INFORMATION TECHNOLOGY

DOI 10.20535/2411-1031.2019.7.2.190539
УДК 004.75

ЯРОСЛАВ ДОРОГИЙ

ТЕХНОЛОГІЯ ПОШУКУ СУМНІВНИХ ЗАПИСІВ ПРИ СТВОРЕННІ ЄДИНОГО РЕЄСТРУ ІДЕНТИФІКАЦІЇ ФІЗИЧНИХ ОСІБ УКРАЇНИ

Одним з найефективніших рішень для захисту персональних даних при побудові Єдиного реєстру ідентифікації фізичних осіб є спільне використання наскрізного ідентифікатора та хеш-кодів, які генеруються з комбінацій персональних даних за допомогою односторонніх хеш-функцій. Це пов'язано з тим, що етап створення єдиного реєстру ідентифікації фізичних осіб не передбачає використання відкритих персональних даних і тому на сервері не дозволяється зберігати персональні дані та допускається лише використання унікальних ідентифікаторів і хеш-кодів. Відповідно до принципів створення вищевказаного реєстру, проаналізовані та використані для генерування хеш-кодів п'ять обов'язкових та п'ятнадцять опціональних типів персональних даних, які зберігаються в реєстрі, а також можливі комбінації полів персональних даних (в роботі використано десять різних комбінацій персональних даних), побудованих на зазначених типах. Розроблено технологію наскрізної ідентифікації особи, яка має можливість відслідковувати помилки в полях з персональними даними при введенні нових даних та при шуканні в реєстрі. Для оцінки запропонованої технології відібрано 100 000 модельованих осіб з випадковими помилками в відповідних полях, що зберігають персональні дані. Ці помилки випадковим чином поміщені в поля бази даних створюваного реєстру, які зберігають персональні дані обов'язкового та опціональних типів. Працездатність запропонованої технології також перевірено шляхом реєстрації нових осіб у реєстрі. Запропонована технологія має високу толерантність до помилок і може правильно ідентифікувати та асоціювати особу, навіть з помилками в декількох полях з персональними даними. Правильні персональні дані, особливо в полях бази даних з персональними даними обов'язкового типу, мають вирішальне значення для уникнення помилкових записів в створюваному реєстрі. У контексті одностороннього хеш-перетворення сумнівний запис з персональними даними може бути ідентифікований шляхом застосування операторів теорії множин на основі хеш-кодів, розрахованих відповідно до визначених комбінацій персональних даних.

Ключові слова: точність даних; персональні дані; приватність; наскрізний ідентифікатор; реєстр; конфіденційність.

Постановка проблеми. Наразі в Україні немає технології наскрізної ідентифікації особи і, відповідно, єдиного ідентифікатора особи, тобто відсутній реєстр, який міг би стати основою для створення умов реалізації принципу "once only" (принципу, за яким первинна реєстрація особи та внесення інформації щодо неї в державні інформаційні ресурси виконується одноразово). Фактично, нині держава (через реєстри декількох державних органів влади) володіє інформацією стосовно фізичних осіб, але жодний з наявних реєстрів не здатний забезпечити достовірну ідентифікацію всіх фізичних осіб. У зв'язку з цим на сьогодні склалася вкрай небезпечна ситуація через масове дублювання інформації про особу в інформаційних ресурсах як органів державної влади, так і в інформаційних ресурсах банківської сфери та комерційного сектору. Як наслідок, надання більшості адміністративних послуг здійснюється з порушенням частини сьомої статті 9 Закону України "Про адміністративні послуги" [1].

У зв'язку із цим, в умовах євроінтеграційних процесів і активної розбудови електронного урядування в Україні перед органами державної влади виникла нагальна потреба в однозначній ідентифікації особи. На цей час найбільш повними реєстрами в частині обробки персональних даних є:

1. Державний реєстр виборців Центральної виборчої комісії (не містить відомостей щодо осіб, які не мають права обирати, – щодо дітей, іноземців, інших категорій осіб).
2. Державний реєстр актів цивільного стану (ДРАЦС) Міністерства юстиції України (особа ідентифікується за декількома полями реєстру).
3. Державний реєстр фізичних осіб Державної фіскальної служби України (ідентифікатором є індивідуальний податковий номер (ІПН), але реєстр не може забезпечити ідентифікацію всіх громадян, оскільки не всі громадяни мають ІПН – діти, громадяни з релігійними переконаннями).
4. Єдиний державний демографічний реєстр (ЄДДР) Державної міграційної служби України (містить близько 12 млн записів).
5. Реєстр застрахованих осіб Державного реєстру загальнообов'язкового державного соціального страхування Пенсійного фонду України (ПФУ).
6. Реєстр пацієнтів електронної системи охорони здоров'я Міністерства охорони здоров'я України.
7. Єдина державна електронна база з питань освіти Міністерства освіти і науки України.
8. Інформаційна база стосовно віз для в'їзду в Україну і транзитного проїзду через її територію (Е-Віза) Міністерства закордонних справ України.

