УДК 342.7:004.056 + 347.15/.18 + 340.64:004.056
ББК Х401:114 + Х401.013

Мищенко Е. Ю., Соколов А. Н.

КОЛИЧЕСТВЕННЫЕ КРИТЕРИИ ИДЕНТИФИКАЦИИ ФИЗИЧЕСКОГО ЛИЦА ПРИ ОБЕЗЛИЧИВАНИИ ПЕРСОНАЛЬНЫХ ДАННЫХ

Результатом обезличивания персональных данных является невозможность идентификации физического лица. Нормативные акты определяют некоторые критерии обезличивания, но все они, как правило, качественные. В статье проанализирована схема идентификации физического лица и дано обоснование применения не только качественных, но и количественных критериев обезличивания. Введены понятия вероятности идентификации и степени обезличивания персональных данных. Показано, что различные атрибуты персональных данных имеют разную значимость при идентификации, а количество атрибутов в группе-идентификаторе растет с ростом объема персональных данных.

Ключевые слова: *персональные данные, обезличивание персональных данных, вероятность идентификации, степень обезличивания.*

Mishchenko E. Y., Sokolov A. N.

QUANTITATIVE CRITERIA OF INDIVIDUAL IDENTIFICATION IN THE PROCESS OF DEPERSONALIZATION OF PERSONAL DATA

The result of depersonalization is the impossibility of the individual identification. Statutory acts define certain criteria of depersonalization, but all of them are generally qualitative. The article describes the scheme of individual identification and proves the application of not only qualitative but also quantitative criteria of depersonalization as well. The terms of identification probability and depersonalization degree are introduced. The article demonstrates that certain attributes of personal data have different effect on identification, and quantity of attributes in group identifier rises with the rise of personal data content.

Keywords: *personal data, depersonalization, identification probability, depersonalization degree.*

В статье «Обезличивание персональных данных: термины и определения»[1] проанализированы различные стороны процесса обработки персональных данных (ПД), их связь с такими субъектами, как Человек (физическое лицо, субъект обработки ПД), Оператор (уполномоченный — орган (или лицо), обрабатывающий или организующий обработку

ПД) и Контролер (федеральный орган исполнительной власти, контролирующий выполнение Закона[2]), рассмотрена терминология процесса обезличивания персональных данных. Показано, что процессы обезличивания и идентификации не могут быть описаны исключительно на качественном уровне.

Целью данной статьи является обоснование применения количественных критериев обезличивания ПД. Чтобы ответить на вопрос, является ли тот или иной набор информации обезличенными ПД, надо подтвердить эффективность обезличивания, то есть доказать, что некий показатель обезличивания изменился после проведения соответствующей процедуры в нужную сторону. Для этого введем понятие *вероятности идентификации* (ВИ) физического лица (ФЛ) в базе ПД — показателя, который в идеальном случае до обезличивания должен быть равен 1, а после обезличивания равен 0.

Значение ВИ = 1 для конкретного ФЛ означает, что его ПД однозначно сопоставлены ему в базе. Однако максимальное значение ВИ = 1 может быть достигнуто только в том случае, если в базе ФЛ нет абсолютно одинаковых, т. е. неразличимых с точки зрения возможностей идентификации. Фактически они могут быть — речь идет о физических близнецах (их около 2 %). И пока вопрос об их полной идентичности (по ДНК, отпечаткам пальцев) не решен, необходимо такую возможность учитывать для любой базы ПД.

Смысл значения ВИ = 0 сложнее для понимания. Фактически оно означает отсутствие возможности сопоставить ПД из базы некоторому ФЛ. При этом возможна ситуация, когда конкретному ФЛ можно сопоставить ПД нескольких «прочих» ФЛ. Чем больше «прочих» ФЛ, тем выше *эффективность обезличивания,* значение ВИ при этом обратно пропорционально количеству «прочих» ФЛ. Это означает, что идеальное значение ВИ = 0 недостижимо. В реальности необходимо требовать выполнения соотношения ВИ < НОРМ, где НОРМ — некое нормативное значение, обратное достаточно большому (тоже нормативному) количеству ФЛ, идентифицированных в обезличенном наборе вместо одного искомого ФЛ. Понятно, что максимальное количество ФЛ — это полное количество ФЛ в базе.

