

И. Ю. Кучин

ЗАЩИТА КОНФИДЕНЦИАЛЬНОСТИ ПЕРСОНАЛЬНЫХ ДАННЫХ С ПОМОЩЬЮ ОБЕЗЛИЧИВАНИЯ

Введение

Современное общество уделяет большое внимание глобальному обмену информацией, которая, вероятно, становится самым важным и востребованным ресурсом. Информация о субъекте собирается постоянно, как только он обращается в лечебное учреждение, оплачивает коммунальные услуги, заключает договор на услуги связи и т. п. Крупные организации, которые после принятия Федерального закона № 152 «О персональных данных» (далее Закон) [1] стало принято называть «операторами персональных данных», не торопятся принимать адекватные меры по их защите. В большей степени это обусловлено несовершенством самого Закона и других законов и нормативных актов, регулирующих это направление.

Долгое время вопросы защиты персональных данных в информационных системах в России никак не регулировались. Все эти годы огромные массивы информации, касающиеся жителей нашей страны, находились фактически в свободном доступе в сети Интернет.

Даже если Закон в итоге заработает и все операторы персональных данных приведут-таки свои информационные системы в соответствие всем требованиям регулирующих органов, а судебные дела за разглашение персональных данных или за необеспечение их конфиденциальности будут доводиться до логического конца, еще долго можно будет за определенную сумму получить практически любую информацию, касающуюся конкретного субъекта.

Законодатель предусматривает возможность установления режима, открытости персональных данных, делая исключения из условия о конфиденциальности для общедоступных массивов сведений и обезличенных персональных данных.

Обезличивание персональных данных

Неразрывная связь персональных данных с личностью их субъекта и возможность идентификации последнего требуют повышенной защищенности в процессе их обработки. В случаях если из них могут быть исключены сведения-идентификаторы, режим конфиденциальности, устанавливаемый для персональных данных, снимается. Законодатель допускает использование обезличенных персональных данных без согласия их субъекта в целях проведения статистических, социологических, исторических, медицинских и других научных и практических исследований [2].

В Законе с обезличиванием связаны следующие статьи:

Статья 3 пункт 8: «*Обезличивание персональных данных – действия, в результате которых невозможно определить принадлежность персональных данных конкретному субъекту персональных данных*».

Статья 7 пункт 2: «*Обеспечения конфиденциальности персональных данных не требуется:*

- 1) *в случае обезличивания персональных данных;*
- 2) *в отношении общедоступных персональных данных*».

При этом остается нераскрытой процедура обезличивания персональных данных, соответствие ее определенному этапу их обработки, степень обезличивания. Все это порождает множество споров, т. к. каждый понимает процедуру обезличивания по-своему.

Здесь важно, что нет никаких сведений о том, на каком этапе обезличиваются данные, равно как нет никаких сведений о степени идентификации субъектов персональных данных (раз уж законопроект использует этот термин) при обработке персональных данных для статистических целей. Для любого человека, знакомого со спецификой работы систем учёта населения, очевидно, что существует огромная разница между статистической обработкой изначально неперсонифицированных (обезличенных) данных и статистической обработкой тех же данных, полученных при условии полной идентификации субъектов персональных данных с последующим обезличиванием уже в процессе обработки. В первом случае риск нарушения конституционных прав невелик, а во втором случае необходимо обеспечение дополнительных мер безопасности на каждом этапе обработки данных, включая этап обезличивания этих данных [3].

Еще туманнее дело обстоит с использованием личных данных в научных целях. В пункте 3 части 2 статьи 6 Закона, в качестве исключения из общего правила согласия субъекта персональных данных на их обработку, указываются случаи, когда «обработка персональных данных осуществляется для статистических или иных научных целей при условии обязательного обезличивания персональных данных». Однако следом, в пункте 6, говорится о допустимости несогласованной обработки, если таковая «осуществляется в целях профессиональной деятельности журналиста либо в целях научной, литературной или иной творческой деятельности при условии, что при этом не нарушаются права и свободы субъекта персональных данных». Из сопоставления этих пунктов неясно, требуется ли при обработке данных в научных целях обезличивание данных либо достаточно, если при этом не будут нарушены права человека. Ведь если для статистических целей предпочтительно использование обезличенной информации, то научные исследования в широком смысле иногда нуждаются в точных сведениях о конкретных людях, особенно применительно к истории, политологии, литературоведению и иным гуманитарным дисциплинам [4].