Для створення технології наскрізної ідентифікації та приведення інформації про особу у відповідність до єдиного ідентифікатора, який відповідатиме формату Унікального номера запису реєстру (УНЗР), потрібно проаналізувати наявну в основних державних інформаційних ресурсах інформацію про особу для створення Єдиного реєстру ідентифікації фізичних осіб (ЄРІФО), уникнення дублювання та виправлення помилок, що виявляться під час аналізу. Створений ЄРІФО стане основою технології первинної наскрізної ідентифікації особи в інформаційних ресурсах – публічного сервісу, побудованого з використанням ідентифікаторів ЄРІФО, з повною прив'язкою до унікального номеру запису реєстру (УНЗР) ЄДДР за наявності і надалі полегшить ідентифікацію особи в момент надання послуг з документування згідно із частиною першою статті 4 Закону України "Про Єдиний державний демографічний реєстр та документи, що підтверджують громадянство України, посвідчують особу чи її спеціальний статус" [2].

Аналіз останніх досліджень і публікацій. Найчастіше, при створенні нового реєстру, дані про особу збираються з декількох реєстрів. У випадку створення ЄРІФО будуть використані реєстри, перелік яких наведено при постановці проблеми. При цьому пропонується розглядати ЄДДР як еталонний реєстр, відомості до якого і надалі можуть вноситися виключно відповідальними особами, а ідентифікатор ЄРІФО використовуватиметься наскрізно, з можливістю створення умов довіреного внесення інформації іншими державними органами влади (наприклад, представниками Міністерства охорони здоров'я України або ДРАЦС при народженні дитини) шляхом використання Системи електронної взаємодії державних електронних інформаційних ресурсів "Трембіта" (далі – СЕВ ДЕІР) [3] або інших засобів електронної взаємодії.

Персональні дані (ПД) з реєстрів використовуються для ідентифікації та агрегування різних типів даних щодо особи (наприклад, біометричних та генетичних даних, клінічних даних, інформації про рухоме та нерухоме майно). Такі дані зазвичай включають в себе технологічний ідентифікатор (наприклад, ідентифікатор пацієнта, номер соціального страхування або індивідуальний податковий номер), прізвище, ім'я та по батькові (ПІБ), дату народження, місце народження, адресу, поштовий індекс [4]. Однак, обмін ПД може

призвести до розкриття особистої інформації і тому обмежений різними нормативно-правовими актами. Тому дослідження методів обміну ПД направлено на забезпечення цілісності, доступності та конфіденційності даних і є дуже важливим завданням при побудові ЄРІФО.

Існують різні способи захисту ПД особи в реєстрах. Серед них доцільно виділити деідентифікацію [5] - [7], деперсоніфікацію [8], анонімізацію даних [9], [10], використання обмеженого набору даних [11] і хеш-перетворення [12], [13]. Використання технології наскрізної ідентифікації є одним з найефективніших методів захисту ПД особи.

Метою статті є аналіз проблематики створення технології наскрізної ідентифікації фізичних осіб та її подальшого використання при побудові Єдиного реєстру ідентифікації фізичних осіб України.

Виклад основного матеріалу дослідження. Створювана технологія наскрізної ідентифікації передбачає трансформування комбінації схем полів з ПД в хеш-коди за допомогою одностороннього алгоритму хешування. Вона може бути використана для ідентифікації особи в реєстрах, без передачі будь-якої частини ПД. Декілька полів з ПД можуть об'єднуватися в різні шаблони з метою полегшення пошуку в різних реєстрах. Технологія наскрізної ідентифікації передбачає, що передача ПД до ЄРІФО відбувається у відкритому або хешованому вигляді, тобто ПД фактично можуть не передаватися до реєстру у відкритому вигляді та зберігаються в реєстрах-донорах. Така ситуація (при використанні варіанту з передачею хешованих даних) буде зберігатися до введення в дію ЄРІФО. З введенням в дію ЄРІФО, хешована інформація в реєстрі трансформується до відкритого стану, ПД в реєстрах-донорах будуть видалені внаслідок нормативної вимоги та на виконання принципу одноразового внесення даних "once only" та буде нормативно узгоджена процедура наповнення ЄРІФО новими даними з реєстрів-донорів державних установ України, які виконують первинну реєстрацію особи (народження дитини, смерть особи, в'їзд іноземного громадянина на територію України).

