественно отличались от рассчитанных из оригинального массива данных.

Обмен данными. Данный метод относится к методам модификации и предполагает, что конфиденциальные переменные для отдельных записей меняются местами. Такие изменения должны производиться таким образом, чтобы внешние параметры и характеристики наборов данных оставались неизменными.

Другой вариант обмена данными предполагает перестановку рангов. Значения одной переменной ранжируются в порядке возрастания, затем каждое ранжированное значение меняется местами с другим значением, случайно выбранным в некотором ограниченном диапазоне.

Для того, чтобы избежать избыточной модификации данных, используется *метод вменения* значений ближайшей кластерной единицы. Вменение следует применять с использованием значений ближайших (относительно функции расстояния, использованной в алгоритме кластеризации) не подверженных риску соседей, в противном случае увеличение неопределенности может оказаться недостаточным.

Метод сводится к следующей процедуре:

1. Пусть $x^{конф}$ — значение, которое требуется защитить.

2. Находится ближайшая кластерная единица $x_{кл}$, для которой:

$$\mathrm{d}\left(x^{\text{конф}}, x_{\text{кл}}\right) = \min_{x_c \in C}\left(\mathrm{d}\left(x^{\text{конф}}, x_c\right)\right),$$

где C — множество всех кластерных единиц.

Защищённым значением $x^{конф}$ будет являться $x_{кл}$.

Микроагрегирование. В основе этого метода лежит следующее правило. Набор данных, представляющий агрегированные данные, может быть опубликован, если отдельные записи соответствуют группам в составе *k* или более объектов (принцип *k*-анонимности), ни один из этих объектов не является доминирующим в группе (т. е. не определяет групповые показатели), а *k* — пороговое значение. Базовый *принцип* микроагрегирования состоит в строгом соблюдении правил конфиденциальности и обусловливает выполнение подмены индивидуальных значений значениями, рассчитанными для малых множеств (микроагрегатов).

Для получения микроагрегатов исходная совокупность единиц наблюдения определенным образом разделяется на небольшие группы ближайших друг к другу объектов размером не менее *k*. Классические алгоритмы микроагрегирования требуют, чтобы все группы (возможно, кроме одной) имели размер *k*. Если количество всех объектов *N* кратно *k*, то создается $n = N/k$ групп по *k* объектов в каждой. Если *N* не кратно *k*, то последняя группа, содержащая менее *k* объектов, объединяется с предыдущей группой и, таким образом, содержит более чем *k* объектов. Затем для каждой группы рассчитывается среднее значение переменной, после чего это значение используется вместо оригинальных данных для всех единиц данной группы. Таким образом, реальный объект заменяется некоторым суррогатным объектом. Особое внимание при этом необходимо уделять тем объектам, которые по своим статистическим и информационным показателям [12, 15] значительно отличаются от других.

Добавление шума. Метод добавления шума используется для численных переменных. К истинному значению некоторой переменной из набора данных добавляется случайная величина и получившееся значение заменяет истинное. Распределение случайной величины выбирается отдельно для каждого значения модифицируемой переменной. В общем случае случайное значение должно иметь нулевое среднее, а в случае недопустимости отрицательных значений переменной они заменяются на нулевые. Степень защищённости набора данных при таком подходе определяется величиной вводимого шума. Необходимо заметить, что искажения набора данных будут увеличиваться при повышении защищённости.

Округление. Метод используется для численных значений и предполагает замену исходных истинных значений округлёнными. Округлённые значения выбираются из одномерного или многомерного массива точек округления.

Протокол анонимизации конфиденциальных данных

Наборы данных/результаты исследований могут быть представлены в бумажном и электронном виде и помещаются соответственно в бумажные или электронные папки (электронные директории).

Обобщенная схема процесса обработки наборов данных/результатов исследований от момента получения до момента публикации приведена на *рис. 4*.

Предварительным шагом для инициализации процедуры публикации набора данных/результата исследования является поступление запроса на передачу информации. После чего с Обладателем набора данных заключается Соглашение об информационном вза-

публикации информация состоит непосредственно из набора данных/результата исследования и Паспорта набора данных/результата исследования как аннотации — формализованного описания данных. Паспорт набора данных/результата исследования проверяется на предмет отсутствия в нём защищаемой законодательством информации по наличию в Паспорте набора данных/результата исследования соответствующей отметки и собственноручной подписи Обладателя набора данных — лица, предоставляющего эти данные.