Кроме того, в связи с определенными затратами, связанными с защитой персональных данных, в Интернете все чаще появляются мнения, что недоработки в Законе, связанные с обезличиванием, оператор может использовать как «лазейку» – он может отказаться от обеспечения конфиденциальности, заявив, что данные, хранящиеся в его информационной системе, являются обезличенными, предварительно удалив или на время «припрятав» некоторые поля в базе данных. И действительно, в условиях отсутствия каких-либо критериев обезличивания доказать что-либо и привлечь оператора к ответственности будет очень сложно.

На Западе обсуждаемая проблема была давно осознана. Учеными из университетов многих европейских государств ведутся исследования и предлагаются различные алгоритмы обезличивания персональных данных [5–7]. В России данная проблема остается неизученной. Подавляющее большинство специалистов по защите персональных данных представляют себе процедуру обезличивания как замену полей, содержащих фамилию, имя, отчество, а иногда и адрес субъекта на некий идентификатор в некоторой базе данных, информацию в которой признают обезличенной. При этом создается еще одна база данных, в которой происходит «привязка» идентификатора к удаленным из исходной базы данных, критичных для опознавания их владельца.

По нашему мнению, описываемый способ не может считаться методом «обезличивания», т. к. в самом определении этого понятия в Законе подразумевается необратимость данного процесса. Обезличенные данные ни при каких обстоятельствах не должны стать персональными, что подтверждается применением наречия «невозможно». В приведенном методе в любой момент по желанию оператора персональных данных либо в случае компрометации «второй» базы данных, информация, хранящаяся в «первой», из обезличенной вновь превращается в персональную.

Более того, даже надежное хранение базы данных с идентификаторами вовсе не гарантирует, что после проведения определенного анализа над «обезличенной базой данных» злоумышленник не сможет сделать выводы о принадлежности данных конкретному владельцу.

В качестве примера рассмотрим упрощенную и, тем не менее, вполне реальную ситуацию. Имеется база данных пациентов некоторого лечебного заведения (табл. 1).

Таблица 1

Персональные данные пациентов

Фамилия	Имя	Отчество	Пол	Дата рождения	Паспортные данные	Место работы	Диагноз
Иванов	Глеб	Данилович	м	10.01.79	1202 255469	ООО «АБВ»	Гипертоническая болезнь
Степанова	Елена	Федоровна	ж	06.07.85	1002 103450	АГТУ	Язвенная болезнь желудка
Петров	Денис	Сергеевич	м	24.03.51	2306 090877	Магазин № 35	Ишемическая болезнь сердца
Семенова	Лариса	Дмитриевна	ж	03.11.90	3101 045067	ООО «Успех»	Пневмония
Селезнева	Жанна	Олеговна	ж	29.02.88	0345 789341	Школа № 2	Пиелонефрит

Сведения, хранящиеся в представленной базе данных, критичны, т. к. обрабатывается «специальная категория» персональных данных – сведения о здоровье субъекта. Предположим, что оператор (в данном случае медицинское учреждение) для предоставления отчетности либо для сбора статистики должен передать определенный набор сведений третьему лицу. Для сохранения конфиденциальности критичных сведений принимается решение об удалении из базы данных следующих полей: «Фамилия», «Имя», «Отчество», «Дата рождения» и «Паспортные данные». Другими словами, происходит процедура «обезличивания» (табл. 2).