З метою належного функціонування технології наскрізної ідентифікації та ЄРІФО, ПД мають бути попередньо зібрані та відфільтровані. Як зазначено вище, еталонним реєстром обрано Єдиний державний демографічний реєстр України та, відповідно, УНЗР обрано форматом наскрізного ідентифікатора для інформаційної взаємодії. Для використання в інформаційних ресурсах та інформаційних системах також буде згенерований унікальний технологічний ідентифікатор для кожного верифікованого запису ЄРІФО. Основною функцією використання еталонного реєстру в цій технології є остаточна верифікація особи, тобто якщо дані про певну особу співпадають з даними в ЄДДР, то особа вважається верифікованою і всі персональні дані про особу в усіх реєстрах повинні бути видалені та, відповідно, внесено інформацію, про наскрізний ідентифікатор ЄРІФО (публічна частина – УНЗР, технологічна – згенерований унікальний ідентифікатор), що надасть реєстрам можливість в будь-який момент отримати персональну інформацію в межах компетенції та повноважень.

Первинне формування ЄРІФО. Первинне формування записів ЄРІФО відбувається з використанням 20 полів ПД для ідентифікації особи, включаючи 5 обов'язкових полів БД і 15 опціональних полів (див. табл. 1). Як правило, вони є унікальними для особи і не змінюються протягом життя. Кожне поле ПД має свою апроксимовану ймовірність ідентифікування двох різних осіб. Це означає, що 2 різні особи можуть бути випадковим чином ідентифіковані в межах популяції, і які мають такі ж значення для цього поля ПД.

Кожне поле ПД програмно нормалізується з метою їх приведення до уніфікованого формату (тільки великі літери і цифри, без пробілів, знаків пунктуації). Всі поля нормалізованої база даних хешуються за допомогою алгоритму SHA-512. Саме цей тип хешування обрано через його поширеність та наявність підтримки у сучасних СКБД. Саме за такою процедурою пропонується знеособити всі вищезазначені державні реєстри з метою їх

подальшої передачі до розробника ЄРІФО та використання для первинного його наповнення у разі застосування хешованих ПД. Наступним кроком передбачається поступове внесення всіх записів зі знеособлених або відкритих реєстрів в ЄРІФО шляхом додаткової обробки кожного запису. Така обробка передбачає створення додаткової таблиці БД, що буде містити записи для кожного суб'єкта реєстру у вигляді додаткових полів з хеш-кодами від шаблонів комбінацій персональних даних, вже попередньо оброблених та представлених у вигляді хеш-кодів (див. табл. 2). Фактично, у разі використання знеособлених реєстрів, для зведення записів реєстрів в єдину базу даних пропонується застосовувати подвійне хешування.

Таблиця 1 – Поля ПД

Тип	**Ім'я**	**Значення**
Обов'язковий	Family (F)	Прізвище
	Name (N)	Ім'я
	Patronymic (PT)	По батькові
	Address (AD)	Місце народження
	DateOfBirth (DOB)	Дата народження
Опціональний	Sex (SX)	Стать
	Country (CNTR)	Країна видачі документа особи
	DocSeries (DS)	Серія документа особи
	DocNumber (DN)	Номер документа особи
	DocDataGet (DDG)	Дата видачі документа особи
	DocExpDate (DED)	Дата закінчення дії документа особи
	DocIssued (DI)	Ким видано документа особи
	IPN	Індивідуальний податковий номер
	SocNum (SN)	Номер соціального страхування
	MotherFN (MFN)	Ім'я матері особи
	MotherLN (MLN)	Прізвище матері особи
	MotherDateOfB (MDOB)	Дата народження матері особи
	FatherFN (FFN)	Ім'я батька особи
	FatherLN (FLN)	Прізвище батька особи
	FatherDateOfB	Дата народження батька особи

Кожен шаблон комбінації перетворюється в 64-байтовий хеш-код за допомогою одностороннього алгоритму хешування SHA-512. До кожного результуючого коду додається додатковий байт, який вказує кількість відсутніх полів ПД для хеш-коду. Кожна комбінація є достатньою для того, щоб впевнено розрізняти особу. Далі, генерується випадковий унікальний код UID та тимчасовий УНЗР і зв'язується з хеш-кодами комбінацій. UID і його зв'язані хеш-коди зберігаються на сервері і використовуються для визначення особи.

*Правило співставлення хеш-кодів та особи***.** Кожний хеш-код складається з 64-байтового хеш-значення, яке обчислюється з шаблону комбінації ПД за допомогою одностороннього алгоритму хешування та додаванням 1 додаткового байту, яке вміщує кількість пропущених полів ПД у хеш-коді. Таким чином, будь-яка помилка в полі ПД, що використовуються в комбінації, призведе до помилки співставлення хеш-коду.

