УДК 004.02

ЗАХИСТ ПЕРСОНАЛЬНИХ ДАНИХ МЕТОДАМИ ЗНЕОСОБЛЕННЯ

Яковенко А.В., Любевич К.А.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Визначено сутність поняття захисту персональних даних, та знеособлення персональних даних. Досліджено законодавчі норми регулювання захисту персональних даних. Описано яким вимоги повинні відповідати методи знеособлення персональних даних. Розглянуто існуючі методи знеособлення персональних даних та проаналізовано їх переваги і недоліки. Запропоновано досконаліший метод знеособлення персональних даних з використанням ETL та хешування.
Ключові слова: персональні дані, захист інформації, знеособлення, захист персональних даних, хешування.

Постановка проблеми. З розвитком процесів інформатизації, створенням нових та інтеграцією існуючих медичних інформаційних систем, орієнтованих на полегшення роботи лікарів та обслуговування населення, все більше уваги приділяється організації обробки персональних даних (ПД).

Широкий вибір технологій дозволяє сучасній людині отримувати актуальну інформацію практично про все. Чим більше персональних даних людини міститься в інформаційному світі, тим більший можливий вплив на життя цієї людини, і відповідно, величина ризику порушення її прав. Активний розвиток відкритих інформаційних систем значно спрощують витік та інші форми незаконного доступу до персональних даних суб'єктів, що робить задачу забезпечення безпеки інформації від зовнішніх і внутрішніх загроз розкрадання, руйнування і/або модифікації особливо актуальною.

Особлива увага приділяється питанням захисту ПД в інформаційних системах. Одним із способів захисту, з точки зору законодавства, є знеособлення, оскільки дозволяє усунути об'єкт атаки.

Відповідно до ст. 2 Закону України «Про захист персональних даних» [1] – знеособлення персональних даних являє собою вилучення даних, що дозволяють ідентифікувати особистість.

Персональні дані – складають важливу частину інформаційного простору, що містить будь-яку інформацію, яка стосується прямо або побічно фізичних осіб – суб'єктів персональних даних [2].

До даних, за якими людина може бути ідентифікована, відносяться всі паспортні дані, а також деяка інша інформація:

– прізвище, ім'я та по батькові;
– вік або дата та місце народження;
– місце проживання;
– ідентифікаційний номер (код);
– соціальний статус;
– пільги відповідно до закону (одинокі матері, жінки з дітьми до трьох років, чорнобильці, неповнолітні, пенсіонери тощо);
– факт звернення по медичну допомогу, отримання медичної допомоги чи медичних послуг особою-пацієнтом, участь у клінічних дослідженнях лікарських засобів та інші.

Виділення таких даних в окрему підмножину обумовлено особливими вимогами до організації їх обробки, пов'язаними з можливістю нанесення шкоди суб'єктам ПД. Тому розвивається і вдосконалюється законодавча база, яка регламентує правила обробки ПД і реалізацію прав громадян на конфіденційність інформації.

Аналіз останніх досліджень і публікацій. Вимоги до захисту в інформаційних системах, відповідно до законодавчих документів, враховують категорію і кількість ПД, специфіку вирішуваних завдань і ряд інших показників. Виконання цих вимог, як правило, пов'язане з істотними матеріальними і фінансовими витратами, викликаними необхідністю створення системи захисту, забезпеченням високої кваліфікації персоналу, отриманням дозвільних документів, що не завжди можливо для великого числа користувачів інформації.

Згідно з вимогами законодавства по знеособленим даними повинна бути відсутньою можливість відновити приналежність персональних даних суб'єкта персональних даних без використання додаткової інформації, до якої вони відносилися до знеособлення [3].

Знеособлення персональних даних повинно забезпечувати не тільки захист даних від несанкціонованого використання, але й можливість їх обробки, тобто данні після знеособлення повинні мати ряд властивостей, до яких відносяться [4]:

· Повнота – збереження всієї інформації про конкретних суб'єктів або груп суб'єктів.

· Структурованість – збереження структурних зв'язків між знеособленими даними конкретного суб'єкта, що відповідають зв'язкам до знеособлення.