Перед тем как обеспечить доступ пользователей к набору данных/результату исследования, который содержит конфиденциальную информацию, набор данных/результат исследования должен быть подвергнут обработке (*преобразованию*) для обеспечения недопустимости утечки конфиденциальных данных. Кроме того, порядок доступа пользователей должен регулироваться лицензионными договорами с правообладателями и частью IV Гражданского кодекса РФ[7].

Непосредственно обработка данных при анонимизации включает в себя два этапа. На *первом* этапе используются методы сокращения данных (выборка, локальное подавление, сокращение детализации), на *втором* этапе — методы модификации (микроагрегирование, обмен данными и добавление шума). Для повышения уровня защиты конфиденциальных сведений после проведённых преобразований может использоваться округление.

Протокол анонимизации данных можно представить в виде следующей процедуры и регламента:

***Шаг* 1. Анализ исходного набора данных.**

На этом этапе при необходимости проводится преобразование файла набора данных в машиночитаемый формат. Проводится выявление и анализ набора идентифицирующих переменных и их классификация (прямые, косвенные, конфиденциальные и др.).

Из исходного файла данных на основе *семантического анализа* текста формируется обобщенный вектор ключевых элементов, определяющий режим доступа к данным и представляющий собой одномерный массив *V[N]*, в общем случае неограниченной длины, где *N* — количество ключевых элементов от всех возможных источников.

Каждая ячейка *V[i]*, *i*=1…*N*, содержит число, указывающее на количество появлений в файле *i*-го ключевого элемента.

По каждому ключевому элементу принимается решение о наличии в нём конфиденциальных сведений. Если принимается решение, что *i*-й ключевой элемент неактуален и в дальнейшем не используется, ячейка *V[i]* остается зарезервированной под *i*-й ключевой элемент, но ее значение всегда равно нулю. Если принимается решение, что неактуальны и в дальнейшем не используются первые *k* ключевых элементов, то вводится константа *k* так называемого *стартового отступа* вектора *V[N]*, в котором первые *k* ячеек принимают нулевые значения.

Такой подход позволяет не использовать *k* начальных ячеек вектора *V[N]*, при этом реальная длина вектора уменьшается на *k*. При обработке используется стартовый отступ *k* вектора *V[N]*, что позволяет модифицировать обобщенный вектор ключевых элементов, не модифицируя сами методики определения конфиденциальных данных.

***Шаг* 2. Предобработка переменных.**

Из набора выявленных на предыдущем этапе переменных удаляются прямые идентификаторы, а также переменные, которые могут привести к спонтанной идентификации либо к идентификации на основе сведений из внешних источников.

Проводится сокращение детализации отдельных сильно идентифицирующих численных переменных, на основе которых создаются соответствующие новые категориальные переменные.

Для сформированного массива исходных данных вычисляются основные статистические характеристики. Для последующего выбора методов деперсонификации и их параметров проводится изучение пользовательских предпочтений.

***Шаг* 3. Оценка риска раскрытия конфиденциальных сведений**.

Оценивается риск деанонимизации для различных комбинаций ключевых категориальных конфиденциальных переменных набора данных. Пороговое значение риска определяется экспертными методами.

Результатом обработки конкретной переменной может быть, в частности, число, называемое уровнем конфиденциальности, которое можно интерпретировать, например, как: номер регламентируемой папки; ценность информации документа; достаточный уровень для понижения класса конфиденциальности путем модификации информации документа; необходимый уровень для повышения класса конфиденциальности путем модификации повторной обработки документа или его Паспорта другой методикой.

***Шаг* 4. Выбор и применение методов анонимизации.**

В том случае, если на этапе оценки риска деанонимизации ключевых категориальных конфиденциальных переменных получено значение, превышающее пороговое, проводится глобальное перекодирование этих переменных с последующим выполнением процедур оценки риска.