Таблиця 2 – Шаблони комбінацій ПД

Хеш-код	Шаблони комбінацій
1	Family + Name + Patronymic + Sex^{o} + Country^{o} + IPN^{o}
2	Family + Name + Patronymic + Address + IPN^{o}
3	Family + Name + Patronymic + DateOfBirth + SocNum^{o}
4	Family + Name + Patronymic + Address + IPN^{o} + SocNum^{o}
5	Family + DateOfBirth + DocSeries^{o} + DocNumber^{o} + DocExpDate^{o}
6	Family + Name + Patronymic + Sex^{o} + IPN^{o}
7	Family + Name + Patronymic + DocSeries^{o} + DocNumber^{o} + SocNum^{o}
8	Family + Name + Patronymic + FatherFN^{o} + FatherLN^{o} + FatherDateOfB^{o}
9	Family + Name + MotherFN^{o} + MotherLN^{o} + MotherDateOfB^{o}
10	Family + Name + Address + MotherFN^{o} + MotherLN^{o} + IPN^{o}

^{o} – поле, яке НЕ є обов’язковим.

Пропонується використовувати 3 типи хеш-кодів: ідеальний, хороший і поганий. Для кожного хеш-коду використовуються 2 параметри для визначення його типу: нижній поріг (L) і верхній поріг (U) (див. табл. 3). Для ідеального хеш-коду вимагається, щоб кількість пропущених полів ПД дорівнювала або була меншою L. Кількість відсутніх полів ПД для генерації хорошого хеш-коду обмежується інтервалом (L, U). Якщо кількість пропущених полів ПД більше U, то хеш-код буде визначений як поганий. Співпадіння 2 ідеальних хеш-кодів для двох записів БД свідчить про те, що вони ймовірно належать одному і тому ж суб’єкту, а співпадіння між двома хорошими хеш-кодами для двох записів визначають ці хеш-коди ймовірними кандидатами, що належать одному і тому ж суб’єкту.

Таблиця 3 – Пороги відсутніх полів для визначення типу хеш-коду

Параметри	Хеш-код 1	Хеш-код 2	Хеш-код 3	Хеш-код 4	Хеш-код 5	Хеш-код 6	Хеш-код 7	Хеш-код 8	Хеш-код 9	Хеш-код 10
Нижній поріг	1	0	0	1	1	1	1	1	1	1
Верхній поріг	2	1	1	2	2	2	2	2	2	2

При введенні нових записів до реєстру автоматично підраховується кількість ідеальних та хороших співпадінь. Це надасть можливість з’ясувати, чи є вже дана особа у реєстрі. Для

цього використовуються 3 параметри для визначення відповідності особи: поріг для ідеального співпадіння (*P*), поріг для хорошого співпадіння (*G*) і поріг для змішаного співпадіння (*X*). Два записи збігаються один з одним, якщо кількість ідеальних співпадінь більша або рівна P, або кількість хороших співпадінь більша або рівна G, або сума ідеальних і хороших співпадінь більша або рівна X. Для створення ЄРІФО встановлено такі порогові значення $P = 2$, $G = 3$, $X = 3$.

Імовірні шаблони комбінації ПД для ідеальних або хороших хеш-кодів. Хеш-коди генеруються з комбінацій полів ПД в реєстрі, тому кожен з них може розглядатися як набір перетворених полів ПД. Крім того, у різних хеш-кодах є поля, що перекриваються. Завдяки цьому можливе використання теорії множин для систематичної перевірки сумнівних полів ПД. Доки хеш-код узгоджується, його відповідні поля з ПД можуть бути виключені операціями над множинами з пошуку сумнівних полів ПД. Оскільки дозволено мати відсутні значення опціональних полів ПД, то спочатку необхідно проаналізувати всі ймовірні схеми поєднання полів ПД для ідеальних та хороших хеш-кодів, після чого можна розробити алгоритм перевірки сумнівних ПД.

Особа ідентифікується тільки за допомогою ідеальних або хороших хеш-кодів. Відсутні поля ПД можуть впливати на збіг хеш-коду. Під час реєстрації особи, при розгляді відсутніх полів, можна уникнути деяких неправильних співставлень. Наприклад, хеш-код 8 з табл. 2 генерується з комбінації обов'язкових полів Family, Name, Patronymic і опціональних полів FatherFN, FatherLN і FatherDateOfB. Нехай особа зареєстрована вперше, поле FatherDateOfB пропущено, а інші поля введені правильно – в результаті згенеровано хеш-код 8^0. Далі, коли особа повторно зареєстрована в іншому реєстрі, надані правильні значення всіх вищевказаних полів ПД, включаючи FatherDateOfB, відповідно, згенеровано хеш-код 8^1. Оскільки поле FatherDateOfB пропущено при генерації хеш-коду 8^0, то хеш-код 8^1 не відповідатиме хеш-коду 8^0. Однак, існує ідеальна відповідність між хеш-кодом 8^1 і хеш-кодом 8^0. Якщо поле FatherDateOfB вважати відсутнім полем для генерування хеш-коду 8^1, то він буде ідеально зіставлятися з попереднім хеш-кодом 8^0 і таким чином, це надасть можливість уникнути неправильного порівняння хеш-кодів 8. Отже, всі ідеальні або хороші хеш-коди осіб, які зареєстровані в реєстрах, повинні бути ретельно проаналізовані на рахунок правильного зіставлення.