Таблица 2

«Обезличенная» таблица с персональными данными пациентов

Фамилия	Имя	Отчество	Пол	Дата рождения	Паспортные данные	Место работы	Диагноз
			м			ООО «АБВ»	Гипертоническая болезнь
			ж			АГТУ	Язвенная болезнь желудка
			м			Магазин № 35	Ишемическая болезнь сердца
			ж			ООО «Успех»	Пневмония
			ж			Школа № 2	Пиелонефрит

Возможно, большая часть записей в такой базе действительно будет обезличенной. Однако, если, например, у нас имеется база данных, содержащая некие общедоступные сведения обо всех сотрудниках ООО «АБВ» (табл. 3), то, после несложных умозаключений и при наличии обезличенной базы данных (табл. 2), можно сделать однозначный вывод о том, что Иванов Глеб Данилович страдает гипертонической болезнью.

Таблица 3

Общедоступные данные сотрудников ООО «АБВ»

Фамилия	Имя	Отчество	Должность	...
Сидорова	Мария	Петровна	Директор	
Смирнова	Екатерина	Евгеньевна	Бухгалтер	
Иванов	Глеб	Данилович	Водитель	
Лаврентьева	Светлана	Алексеевна	Экономист	
Никулина	Инна	Олеговна	Менеджер	

Определение k -анонимности

Способ идентификации обезличенных баз данных подобным методом называется «атакой на основе связей» и хорошо известен исследователям, занимающимся данным направлением на Западе [6].

В то время как рассмотренная ситуация представляет собой пример однозначной идентификации, в некоторых случаях связи между данными позволяют выделить группу субъектов, к одному из которых относятся эти данные. Так, исследователи из Миланского университета в своей статье под названием « k -Anonymity» [7] вводят понятие k -анонимность, основное положение которого гласит: «Каждое опубликование информации должно быть таким, что любая комбинация значений идентификаторов может быть неточно соотнесена по меньшей мере с k субъектов». При этом термин «идентификатор» расшифровывается как «набор полей, имеющийся в обезличенной таблице, встречающийся также в других базах данных и вследствие этого уязвимый к «атакам на основе связей».

Для соблюдения k -анонимности необходимо, чтобы каждый идентификатор имел по меньшей мере k вхождений в базу данных. Это требование – достаточное условие для выполнения k -анонимности.

Например, в случае с персональными данными в табл. 1, при любом идентификаторе данная таблица будет удовлетворять лишь 1-анонимности, т. к. каждое поле таблицы содержит хотя бы одно уникальное значение. Это означает, что в худшем случае любое опубликование любой части данных из рассматриваемой таблицы однозначно идентифицирует их владельца.

Обязательное требование, связанное с k -анонимностью, заключается в предварительном определении идентификатора. Идентификатор зависит от внешней информации, доступной злоумышленнику, т. к. это определяет возможность проведения атаки на основе связей (предполагается, что не все возможные открытые таблицы с персональными данными доступны каждому

злоумышленнику) и различные идентификаторы могут потенциально существовать для заданной таблицы. Ради простоты первоначальное условие k -анонимности допускает, что у таблицы с персональными данными есть единственный идентификатор, состоящий из всех полей таблицы, доступных также извне и содержащих не более чем одну запись для каждого субъекта. Следовательно, хотя нахождение правильного идентификатора для таблицы с персональными данными может стать весьма трудной задачей, допускается, что он все-таки найден и определен.

Среди техник, предлагаемых для сохранения анонимности данных при их обезличивании, исследователи [7] особенно выделяют 2 техники: обобщение и сокращение. Первая заключается в замене значений полей БД на более общие (а значит, менее точные) значения. С этой целью вводится понятие «домен». Вторая позаимствована в [5] и может применяться в связке с обобщением для достижения k -анонимности – это удаление записей. Интуитивно название метода говорит о том, что этот дополнительный метод может уменьшить число обобщений, необходимых для выполнения ограничений k -анонимности. Удаление, таким образом, используется для умеренного обобщающего процесса, когда ограниченное число «выбросов» (записей, встречающихся менее k раз) заставит значительно увеличивать число обобщений.