· Релевантність – можливість виконання запитів на обробку персональних даних і отримання відповідей в єдиній семантичній формі.

· Семантична цілісність – збереження семантики персональних даних при їх знеособленні.

· Можливість застосування – можливість вирішення задач обробки персональних даних, оброблюваних в інформаційних системах персональних даних, без попереднього дезнеособлення всього обсягу записів про суб'єктів.

· Анонімність – неможливість однозначної ідентифікації суб'єктів даних, отриманих в результаті знеособлення, без застосування додаткової інформації.

Сучасні методи знеособлення персональних даних розвиваються у наступних напрямках:

– зменшення переліку оброблюваних відомостей;
– заміна частини відомостей ідентифікатором/ами;

– заміна чисельних значень мінімальним, середнім або максимальним значенням;

– пониження точності деяких відомостей;

– розмежування інформації та обробки в різних інформаційних системах.

Одним з найбільш актуальних методів знеособлення ПД варто виділити заміну частин ідентифікаторів, а також зниження рівня точності наданих відомостей.

Виділення не вирішених раніше частин загальної проблеми. Багато з перерахованих методів не гарантують неможливість отримання персональної інформації (дезнеособлення) шляхом використання контексту обробки і даних, розміщених в інших системах, які можна пов'язати зі знеособленими, оскільки ці методи, як правило, зберігають зв'язок між різними даними, що відносяться до одного і того ж суб'єкту [5].

Мета статті. Головною метою цієї статті є дослідження методів та алгоритмів знеособлення персональних даних суб'єкта персональних даних.

Виклад основного матеріалу. В роботі [6] запропоновано метод, що базується на хешуванні та декілька алгоритмів знеособлення та реідентифікації суб'єкту ПД в автоматизованих системах.

На рисунку 1 графічно представлений алгоритм знеособлення із застосуванням хеш-функції.

Метод застосування хешування для проведення процедури знеособлення БД, що містить персональні дані, полягає в тому, що в існуючій базі даних визначається, які ідентифікатори дозволяють однозначно визначити суб'єкта персональних даних. Далі здійснюється робота з полями, що містять такі ідентифікатори. Робота виконується циклічно для кожного запису БД. Створюється текстова змінна, яка включає значення однозначних ідентифікаторів, а потім значення і всіх інших ідентифікаторів цього запису. Вираховується хеш-значення даної змінної. Далі відбувається створення окремого файлу, куди послідовно переносяться всі однозначні ідентифікатори, при цьому кожна комірка, дані з якої будуть перенесені, або обнуляється або на вимогу користувача – накладається маска. Все це виконується та повторюється для всіх записів бази даних.

Із недоліків даного алгоритму наведеного в роботі [6] можна виділити наступні:

– запис значень однозначних ідентифікаторів на зовнішній носій – однозначні ідентифікатори повинні зберігатися тільки в захищеному середовищі та доступні для модифікації тільки авторизованим особам;

– проведення процедури хешування в тому ж місці де знаходяться вихідні дані для знеособлення.

Метод введення ідентифікаторів, що представлений у роботі [7] полягає в заміні частини відомостей (значень персональних даних) ідентифікаторами зі створенням таблиці (довідника) відповідності ідентифікаторів вихідними даними. Тобто після застосування даного методу єдина база (БД) розпадається на дві бази:

1) таблиця відповідності, в якій деякий набір ідентифікуючих фізичну особу (ФО) атрибутів однозначно зіставляється з деяким абстрактним атрибутом. Тобто для кожної ФО набір значущих атрибутів ідентифікації відповідає деякому службовому унікальному ідентифікатору. Причому обсяг цієї бази відповідає кількості ФО;