1. Схема идентификации. Нормативного определения термина «идентификация» не существует, поскольку его сущность определяется характеристиками всех сторон, принимающих участие в процессе идентификации. Кроме этого, идентификация решает две задачи: частную (принадлежат ли данные атрибуты заранее определенному ФЛ — «уточнение») и общую (какому именно ФЛ принадлежат данные атрибуты — «поиск»). При этом общая задача решается либо многократным повторением частного решения (перебор небольшого количества вариантов), либо поэтапным сужением области поиска с последующим перебором.

Рассмотрим схему взаимодействия сторон в процессе идентификации (рис. 1). На схеме представлены:

1) область поиска — набор информации, в рамках которого надо идентифицировать Человека (ПД). Для частной задачи — это ПД одного ФЛ, для общей задачи — это ПД некоторого количества ФЛ. Для области поиска как одной из сторон взаимодействия в про-

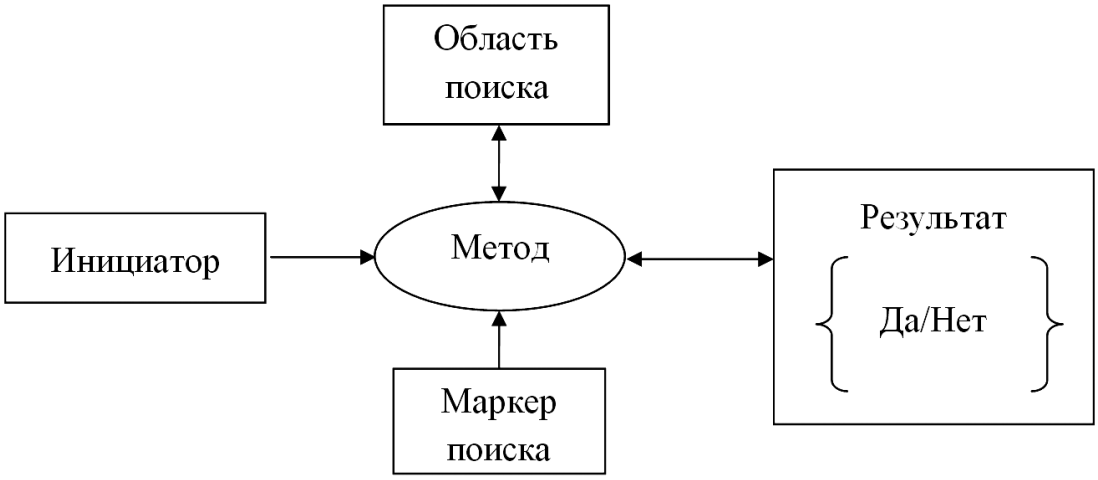

Рис. 1. Схема взаимодействия сторон в процессе идентификации

цессе идентификации можно ввести аналогию — базу данных (БД), где одна запись соответствует одному ФЛ, причем в этой записи заданы значения всех свойств из определенного набора атрибутов, т. е. пустых полей нет;

2) маркер поиска (МП) — набор информации об одном ФЛ (атрибуты неизвестного Человека, которого надо идентифицировать), причем заданы значения всех его атрибутов. Аналогия для МП — одна запись базы данных. МП задает цель поиска, а целью может быть идентификация группы ФЛ (например, все пациенты, больные диабетом), но для простоты в качестве цели поиска мы будем рассматривать ПД одного ФЛ;

3) инициатор процесса — является движущей силой идентификации. Им может быть кто угодно — от контролирующих органов до злоумышленников. У органов власти и силовых структур нет задачи проверки эффективности обезличивания ПД, т. к. они могут получить ПД официально. Проверка требуется только при проведении специальных испытаний, поэтому инициаторами, скорее всего, будут злоумышленники, не имеющие официального доступа к ПД;

4) метод идентификации — любой алгоритм идентификации, определяемый и применяемый инициатором, вне зависимости от достигаемого результата. Для частной задачи Метод представляет собой простое сравнение каждого атрибута, для общей задачи — должен учитывать Результат (двойная стрелка), который можно достигнуть путем нескольких итераций: при отрицательном результате изменяем область поиска (двойная стрелка) и продолжаем работу;

5) результат идентификации — может быть положительный или отрицательный. Для частной задачи он окончательный, для общей задачи — промежуточный, количество этапов при этом зависит от Метода.