Если значение оценки риска деанонимизации ключевых категориальных конфиденциальных переменных *не превышает* пороговое, то принимается решение по тем переменным, риск деанонимизации которых высок. Они могут быть оставлены в массиве данных в неизменном виде или перекодированы, например, методом кодирования сверху и снизу.

Если на этапе оценки риска деанонимизации ключевых числовых конфиденциальных переменных получено значение, *превышающее* пороговое, проводятся

[7] Гражданский кодекс Российской Федерации (ГК РФ), Принят Государственной думой 30 ноября 1994 г. № 151-ФЗ.

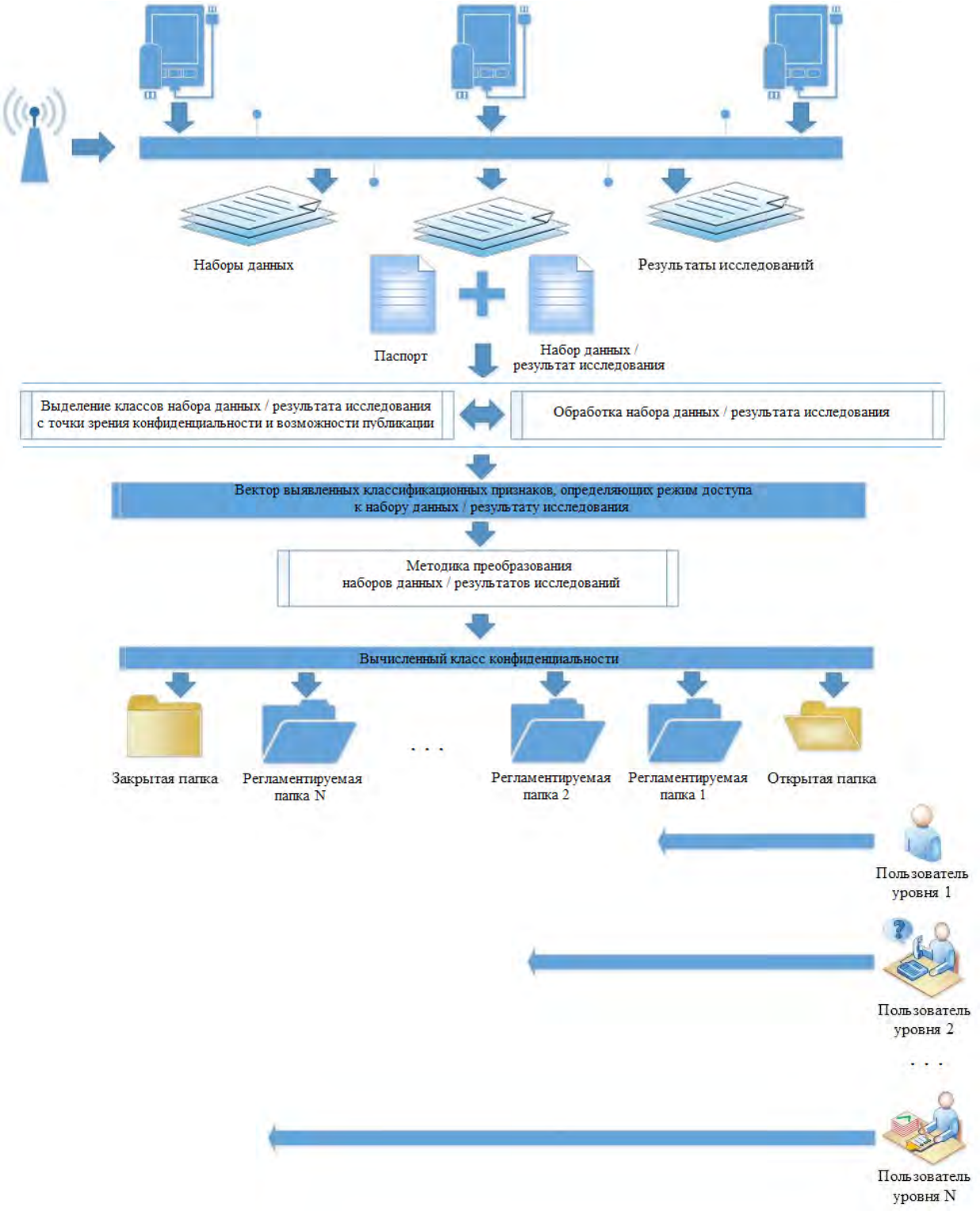

Рис. 4. Обобщенная схема обработки наборов данных

процедуры модификации соответствующих переменных, включающие, например, обмен данных, микроагрегирование и добавление шума. Выбор параметров методов анонимизации проводится на основе обеспечения требуемого уровня защищённости данных экспертными методами. После анонимизации необходима повторная оценка риска раскрытия конфиденциальных сведений на основе алгоритмов кластеризации.

Если на этапе оценки риска деанонимизации ключевых числовых конфиденциальных переменных получено значение, *не превышающее* пороговое, то принимается решение по тем переменным, риск деанонимизации которых высок. Они должны быть модифициро-

ваны с помощью локального подавления, микроагрегирования или обмена данными.

В том случае, если необходимо обеспечить соответствие набора данных ранее опубликованным сведениям, производится корректировка набора для сохранения итоговых значений отдельных показателей и их комбинаций для каждой из публикуемого набора категориальных переменных.

Для повышения уровня защиты конфиденциальных сведений в случае необходимости производится округление числовых переменных, не признанных идентифицирующими.

***Шаг* 5. Оценка качества преобразований наборов данных.**

Контроль обеспечения конфиденциальности данных после проведения анонимизации проводится с применением алгоритмов оценки риска.

Проверка предполагаемых к публикации переменных, которые могут привести к спонтанной идентификации, экспертами — опытными специалистами по обследованию. Если таковые будут выявлены — использование индивидуальных методов защиты.