Кожен хеш-код генерується з різних шаблонів комбінацій полів ПД, які є опціональними або необхідними. На основі розроблених шаблонів комбінацій ПД можуть бути проаналізовані та ідентифіковані всі ймовірні ідеальні або хороші хеш-коди, визначені правила співставлення хеш-кодів (див. табл. 4). Наприклад, хеш-код 5 генерується з комбінації з полів Family, DateOfBirth, DocSeries, DocNumber і DocExpDate. З них поля Family і DateOfBirth є обов'язковими полями, а інші 3 поля є опціональними. Відповідно до правил зіставлення хеш-кодів, ідеальний хеш-код 5 може мати 1 відсутнє поле та хороший хеш-код 5 може мати 2 відсутні поля. Тобто ідеальний хеш-код 5 може містити 1 відсутнє поле з DocSeries, DocNumber або DocExpDate, і хороший хеш-код 5 може використовувати тільки одне з цих полів ПД. Отже, є 4 варіанти ймовірних ідеальних і 3 ймовірних хороших хеш-коди 5.

Таблиця 4 – Імовірні шаблони комбінації ПД для ідеальних або хороших хеш-кодів

Індекс	**Хеш-код**	**Комбінації полів з ПД**						**Відсутні поля**	**Тип хеш-коду**
1	1	F	N	PT	SX^o	$CNTR^o$	IPN^o	-	Ід
2		F	N	PT	SX^o	$CNTR^o$	-	IPN^o	Ід
3		F	N	PT	-	$CNTR^o$	IPN^o	SX^o	Ід
4		F	N	PT	SX^o	-	IPN^o	$CNTR^o$	Ід

Продовження таблиці 4

Індекс	**Хеш-код**	**Комбінації полів з ПД**						**Відсутні поля**	**Тип хеш-коду**
5		F	N	PT	SX^o	-	-	IPN^o CNTR^o	Хор
6		F	N	PT	-	CNTR^o	-	IPN^o SX^o	Хор
7		F	N	PT	-	-	IPN^o	CNTR^o SX^o	Хор
8	2	F	N	PT	AD	IPN^o	-	-	Ід
9		F	N	PT	AD	-	-	IPN^o	Хор
10	3	F	N	PT	DOB	SN^o	-	-	Ід
11		F	N	PT	DOB	-	-		Хор
12	4	F	N	PT	AD	IPN^o	SN^o	-	Ід
13		F	N	PT	AD	-	SN^o	IPN^o	Ід
14		F	N	PT	AD	IPN^o	-	SN^o	Ід
15		F	N	PT	AD	-	-	IPN^o SN^o	Хор
16	5	F	DOB	DS^o	DN^o	DED^o	-	-	Ід
17		F	DOB	-	DN^o	DED^o	-	DS^o	Ід
18		F	DOB	DS^o	-	DED^o	-	DN^o	Ід
19		F	DOB	DS^o	DN^o	-	-	DED^o	Ід
20		F	DOB	DS^o	-	-	-	DN^o DED^o	Хор
21		F	DOB	-	DN^o	-	-	DS^o DED^o	Хор
22		F	DOB	-	-	DED^o	-	DS^o DN^o	Хор
23	6	F	N	PT	SX^o	IPN^o	-	-	Ід
24		F	N	PT	-	IPN^o	-	SX^o	Ід
25		F	N	PT	SX^o	-	-	IPN^o	Ід
26		F	N	PT	-	-	-	SX^o IPN^o	Пог
27	7	F	N	PT	DS^o	DN^o	SN^o	-	Ід
28		F	N	PT	DS^o	DN^o	-	SN^o	Ід
29		F	N	PT	-	DN^o	SN^o	DS^o	Ід
30		F	N	PT	DS^o	-	SN^o	DN^o	Ід
31		F	N	PT	DS^o	-	-	DN^o SN^o	Хор
32		F	N	PT	-	DN^o	-	DS^o SN^o	Хор
33		F	N	PT	-	-	SN^o	DS^o DN^o	Хор
34		F	N	PT	-	-	-	DS^o DN^o SN^o	Хор
35	8	F	N	PT	FFN^o	FLN^o	FDOB^o	-	Ід
36		F	N	PT	-	FLN^o	FDOB^o	FFN^o	Ід
37		F	N	PT	FFN^o	FLN^o	-	FDOB^o	Ід