2. Взаимодействие Базы Данных и Маркера Поиска. Маркер поиска взаимодействует с Базой Данных в рамках Метода. Идентификация — процесс поиска в БД всех записей о ФЛ, для которых значения всех атрибутов из МП (имеющихся в БД) совпадают с соответствующими значениями из БД. Сравнивать можно только те атрибуты МП, семантика которых совпадает с семантикой атрибутов БД (нет смысла сравнивать имя ФЛ с адресом проживания, но и адрес проживания с адресом деятельности тоже сравнивать бессмысленно). Для описания этого взаимодействия введем следующие количественные критерии:

1. Объем базы (ОБ) — количество записей в базе данных о ФЛ. Чем больше объем, тем меньше ВИ для имеющегося маркера поиска ФЛ (больше вероятность совпадений).

2. Количество атрибутов БД (КБ) — ассортимент свойств ФЛ в базе.

3. Количество атрибутов МП (КМ) — ассортимент свойств ФЛ в маркере. В общем случае КМ не равно КБ.

4. Количество атрибутов поиска (КП) — ассортимент свойств ФЛ, являющихся пересечением ассортимента БД и МП. В идеальном варианте все атрибуты МП входят в состав БД. В случае, когда КП меньше КМ, совокупность атрибутов поиска составляет набор поиска.

5. Атрибут АХ1, АХ2… — любое свойство (характеристика) ФЛ. Здесь и далее через Х обозначена принадлежность атрибута либо базе (АБ1, АБ2…), либо маркеру (АМ1, АМ2…), либо набору поиска (АН1, АН2…), а цифрой — порядковый номер атрибута. Аналогия — поле базы данных.

6. Название атрибута — НХ1, НХ2… в базе данных (НБ1, НБ2…) и в маркере (НМ1, НМ2…) или в наборе поиска (НН1, НН2…) отражает его семантику.

7. Значение атрибута АХ (ЗХ) — некая величина, соответствующая его семантике. Значение удобно обозначить аналогично названию атрибута (ЗХ1, ЗХ2…), а различные значения одного атрибута — ЗХ1-1, ЗХ1-2, ...

8. Диапазон значений атрибута ЗХ1 — множество его значений, имеющее верхнюю (ЗХ1макс) и нижнюю (ЗХ1мин) границы. Значение атрибута может не иметь количественной семантики (семейное положение — холост / женат), в этом случае оно будет задано перечислением (ЗХ1-1, ЗХ1-2, …). Здесь и далее последняя цифра — номер варианта значения. Количество вариантов значений — от одного до полного количества записей ОБ.

9. Количество записей, имеющих атрибут с данным вариантом значения (КЗХ1-1), совпадает с количеством записей, найденных в процессе идентификации по конкретному атрибуту. Если для идентификации используется несколько атрибутов, то данная характеристика будет интегральной, и гораздо удобней использовать обозначение КИ. Нижней границей является КИ = 0 (ни одной записи не найдено). Теоретически это означает, что неправильно выбрана БД — в идеале Инициа-

тор должен быть априори уверен в успехе, но на практике при решении общей задачи это не так. В этом случае идентификация считается неуспешной (и это может быть признано решением задачи). Верхней границей является значение КИ = 1 (найдена ровно одна запись) — идентификация считается успешной. Если найдено несколько записей (КИ > 1) — идентификация считается условной.

10. Вес значения — отношение количества ФЛ, имеющих атрибут с данным вариантом значения, к общему количеству ФЛ в базе (ВЗХ1-1 = КЗХ1-1 / ОБ).

11. *Вероятность идентификации* (ВИ) — величина, обратная количеству найденных записей (ВИ = 1 / КИ). Отметим, что ВИ может не только рассчитываться как интегральная характеристика для всех атрибутов НП, но и определяться для каждого атрибута НП отдельно (ВИ1, ВИ2, …).

12. *Степень обезличивания* (СО) — интегральная характеристика базы ПД, являющаяся дополнением максимальной вероятности идентификации до единицы (СО = 1 — ВИ-макс) для некоторого достаточно большого (нормативного) количества операций идентификации (КСО). Если хотя бы в одном случае ВИ = 1 (успешная идентификация), то СО = 0. Очевидно, что КСО зависит от размера базы ОБ. Для оценки обезличивания также целесообразно ввести соответствующее нормативное значение СОнорм.

Проанализируем количественную зависимость значений ВИ и КИ от всех прочих критериев.