Например, если мы удалим записи, выделенные курсивом в табл. 1, то по идентификатору «Пол» данная таблица будет удовлетворять 3-анонимности.

Кроме названных методов защиты обезличенных данных, есть также такие часто используемые технологии, как дискретизация, замена данных, добавление шума, при которых некоторые основные статистические характеристики набора сохраняются. Однако очень часто требуется, чтобы «правильность данных» сохранялась даже на уровне конкретной записи. Это условие «правильности данных» исключает использование тех техник, которые «портят информацию», даже несмотря на то, что общие статистические характеристики набора сохраняются.

Исследователи k -анонимности представляют обобщающий процесс как некую иерархию – граф, вершинами которого являются значения доменов (обобщенные значения), а корнем – максимально обобщенное значение. Данный граф является взвешенным, причем вес каждого обобщающего шага равен 1, а значит, 2 ближайшие вершины графа отличаются по весу ровно на 1.

Развивая тему k -анонимности, можно высказать идею, что после проведения определенного анализа конкретной базы данных и условий ее функционирования можно задать некий предел обезличивания, который будет приемлем в заданных условиях. Таким образом, необходимо подниматься вверх по обобщающему графу лишь до определенной вершины, вес которой не будет превышать заданный порог обезличивания, т. к. верхние уровни обобщающей иерархии содержат слишком обобщенные, а значит, неточные значения. В то же время вес обобщенных данных должен иметь достаточно большое значение для гарантии сохранения конфиденциальности данных. Возникает задача по определению порога обезличивания с учетом выполнения двух перечисленных и в принципе несовместимых требований.

Более того, возможны ситуации, когда совместное выполнение двух указанных требований может оказаться невозможным, и, следовательно, для обезличивания должны быть применены другие методы. Отметим, что существующая система взвешивания обобщенных графов, связанных с k -анонимностью, является грубой, поскольку не учитывает специфические особенности отдельных данных и групп данных. Следовательно, необходимо произвести дифференциацию всех данных по степени необходимости их обезличивания, в частности дополнить существующую модель обезличивания на основе k -анонимности процедурой взвешивания отдельных данных и групп данных. Указанный вопрос требует отдельной проработки.

Заключение

Принимая во внимание тот факт, что конфиденциальность данных может не обеспечиваться в случае их обезличивания, потенциальный канал утечки информации будет существовать до тех пор, пока процедура обезличивания не будет описана и изучена.

СПИСОК ЛИТЕРАТУРЫ

1. *Федеральный закон «О персональных данных» № 152 от 27.02.2006.*
2. *Петров М. И.* Комментарий к Федеральному закону от 27 июля 2006 г. № 152-ФЗ «О персональных данных». – М.: Юстицинформ, 2007.

3. Горелишвили Д. Постатейный комментарий к проекту Закона России «О персональных данных», http://www.livejournal.com/users/david_gor/.
4. Левинсон Л. Закон о персональных данных. Комментарий эксперта / <http://ikd.ru/Campaign/politrights/Article.2006-11-27.5691>.
5. Samarati P. Protecting respondents' identities in microdata release // IEEE Transactions on Knowledge and Data Engineering. – 2001. – 13 (6). – P. 1010–1027.
6. Anderson R. A security policy model for clinical information systems // In Proc. of the 1996 IEEE Symposium on Security and Privacy Oakland, CA, USA. – 1996. – P. 30–43.
7. *k-Anonymity* / V. Ciriani, S. De Capitani di Vimercati, S. Foresti, P. Samarati // Springer US, Advances in Information Security. – 2007.

Статья поступила в редакцию 17.05.2010

**PROTECTION
OF PERSONAL DATA CONFIDENTIALITY
USING DEPERSONALIZATION**

I. Yu. Kuchin

The aspects, concerning data depersonalization are examined. Personal data protection law connected with the procedures of depersonalization is analyzed. The notion "*k*-anonymity" is considered. The idea of the formation of the requirements to its achievement using weighted graph is suggested.

Key words: Federal Law N 152, personal data, depersonalization, *k*-anonymity.