Продовження таблиці 4

Індекс	**Хеш-код**	**Комбінації полів з ПД**						**Відсутні поля**	**Тип хеш-коду**
38		F	N	PT	FFN^{o}	-	FDOB^{o}	FLN^{o}	Ід
39		F	N	PT	FFN^{o}	-	-	FLN^{o} FDOB^{o}	Хор
40		F	N	PT	-	FLN^{o}	-	FFN^{o} FDOB^{o}	Хор
41		F	N	PT	-	-	FDOB^{o}	FFN^{o} FFN^{o}	Хор
42	9	F	N	MFN^{o}	MLN^{o}	MDOB^{o}	-	-	Ід
43		F	N	-	MLN^{o}	MDOB^{o}	-	MFN^{o}	Ід
44		F	N	MFN^{o}	MLN^{o}	-	-	MDOB^{o}	Ід
45		F	N	MFN^{o}	-	MDOB^{o}	-	MLN^{o}	Ід
46		F	N	MFN^{o}	-	-	-	MLN^{o} MDOB^{o}	Хор
47		F	N	-	MLN^{o}	-	-	MFN^{o} MDOB^{o}	Хор
48		F	N	-	-	MDOB^{o}	-	MFN^{o} MLN^{o}	Хор
49	10	F	N	AD	MFN^{o}	MLN^{o}	IPN^{o}	-	Ід
50		F	N	AD	MFN^{o}	MLN^{o}	-	IPN^{o}	Ід
51		F	N	AD	-	MLN^{o}	IPN^{o}	MFN^{o}	Ід
52		F	N	AD	MFN^{o}	-	IPN^{o}	MLN^{o}	Ід
53		F	N	AD	MFN^{o}	-	-	MLN^{o} IPN^{o}	Хор
54		F	N	AD	-	MLN^{o}	-	MFN^{o} IPN^{o}	Хор
55		F	N	AD	-	-	IPN^{o}	MFN^{o} MLN^{o}	Хор

Оскільки хеш-код формується з комбінації полів ПД, він пов’язаний з набором полів ПД. Як тільки він збігається з одним з хеш-кодів ідентифікованої особи, відповідний набір полів ПД також є таким, що збігається, і ці поля ПД вважатимуться перевіреними.

Уточнення інформації ПД щодо сумнівних записів. Відповідно до етапу первинного формування ЄРІФО, кожний запис реєстру може знаходитися в одному з чотирьох станів:

а) “verified” – запис верифікований з еталонним реєстром;

б) “approved” – запис визнано верифікованим через зіставлення даних з різних реєстрів, але для запису відсутній аналогічний запис в еталонному реєстрі;

в) “unapproved” – запис визнано не верифікованим через брак ідентичних даних в різних реєстрах;

г) “unknown” – запис щодо особи, який створено вперше і відсутні інші джерела для порівняння (наприклад, інформація щодо новонародженої дитини).

Після зіставлення реєстрів для кожного запису з статусами “approved” і “verified” генерується унікальний ідентифікатор та сформоване значення тимчасового УНЗР.

Для записів “unapproved” потрібно провести ручну верифікацію шляхом запрошення до органів розпорядників реєстрів відповідних осіб для проходження ідентифікації в режимі опитування.

Розгортання публічного сервісу ідентифікації. На момент закінчення первинного наповнення ЄРІФО та початку етапу уточнення інформації ПД щодо сумнівних записів,

шляхом використання прикладних програмних інтерфейсів має вже бути побудована та налаштована інформаційна взаємодія з метою синхронізації внесених змін між реєстрами-наповнювачами та ЄРІФО. Відповідно, така синхронізація повинна бути двосторонньою.

Штатне функціонування публічного сервісу ідентифікації. *Транзакції на отримання інформації.* В штатному режимі ЄРІФО на запити інших інформаційних систем та електронних інформаційних ресурсів щодо персональних даних через прикладний програмний інтерфейс видає останню актуальну версію значень ПД відносно запитуваної особи шляхом використання унікального ідентифікатора. При правильно налаштованій синхронізації такі запити виконуються тільки тими інформаційними реєстрами та електронними інформаційними ресурсами, які не приймали участі в первинному формуванні ЄРІФО і з якими відсутня будь-яка синхронізація. Фактично, у майбутньому, після видалення базового набору ПД з інших реєстрів з метою реалізації у всіх інформаційних системах та електронних інформаційних ресурсах принципу "once only" та відсутності дублювання інформації, така процедура використовуватиметься усіма інформаційними системами та електронними інформаційними ресурсами.