2.1. Зависимость от объема базы ПД. Объем базы (ОБ) определяется ее функциональным назначением и масштабом. От функционального назначения зависит количество записей об одном ФЛ, содержащихся в базе. Масштаб определяется количеством различных ФЛ в базе. Конечно, реальные базы ПД содержат не одну запись о конкретном ФЛ, а несколько (например, БД посещений поликлиники пациентами), но для простоты и ужесточения условий идентификации мы здесь будем рассматривать только БД-справочники, где одному реальному ФЛ соответствует ровно одна запись. Для данного случая ВИмин=1/ОБ и определяющее влияние на ВИ будет иметь масштаб. При определении масштаба надо учесть один нюанс — в базе могут содержаться ПД умерших людей — они тоже должны защищаться в соответствии с законом[2]. Если считать, что ПД умерших ФЛ не удаляются из базы, то значение ОБ увеличивается ежегодно в среднем на 1 % в соответствии с ростом рождаемости.

БД по масштабу можно классифицировать следующим образом:

1. Масштаб всей планеты — в мире живет около 7 миллиардов человек. Минимальная ВИ (для живущих) будет равна 1/7000000000. Это значение хоть и мало, но все-таки больше 0.

2. В масштабе нашей страны (даже с учетом умерших) ОБ можно принять равным 200 млн, в регионе — 10 млн, в районном центре — 100 тыс. и т. д. Поэтому минимальное значение ВИ в базе ПД небольшого предприятия будет, например, 1/100 или 1/20. Возникает вопрос: имеет ли смысл обезличивать такие маленькие базы? Ответом на него будет принятое нормативное значение ВИ.

2.2. Зависимость от количества атрибутов. В общем случае количество атрибутов базы КБ превышает их количество в Маркере КМ. Сравнение количества не имеет смысла, если не установлено соответствие семантики атрибутов БД и МП (при этом не ясно, что и с чем сравнивать). И хотя эта проблема относится не к количеству, а к значениям атрибутов, будем считать, что это соответствие установлено (т. е. количество атрибутов поиска КП определено).

ВИ слабо зависит от количества атрибутов и гораздо сильнее зависит от их значений. При прочих равных условиях чем меньше отношение КМ к КБ, тем меньше ВИ.

2.3. Зависимость от значения атрибута. Значение атрибута является решающим критерием для идентификации. В зависимости от семантики оно может иметь дискретный или непрерывный характер, но фактически, как правило, дискретный (существует определенный шаг, или точность значения, например, целесообразный шаг для роста — 10 см, для веса — 5 кг).

Количество возможных вариантов значений кроме шага определяется еще и диапазоном. Чем больше диапазон значения атрибута, тем больше ВИ с использованием этого атрибута.

На простом примере продемонстрируем упрощенный расчет ВИ (без учета веса значений) по таким атрибутам, как рост (диапазон — от 120 см до 200 см, 9 вариантов значений) и вес (диапазон — от 40 кг до 120 кг, 17 вариантов значений) человека. В нашем случае название атрибутов НН1 = НБ1 = «рост», НН2 = НБ2 = «вес». Ищем ФЛ со значениями ЗН1 =

170 см, ЗН2 = 90 кг. Значения атрибутов в базе ЗБ1-1 = 120 см, …, ЗБ1-9 = 200 см; ЗБ2-1 = 40 кг, …, ЗБ2-17 = 120 кг. В базе объемом ОБ = 1000 при равном весе значений будет найдено: записей ФЛ КЗБ1-6 = 1000 / 9 = 111 (ВИ1 = 1/111); записей ФЛ КЗБ2-11 = 1000 / 17 = 59 (ВИ2 = 1/59). Таким образом, по весу идентифицировать человека проще, чем по росту.

Показанный расчет является упрощенным, поскольку различные варианты значений атрибута встречаются в базе с разной вероятностью (вес значений различен), что оказывает значительное влияние на ВИ. В идеальном случае для определения ВЗ атрибута в БД должна быть рассчитана функция распределения значений по этому атрибуту. Чем меньше ВЗ, тем больше ВИ с использованием данного значения этого атрибута.