***Шаг* 6. Составление краткого описания изменений набора данных.**

Все проводимые с набором данных преобразования должны быть включены в Паспорт набора данных [4]. Краткое описание преобразований должно включать, какие именно методы и к каким переменным были применены. При этом технические подробности, в том числе в части использования индивидуальных методов защиты, которые позволили бы восстановить идентифицирующие переменные, должны отсутствовать.

Необходимо оценить уровень потери информации и занести в Паспорт набора данных степень модификации данных в результате использования методов анонимизации.

Заключение

Рассмотренный протокол анонимизации наборов данных для публикации в открытых источниках включает процедуры преобразования наборов данных в машиночитаемые форматы, а также обеспечивает выбор эффективных методов деперсонификации для сокрытия конфиденциальных сведений. Выбор того или иного метода зависит от уровня оценки риска деанонимизации отдельных ключевых конфиденциальных переменных. По результатам применения методов преобразования наборов данных краткое описание преобразований заносится в Паспорт набора данных. Для оценки качества набора данных после проведения преобразований оценивается уровень потери информации по отношению к исходному набору данных.

Особенностью предложенного протокола является комбинирование методов преобразования наборов данных, позволяющее получать данные, менее строгие с точки зрения конфиденциальности и пригодные для публикации без угрозы их раскрытия.

Реализация предложенных процедур анонимизации конфиденциальных данных рекомендуется в составе информационно-программного обеспечения существующих и перспективных платформ организаций, принимающих, обрабатывающих и публикующих наборы данных с целью повышения уровня публичной открытости и доступности наборов данных и обеспечения информационной безопасности.

Рецензент: Омельченко Виктор Валентинович, *доктор технических наук, профессор, заслуженный деятель науки и техники РСФСР, советник секретариата научно-технического совета ВПК «НПО Машиностроения», г. Москва, Российская Федерация.*
E-mail: omvv@yandex.ru