Транзакції на створення нових записів. Транзакції щодо створення нових записів ЄРІФО виконуються у штатному режимі посадовими особами державних органів, яким надано таке право. Кожне таке внесення повинно контролюватися в автоматичному режимі засобами логічного контролю на коректність введення інформації (тип даних, мова введення) та повинно відбуватися з прив'язкою до деякого набору вивірених довідників (наприклад, довідник поштових адрес). Кожна така успішна транзакція завершується створенням для запису БД унікального ідентифікатора та формуванням для нього тимчасового УНЗР.

Транзакції на внесення змін. Кожна транзакція на внесення в ЄРІФО змін до існуючих записів відбувається згідно з алгоритмом консенсусу.

Алгоритм консенсусу. У алгоритмі консенсусу для перевірки транзакцій власники реєстрів у ході голосування вибирають валідаторів транзакцій, які будуть відповідати за певний набір полів ПД. Вага кожного голосу визначається сумою рангів полів ПД, щодо яких вносяться зміни. Завдяки цьому можна досягти високої якості щодо внесеної інформації.

Алгоритм консенсусу залишається децентралізованим в тому сенсі, що всі реєстри беруть участь у виборі набору полів ПД, які перевіряються транзакціями, а, також, він є централізованим через те, що всі рішення приймає невелика група реєстрів-учасників. Реалізація алгоритму підтримує постійний процес голосування і систему тасування, яка може вибирати, у разі збільшення кількості учасників, випадкових валідаторів.

Наприклад, при внесенні змін до поля "Прізвище" будь-яким реєстратором, якому надано право вносити зміни до ЄРІФО, згідно алгоритму консенсусу валідатором буде обрано Міністерство внутрішніх справ (МВС), яке і буде приймати остаточне рішення щодо внесення даних змін до реєстру. А при внесенні змін до полів ПД "Прізвище" та "Місце народження", валідаторами будуть МВС та Міністерство юстиції (МЮ), і саме ці дві структури прийматимуть рішення щодо внесення або невнесення вказаних змін ПД.

Пропонується така схема розподілення полів ПД, їх ваг та валідаторів (див. табл. 5).

Таблиця 5 – Поля ПД, ваги, валідатори

Тип	Поле ПД	Вага	Валідатор
Обов'язковий	Прізвище	5	МВС
	Ім'я	5	МВС
	По батькові	5	МВС
	Місце народження	5	МЮ
	Дата народження	5	МЮ

Продовження таблиці 5

Тип	Поле ПД	Вага	Валідатор
Опціональний	Стать	2	МЮ
	Країна видачі документа особи	2	МВС
	Серія посвідчення особи	2	МВС
	Номер документа особи	2	МВС
	Дата видачі документа особи	2	МВС
Опціональний	Дата закінчення дії документа особи	2	МВС
	Ким видано документа особи	2	МВС
	Індивідуальний податковий номер	5	МінФін
	Номер соціального страхування	5	МінСоц
	Ім’я матері особи	3	МЮ
	Прізвище матері особи	3	МЮ
	Дата народження матері особи	3	МЮ
	Ім’я батька особи	1	МЮ
	Прізвище батька особи	1	МЮ
	Дата народження батька особи	1	МЮ

Висновки. Для оцінки запропонованої технології відібрано 100 000 записів даних щодо осіб з випадковими помилками в ПД. Ці помилки були поміщені в обов’язкові та опціональні поля БД створюваного реєстру. В результаті цього у 92331 записах ідентифіковано внесені помилки, решта записів ідентифіковані як нові записи. Працездатність запропонованої технології також перевірено шляхом реєстрації нових осіб у реєстрі. Як результат, представлена технологія в найкращому випадку дозволяє точно визначити місце помилки для конкретного запису ПД в БД, в гіршому – звузити область пошуку відповідної помилки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Верховна Рада України. VI скликання, 11 сесія. (2012, Вер. 06). *Закон № 5203-VI, Про адміністративні послуги.* [Електронний ресурс]. Доступно: http://zakon2.rada.gov.ua/laws/show/5203-17. Дата звернення: 06.09.19.

[2] Верховна Рада України. VI скликання, 11 сесія. (2012, Лис. 20). *Закон № 5492-VI, Про Єдиний державний демографічний реєстр та документи, що підтверджують громадянство України, посвідчують особу чи її спеціальний статус.* [Електронний ресурс]. Доступно: https://zakon.rada.gov.ua/laws/card/5492-17. Дата звернення: 06.09.19.

[3] Кабінет Міністрів України. (2016, Вер. 08). *Постанова Кабінету Міністрів України № 606.* [Електронний ресурс]. Доступно: https://zakon.rada.gov.ua/laws/show/606-2016-%D0%BF. Дата звернення: 06.09.19.

[4] Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, HHS. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed on: 06.09.19.

[5] J.B. Freymann, J.S. Kirby, J.H. Perry, D.A. Clunie, and C.C. Jaffe, “Image data sharing for biomedical research-meeting HIPAA requirements for de-identification”, *Digit Imaging*, № 25 (1), pp. 14-24, 2012. doi: 10.1007/s10278-011-9422-x.