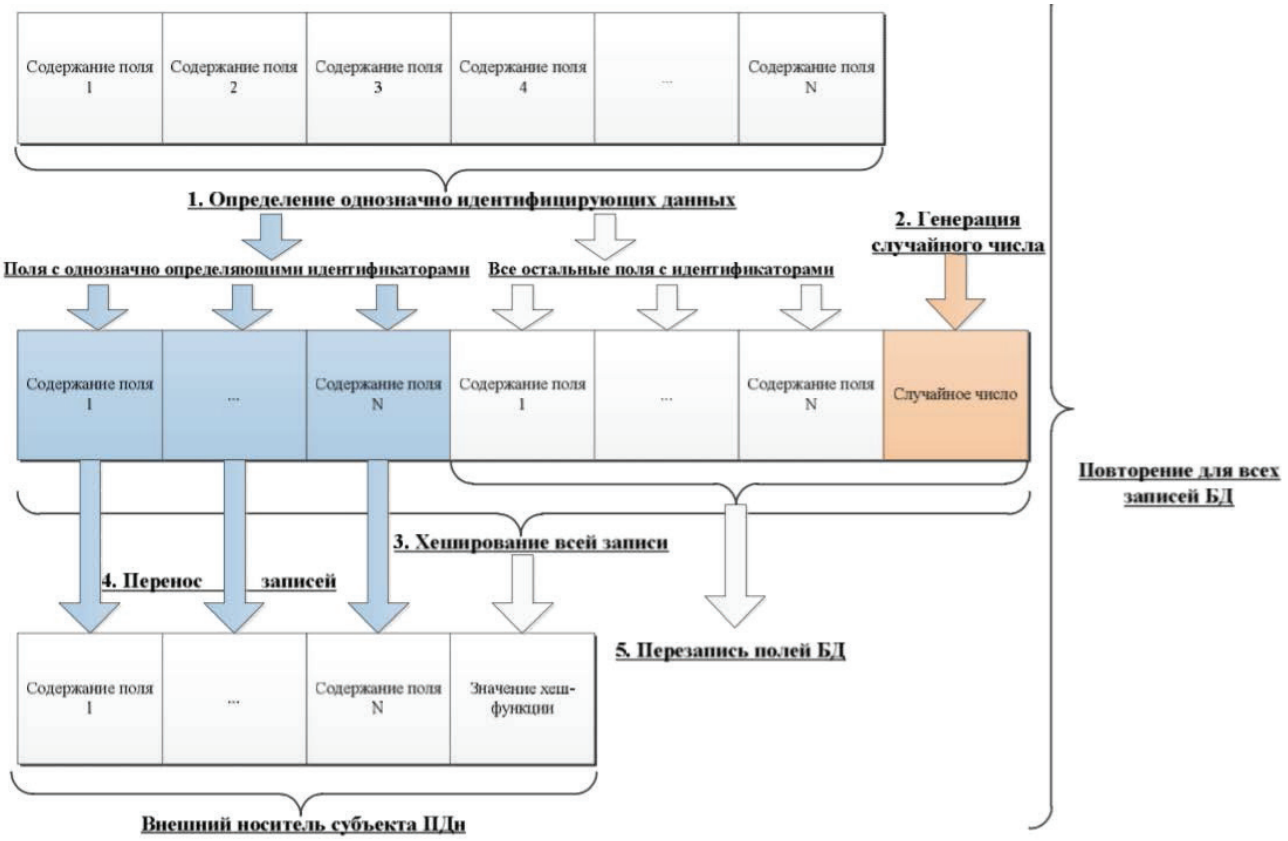

Рис. 1. Алгоритм знеособлення із застосуванням хеш-функції

Джерело: [6]

2) база інших даних, в якій певному службовому ідентифікатору (не унікальному, так як зустрічається в цій базі кілька разів) однозначно зіставляється набір інших даних – не значимі з точки зору ідентифікації, але визначають суть обробки. Причому обсяг цієї бази може набагато перевищувати значення попередньої бази [7].

Для прикладу, наведено шістьох осіб зі зміненими ПД та проведено знеособлення за методом введення ідентифікаторів (таблиця 1).