Литература

1. Борисов Р.С. Эффективный алгоритм управления переработкой судебной статистической информации // Правовая информатика. 2018. № 1. С. 15—22. DOI: 10.21681/1994-1404-2018-1-15-22 .
2. Борисов Р.С., Ефименко А.А. Анализ федерального законодательства об ограничении публикации открытых данных // Государственная власть и местное самоуправление. 2022. № 2. С. 42—47. DOI: 10.18572/1813-1247-2022-2-42-47 .
3. Борисов Р.С., Ефименко А.А. Классификатор правовых актов для установления правового режима публикуемой информации // Правовая информатика. 2021. № 4. С. 31—45. DOI: 10.21681/1994-1404-2021-4-31-45 .
4. Борисов Р.С., Ефименко А.А. Паспорт наборов данных и результатов исследований для публикации в открытых источниках // Правовая информатика. 2022. № 2. С. 66—79. DOI: 10.21681/1994-1404-2022-2-66-79 .
5. Борисов Р.С., Ефименко А.А. Протокол обработки наборов данных для их публикации в открытых источниках // Правовая информатика. 2021. № 2. С. 59—70. DOI: 10.21681/1994-1404-2021-2-59-70 .
6. Ващекин А.Н., Дзедзинский А.В. Правовое регулирование отношений в цифровом пространстве // Правосудие. 2020. № 2. С. 108—114.

7. Ефименко А.А. Интегрированная интеллектуальная технология оптимизации параллельных алгоритмов в высокопроизводительных вычислительных системах // Труды Междунар. науч.-прак. конф. «Современные тенденции в науке, технике, образовании» (31 января 2016 г.). Смоленск : Новаленсо, 2016. С. 59—61.
8. Коваленко А.О. Протокол рациональной переработки и правовые режимы судебной информации // Правовая информатика. 2019. № 2. С. 49—56. DOI 10.21681/1994-1404-2019-2-49-56 .
9. Ловцов Д.А. Теория защищенности информации в эргасистемах : монография. М. : РГУП, 2021. 276 с. ISBN 978-5-93916-896-0.
10. Ловцов Д.А. Проблема гарантированного обеспечения информационной безопасности крупномасштабных автоматизированных систем // Правовая информатика. 2017. № 3. С. 66—74. DOI: 10.21681/1994-1404-2017-3-66-74 .
11. Ловцов Д.А. Системология правового регулирования информационных отношений в инфосфере. М. : РГУП, 2016. 316 с. ISBN 978-5-93916-505-1.
12. Ловцов Д.А., Богданова М.В., Лобан А.В., Паршинцева Л.С. Статистика (компьютеризированный курс) / Под ред. Д.А. Ловцова. М. : РГУП, 2020. 400 с. ISBN 978-5-93916-834-2.
13. Ловцов Д.А., Федичев А.В. Архитектура национального классификатора правовых режимов информации ограниченного доступа // Правовая информатика. 2017. № 2. С. 35—54. DOI: 10.21681/1994-1404-2017-2-35-54 .
14. Федосеев С.В. Применение современных технологий больших данных в правовой сфере // Правовая информатика. 2018. № 4. С. 50—58. DOI: 10.21681/1994-1404-2018-4-5-58 .
15. Lovtsov D.A. Informational Indices of the Efficiency of Control Systems for Complex Dynamic Objects // Automation and Remote Control, 1995. Vol. 55. No. 12. Part 2. Pp. 1824—1829.

A DATA SETS ANONYMISATION PROTOCOL FOR THEIR PUBLICATION IN OPEN SOURCES

Roman Borisov, *Ph.D. (Technology), Associate Professor at the Department of Information Technology Law, Informatics and Mathematics of the Russian State University of Justice, Moscow, Russian Federation. E-mail: bestseller@bk.ru*

Aleksei Efimenko, *Ph.D. (Technology), Associate Professor at the Department of Information Technology Law, Informatics and Mathematics of the Russian State University of Justice, Moscow, Russian Federation. E-mail: alex192@mail.ru*

Keywords: *open data, data set structure, metadata, protocol, data sets anonymisation methods, confidentiality.*

Abstract

Purpose of the work: developing program and functional anonymisation models for improving the research and methodological basis for regulating the publication in open sources of data sets possibly containing confidential information.

Methods used: system and legal analysis of external and attributive properties and organisational techniques for processing data sets allowing to ensure anonymisation of information when published in open sources.

Study findings: requirements for data sets machine-readable formats are identified. An analysis of the conceptual apparatus in the field of anonymisation and a classification of data sets anonymisation methods are presented. An organisational-cum-technical data sets anonymisation protocol for their publication in open sources is proposed which ensures an iteration procedure for full anonymisation depending on the processed data confidentiality level.