Продемонстрируем вышесказанное на следующем примере. Опыт говорит нам, что людей с ростом 170 см гораздо больше (это средний рост взрослого человека — максимум кривой распределения, ВЗБ1-6 = 25%), чем с ростом 200 см (ВЗБ1-9 = 1%). Соответственно, количество найденных в базе записей: КЗБ1-6 = 1000 · 0,25 = 250 (ВИ1-6 = 1/250); КЗБ1-9 = 1000 · 0,01 = 10 (ВИ1-9 = 1/10). Таким образом, ФЛ с ростом 170 см идентифицировать гораздо труднее, чем с ростом 200 см. С другой стороны, можно сделать вывод, что использование атрибута «рост» для идентификации в целом неэффективно. Эффективным будет использование атрибута, имеющего уникальное значение, при этом ВЗБ5-уник = 1 / ОБ, ВИ5-уник = 1. Если у атрибута все варианты значений уникальные, то этот атрибут полностью идентифицирует ФЛ, т. е. является *идентификатором*. Выявить такие атрибуты в базе очень важно, так как процедура идентификации обязательно будет на них опираться, а процедура обезличивания должна их нейтрализовать.

Таким образом, различные атрибуты имеют разную значимость при идентификации. По этому критерию атрибуты можно предварительно разделить на значимые и незначимые. В группу значимых войдут атрибуты, имеющие постоянные значения на протяжении длительных периодов жизни ФЛ (несколько лет, а в идеале — всю жизнь), имеющие дискретные значения из достаточно большого набора. Например, в эту группу не войдут: рост ФЛ (меняется со временем, большой вес значений), семейное положение (малый диапазон значений с большим весом: «холост / женат / разведен / вдовец»). Сложнее обстоят дела с атрибутом «место проживания» — его значение не всегда документально привязано к Человеку и может отличаться от указанного в документах. Вот серьезные кандидаты на включение в группу значимых: ДНК-анализ, отпечаток пальца, фотография (лицо), фамилия, имя, отчество, дата рождения, место рождения.

К сожалению, ни один из указанных значимых атрибутов в отдельности идентификатором быть не может. Принципиально не меняются только дата и место рождения ФЛ, т. е. для данного места ВИ определяется ежедневной рождаемостью. Но даже в небольшом населенном пункте со средней рождаемостью 1 человек в день вероятность рождения 2-х человек в один день достаточно высока, а в городе с населением 1 млн ежедневно рождается около 50 человек. Атрибуты ФИО и адрес хоть и не часто, но меняются, образ лица меняется постоянно, а у анализа ДНК и отпечатка пальца наибольший вес имеет значение «нет данных» (пример: в нашей стране есть федеральная база ДНК преступников, хотя в Исландии она включает всех жителей).

Все остальные атрибуты, входящие в состав ПД, будут либо незначимыми (нужны для целей обработки, но имеют большой вес значений — т. е. дополнительные сведения, ради которых ИСПДн существует, например, профессия ФЛ), либо косвенно значимыми (для целей обработки они не нужны, поэтому в явном виде отсутствуют, но могут оказывать значительное влияние на идентификацию — например, название предприятия отсутствует в базе данных отдела кадров, но фактически сильно ограничивает набор ПД и резко увеличивает ВИ).

Можно использовать в качестве идентификатора совокупность нескольких значимых атрибутов, но сначала рассмотрим еще одну группу — это такие атрибуты, как ИНН и номер паспорта. Мы не включили их в группу значимых, хотя это общепринятые идентификаторы во многих государствах мира. К упомянутым атрибутам можно добавить номер полиса медицинского страхования, водительского удостоверения, телефона, банковского счета… Проблема заключается в том, что все эти атрибуты не имеют прямого отношения к ФЛ, а являются искусственными ведомственными идентификаторами. По первоначальному замыслу все значения этих атрибутов уникальны, но реально это не так. Вот

причины, не позволяющие принять данные атрибуты в качестве идентификаторов:

1. Ограниченное распространение среди физических лиц (многих атрибутов нет у детей в принципе — в отличие от анализа ДНК, который можно взять у любого ФЛ).

2. Изменяемость значений (номера меняются в связи с утерей документов, сменой места жительства, просто по желанию ФЛ, т. е. не являются уникальными в отличие от некоторых значимых атрибутов).

3. Узковедомственная принадлежность (за рамками ведомственных реестров ограниченного доступа эти атрибуты не имеют смысла — попробуйте идентифицировать иностранца по номеру паспорта — т. е. не являются общепризнанными в отличие от значимых атрибутов).

Указанные причины позволяют признать данные атрибуты лишь условно значимыми, или служебными, т. е. каждый из них будет идентификатором ФЛ только при наличии доступа к соответствующему ведомственному реестру. В принципе каждый Оператор может составить свой реестр ПД и присвоить каждому ФЛ уникальный искусственный идентификатор в этом реестре.