Таблиця 1

Прізвище	Місто	Вік	Стать	Діагноз
Іваненко	Київ	23	Ж	Ішемічна Хвороба
Петренко	Черкаси	25	Ч	Міокардит
Шевченко	Вінниця	27	Ж	Пролапс мітрального клапана
Франко	Київ	26	Ч	Стеноз гирла аорти
Чубенко	Одеса	29	Ч	Міокардит

Джерело: розроблено авторами

Ввівши ідентифікатор для атрибутів "Діагноз", "Місто" та створивши додаткові таблиці-довідники, дані будуть мати наступний вигляд:

Таблиця 2

Прізвище	Вік	Стать	Місто	Діагноз
Іваненко	23	Ж	1	1
Петренко	25	Ч	2	2
Шевченко	27	Ж	3	3
Франко	26	Ч	1	4
Чубенко	29	Ч	4	2

Джерело: розроблено авторами

Таблиця 3

Ідентифікатор	Місто
1	Київ
2	Черкаси
3	Вінниця
4	Одеса

Джерело: розроблено авторами

Таблиця 4

Ідентифікатор	Діагноз
1	Ішемічна Хвороба
2	Міокардит
3	Пролапс мітрального клапана
4	Стеноз гирла аорти

Джерело: розроблено авторами

Із опису реалізації даного алгоритму наведеного у роботі [7] недоліками є:

– не забезпечення властивості релевантності, так як в запиті і відповіді на запит змінюється вид атрибутів персональних даних, які були замінені ідентифікаторами;

– зберігання в записах зв'язок між атрибутами знеособлених даних;

– алгоритм доцільно застосовувати при невеликій кількості атрибутів і невеликому обсязі масиву персональних даних.

В основу методу перемішування, представленого у роботі [8] покладено перестановку окремих значень або груп значень атрибутів персональних даних в масиві ПД. Значення кожного атрибуту нумеруються і розбиваються на підмножини. Усередині кожної i-ї підмножини дані переміщуються шляхом циклічної перестановки (зсуву) на R_i позицій. Потім підмножини кожного j-го атрибуту також перемішуються шляхом циклічного зсуву на C_j.

Проте метод, наведений у роботі [8] має декілька недоліків:

– дані фактично знаходяться в тому ж місці і не змінюють свого складу, що при невеликій кількості атрибутів дозволить легко провести дезнеособлення;

– не виконуються властивості релевантності, структурованості, можливості застосування;

– стійкість до атак збільшується зі збільшенням обсягу масиву персональних даних.

Аналіз недоліків, що було виділено в розглянутих методах, дозволив розробити наступний підхід рис. 2.

Для організації збереження ПД доцільним стало використання підходу ETL (Extraction-Transformation-Load) [9] із трьома рівнями розділення відповідальності.

ETL – це комплекс методів, що реалізують процес перенесення вихідних даних з різних джерел в аналітичний додаток або сховище даних, що його підтримує.

· Вихідні дані розташовані в джерелах найрізноманітніших типів і форматів, створених в різних додатках, в той час як для вирішення завдань аналізу дані повинні бути перетворені в єдиний універсальний формат, який підтримується сховищем даних та аналітичним додатком.

· Дані в джерелах зазвичай надмірно деталізовані, тоді як для вирішення завдань аналізу в більшості випадків потрібні узагальнені дані.

· Вихідні дані, як правило містять різні фактори, які заважають їх коректному аналізу.

В якості алгоритму знеособлення можливим стало використання підходу з хешуванням, що описаний у роботі [6] з деяким доопрацюванням. Модифікація полягає у хешуванні цілісного кортежу ПД для кожного окремого запису в БД із розсекреченням тільки тих атрибутів, які однозначно не визначають приналежність даних певній фізичній особі. Таким чином у відкритий доступ попадуть тільки необхідні дані, проте залишається можливість дезнеособлення по обчисленому хеш-значенню.

Захищена база даних зберігає необроблені персональні дані користувачів (рис. 2). Клієнтом може бути будь-хто (оператор, інші сервери, що потребують дані для своїх потреб і т.д.), хто може запитувати персональні дані і отримувати їх в знеособленому вигляді. Сервер анонімізації виконує функцію проміжної ланки між клієнтом та захищеною БД. Він обробляє та знеособлює ПД перед тим як віддати їх для використання.