References

1. Borisov R.S. Effektivnyi algoritm upravleniia pererabotkoi sudebnoi statisticheskoi informatsii. Pravovaia informatika, 2018, No. 1, pp. 15–22. DOI: 10.21681/1994-1404-2018-1-15-22 .
2. Borisov R.S., Efimenko A.A. Analiz federal'nogo zakonodatel'stva ob ogranichenii publikatsii otkrytykh dannykh. Gosudarstvennaia vlast' i mestnoe samoupravlenie, 2022, No. 2, pp. 42–47. DOI: 10.18572/1813-1247-2022-2-42-47 .
3. Borisov R.S., Efimenko A.A. Klassifikator pravovykh aktov dlia ustanovleniia pravovogo rezhima publikuemoi informatsii. Pravovaia informatika, 2021, No. 4, pp. 31–45. DOI: 10.21681/1994-1404-2021-4-31-45 .
4. Borisov R.S., Efimenko A.A. Pasport naborov dannykh i rezul'tatov issledovanii dlia publikatsii v otkrytykh istochnikakh. Pravovaia informatika, 2022, No. 2, pp. 66–79. DOI: 10.21681/1994-1404-2022-2-66-79 .

5. Borisov R.S., Efimenko A.A. Protokol obrabotki naborov dannykh dlia ikh publikatsii v otkrytykh istochnikakh. Pravovaia informatika, 2021, No. 2, pp. 59–70. DOI: 10.21681/1994-1404-2021-2-59-70 .
6. Vashchekin A.N., Dzedzinskii A.V. Pravovoe regulirovanie otnoshenii v tsifrovom prostranstve. Pravosudie, 2020, No. 2, pp. 108–114.
7. Efimenko A.A. Integrirovannaia intellektual'naia tekhnologiia optimizatsii parallel'nykh algoritmov v vysokoproizvoditel'nykh vychislitel'nykh sistemakh. Trudy Mezhdunar. nauch.-prak. konf. "Sovremennye tendentsii v nauke, tekhnike, obrazovanii" (31 ianvaria 2016 g.). Smolensk : Novalenso, 2016, pp. 59–61.
8. Kovalenko A.O. Protokol ratsional'noi pererabotki i pravovye rezhimy sudebnoi informatsii. Pravovaia informatika, 2019, No. 2, pp. 49–56. DOI 10.21681/1994-1404-2019-2-49-56 .
9. Lovtsov D.A. Teoriia zashchishchennosti informatsii v ergasistemakh : monografiia. M. : RGUP, 2021. 276 pp. ISBN 978-5-93916-896-0.
10. Lovtsov D.A. Problema garantirovannogo obespecheniia informatsionnoi bezopasnosti krupnomasshtabnykh avtomatizirovannykh sistem. Pravovaia informatika, 2017, No. 3, pp. 66–74. DOI: 10.21681/1994-1404-2017-3-66-74 .
11. Lovtsov D.A. Sistemologiia pravovogo regulirovaniia informatsionnykh otnoshenii v infosfere. M. : RGUP, 2016. 316 pp. ISBN 978-5-93916-505-1.
12. Lovtsov D.A., Bogdanova M.V., Loban A.V., Parshintseva L.S. Statistika (komp'iuterizirovannyi kurs). Pod red. D.A. Lovtsova. M. : RGUP, 2020. 400 pp. ISBN 978-5-93916-834-2.
13. Lovtsov D.A., Fedichev A.V. Arkhitektura natsional'nogo klassifikatora pravovykh rezhimov informatsii ogranichennogo dostupa. Pravovaia informatika, 2017, No. 2, pp. 35–54. DOI: 10.21681/1994-1404-2017-2-35-54 .
14. Fedoseev S.V. Primenenie sovremennykh tekhnologii bol'shikh dannykh v pravovoi sfere. Pravovaia informatika, 2018, No. 4, pp. 50–58. DOI: 10.21681/1994-1404-2018-4-5-58 .
15. Lovtsov D.A. Informational Indices of the Efficiency of Control Systems for Complex Dynamic Objects. Automation and Remote Control, 1995. Vol. 55. No. 12. Part 2. Pp. 1824–1829.