Подводя итог, можно сказать, что в качестве идентификатора целесообразно выбирать совокупность значимых атрибутов. Чтобы определить, какие именно атрибуты использовать, необходимо для каждой из возможных групп рассчитать интегральную вероятность идентификации (ВИ). Группа с ВИ = 1 и будет идентификатором.

В качестве примера определим группу идентификаторов ФЛ для города с населением 1 млн человек. Будем принимать в расчет самые трудные варианты. Для атрибута «фамилия» ВИ будет больше, чем 1/100 (с учетом того, что женская фамилия отличается от мужской). Для атрибута «дата рождения» ВИ будет больше, чем 1/50 (365дн · 60 лет / 1 млн). А вот группа из этих двух атрибутов будет иметь ВИ = 1 (это идентификатор). Понятно, что расчет достаточно грубый. Для большей точности необходимо учитывать весовые коэффициенты каждой фамилии и каждой даты рождения в общем наборе информации (объем ПД). Но на конечный результат это вряд ли повлияет.

Не следует забывать и об объеме ПД — для крупного мегаполиса, например с 10 млн жителей, рассмотренная группа атрибутов даст лишь значение ВИ = 1/5. Но если в группу атрибутов добавить либо «имя», либо «инициалы», то этого будет достаточно даже для него.

Если рассматривать объем ПД масштаба страны, в описанную группу атрибутов придется еще добавить, например, «место рождения» и т. д.

Таким образом, количество атрибутов в группе-идентификаторе растет с ростом объема ПД. Подводит данную группу только возможная смена фамилии ФЛ, о которой Инициатор процесса может не знать. В этом случае результат идентификации будет отрицательный (ФЛ не найдено), а ВИ = 0. Поскольку Закон [2] требует подлинности ПД (за это отвечает Оператор) и очевидно, что в рассматриваемом случае база ПД устарела, значение ВИ = 0 должно означать законную (подлинную) смену значения одного из атрибутов группы-идентификатора. Поскольку для «даты рождения» такой процедуры нет, это будет означать, что изменилась «фамилия», и необходимы дополнительные сведения.

Возвращаясь к теме статьи и учитывая сказанное выше, можно сделать вывод, что смена фамилии — один из способов обезличивания ПД Человека, хотя и недостаточно эффективный. Следует отметить, что анализ различных методов обезличивания [3] и методов идентификации (как способа контроля степени обезличивания) выходят за рамки данной работы.

Примечания

[1] Мищенко, Е. Ю. Обезличивание персональных данных: термины и определения / Е. Ю. Мищенко, А. Н. Соколов // Вестник УрФО. Безопасность в информационной сфере. — 2013. — № 1(7) — С. 10—13.

[2] О персональных данных : Федеральный закон Российской Федерации от 27 июля 2006 № 152 (в редакции 2011 года). — http://www.garant.ru.

[3] Об утверждении требований и методов по обезличиванию персональных данных : приказ Роскомнадзора от 5.09.2013 г. № 996. — http://www.garant.ru.

References

[1] Mishchenko E.Yu., Sokolov A.N. Obezlichivanie personal'nykh dannykh: terminy i opredeleniya [Depersonalization of personal data]// Vestnik UrFO. Bezopasnost' v informatsionnoi sfere. — Chelyabinsk: Izd. tsentr YuUrGU Publ., 2013. — No. 1(7) — p.10 – 13.

[2] Federal law of the Russian Federation as of July 27, 2006 No. 152 «On personal data» (editorship as of 2011) [Electronic resource]. URL: http://www.garant.ru

[3] Order of the Federal Supervision Agency for Information Technologies and Communications as of 5.09.2013 No. 996 «On the establishment of regulations and methods on depersonalization of personal data» [Electronic resource]. URL: http://www.garant.ru

Мищенко Евгений Юрьевич, старший преподаватель кафедры безопасности информационных систем ФГБОУ ВПО «Южно-Уральский государственный университет» (национальный исследовательский университет). E-mail: Eug6303@mail.ru

Соколов Александр Николаевич, заведующий кафедрой безопасности информационных систем ФГБОУ ВПО «Южно-Уральский государственный университет» (национальный исследовательский университет). E-mail: ANSokolov@inbox.ru

Evgeny Yurievich Mishchenko, senior lecturer and tutor of the Department of Information System Security of South Ural State University (National Research University). E-mail: Eug6303@mail.ru

Aleksandr Nikolaevich Sokolov, head of the Department of Information system Security of South Ural State University (National Research University). E-mail: ANSokolov@inbox.ru