Висновки і пропозиції. Аналіз методів та алгоритмів знеособлення ПД дозволив виявити, що багато з перерахованих методів не гарантують неможливість отримання персональної інформації, оскільки ці методи, як правило, зберігають зв'язок між різними даними, що відносяться до одного і того ж суб'єкту.

Запропоновано підхід до вирішення проблеми знеособлення ПД, що дозволяє ефективно та прозоро виконати знеособлення персональних даних та виділити збереження персональних даних в окреме, захищене середовище.

Рис. 2. Схема зберігання та обробки персональних даних

Джерело: розроблено авторами

Список літератури:

1. Закон України «Про захист персональних даних» (2297-17 Редакція від 30.01.2018).
2. Захист персональних даних: Правове регулювання та практичні аспекти / М.В. Бем, І.М. Городиський, Г. Саттон, О.М. Родіоненко // Правове регулювання та практичні аспекти: науково-практичний посібник / М.В. Бем, І.М. Городиський, Г. Саттон, О.М. Родіоненко. – К: К.І.С., 2015. – С. 220.
3. McCallister E. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) / E. McCallister, T. Grance, K. Scarfone // Recomendations of the National Institute of Standards and Technology (NIST) U.S. – 2010. – С. 59.
4. Трифонова Ю.В. Возможности обезличивания персональных данных в системах, использующих реляционные базы данных / Ю.В. Трифонова, Р.Ф. Жаринов // Доклады ТУСУРа. – 2014. – № 2. – С. 188–194.
5. Саксонов Е.А. Процедура обезличивания персональных данных [Електронний ресурс] / Е.А. Саксонов, Р.В. Шередин. – 2011. – Режим доступу до ресурсу: http://technomag.edu.ru/jour.
6. Волокитина Е.С. Метод и алгоритмы гарантированного обезличивания и реидентификации субъекта персональных данных в автоматизированных информационных системах : автореф. дис. на здобуття наук. ступеня канд. техн. наук : спец. 05.13.19 "Методы и системы защиты информации, информационная безопасность" / Волокитина Евгения Сергеевна. – Санкт-Петербург, 2013. – 24 с.
7. Мищенко Е.Ю. Количественный анализ процедуры обезличивания персональных данных. Метод введения идентификаторов / Е.Ю. Мищенко, А.Н. Соколов // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2015. – № 3. – С. 18–25.
8. Карпова И.П. О реализации метода обезличивания персональных данных / И.П. Карпова // Вестник компьютерных и информационных технологий. – 2013. – № 6. – С. 56–60.
9. Lane A. Understanding and Selecting Data Masking: How It Works [Електронний ресурс] / Adrian Lane. – 2012. – Режим доступу до ресурсу: https://securosis.com/blog/understanding-and-selecting-data-masking-how-it-works.

Яковенко А.В., Любевич К.А.
Национальный технический университет Украины
«Киевский политехнический институт имени Игоря Сикорского»

ЗАЩИТА ПЕРСОНАЛЬНЫХ ДАННЫХ МЕТОДАМИ ОБЕЗЛИЧИВАНИЯ

Аннотация
Определены сущность понятия защиты персональных данных, и обезличивания персональных данных. Исследованы нормы регулирования защиты персональных данных. Описаны которым требования должны соответствовать методы обезличивания персональных данных. Рассмотрены существующие методы обезличивания персональных данных и проанализированы их преимущества и недостатки. Предложено более совершенный метод обезличивания персональных данных с использованием ETL и хеширования.
Ключевые слова: персональные данные, защита информации, обезличивание, защита персональных данных, хеширования.

Yakovenko A.V., Lyubevich K.A.
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"

PROTECTION OF PERSONAL DATA USING DEPERSONALIZATION

Summary
Determined the essence of the notion of protection of personal data, and depersonalization of personal data. The legislative norms of personal data protection regulation are investigated. Described what requirements should meet the methods of personal data depersonalization. Analyzed existing methods of depersonalization of personal data and their advantages and disadvantages. Proposed more advanced method of depersonalization of personal data using ETL and hashing.
Keywords: personal data, protection of information, depersonalization, protection of personal data, hashing.