[6] O. Uzuner. Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification", *Am Med Inform Assoc*., № 14 (5), pp. 550-563, 2007. doi: 10.1197/ jamia.M2444.
[7] K.El Emam, and etc., "De-identification methods for open health data: the case of the Heritage Health Prize claims dataset", *Med Internet Res*., № 27, 2012. doi: 10.2196/ jmir.2001.
[8] B.S. Elger, and etc., "Strategies for health data exchange for secondary, cross-institutional clinical research", *Comput Methods Programs Biomed*, № 99 (3), pp. 230-251, 2010. doi: 10.1016/j.cmpb.2009.12.001.
[9] Privacy rule and research nih. Clinical research and the HIPAA Privacy Rule, HSS. [Online]. Available: https://privacyruleandresearch.nih.gov/ pdf/clin_research.pdf. Accessed on: 06.09.19.
[10] L. Ohno-Machado, and etc., "iDASH: integrating data for analysis, anonymization, and sharing", *Am Med Inform Assoc*., № 19 (2), pp. 196-201, 2012. doi: 10.1136/amiajnl-2011-000538.
[11] K. Benitez, and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule", *Am Med Inform Assoc*., № 17 (2), pp. 169-177, 2010. doi: 10.1136/ jamia.2009.000026.
[12] C. Quantin, and etc., "Linking anonymous databases for national and international multicenter epidemiological studies: a cryptographic algorithm", *Epidemiol Sante Publique*, № 57 (1), pp. 33-39, 2009. doi: 10.1016/j.respe.2008.10.010.
[13] S.B. Johnson, "Using global unique identifiers to link autism collections", *Am Med Inform Assoc*., № 17 (6), pp. 689-695, 2010. doi: 10.1136/jamia.2009.002063.

Стаття надійшла до редакції 16.09.2019.

REFERENCE

[1] Verkhovna Rada Ukrainy. VI convocation, 11th session. (2012, Sept. 06). *Zakon № 5203-VI, Pro Administratyvni Posluhy.* [Online]. Available: http://zakon2.rada.gov.ua/laws/show/5203-17. Accessed on: 06.09.19.
[2] Verkhovna Rada Ukrainy. VI convocation, 11th session. (2012, Nov. 20). *Zakon № 5492-VI, Pro Yedynyi Derzhavnyi Demohrafichnyi Reiestr Ta Dokumenty Shcho Pidtverdzhuiut Hromadianstvo Ukrainy Posvidchuiut Osobu Chy Yii Spetsialnyi Status.* [Online]. Available: https://zakon.rada.gov.ua/laws/card/5492-17. Accessed on: 06.09.19.
[3] Kabinet Ministriv Ukrainy. (2016, Sept. 08). *Postanova Kabinetu Ministriv Ukrainy № 606.* [Online]. Available: https://zakon.rada.gov.ua/laws/show/606-2016-%D0%BF. Accessed on: 06.09.19.
[4] Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, HHS. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/ special-topics/de-identification/index.html. Accessed on: 06.09.19.
[5] J.B. Freymann, J.S. Kirby, J.H. Perry, D.A. Clunie, and C.C. Jaffe, "Image data sharing for biomedical research-meeting HIPAA requirements for de-identification", *Digit Imaging*, № 25 (1), pp. 14-24, 2012. doi: 10.1007/s10278-011-9422-x.
[6] O. Uzuner. Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification", *Am Med Inform Assoc*., № 14 (5), pp. 550-563, 2007. doi: 10.1197/ jamia.M2444.
[7] K.El Emam, and etc., "De-identification methods for open health data: the case of the Heritage Health Prize claims dataset", *Med Internet Res*., № 27, 2012. doi: 10.2196/ jmir.2001.
[8] B. S. Elger, and etc., "Strategies for health data exchange for secondary, cross-institutional clinical research", *Comput Methods Programs Biomed*, № 99 (3), pp. 230-251, 2010. doi: 10.1016/j.cmpb.2009.12.001.

[9] Privacy rule and research nih. Clinical research and the HIPAA Privacy Rule, HSS. [Online]. Available: https://privacyruleandresearch.nih.gov/ pdf/clin_research.pdf. Accessed on: 06.09.19.
[10] L. Ohno-Machado, and etc., "iDASH: integrating data for analysis, anonymization, and sharing", *Am Med Inform Assoc*., № 19 (2), pp. 196-201, 2012. doi: 10.1136/amiajnl-2011-000538.
[11] K. Benitez, and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule", *Am Med Inform Assoc*., № 17 (2), pp. 169-177, 2010. doi: 10.1136/ jamia.2009.000026.
[12] C. Quantin, and etc., "Linking anonymous databases for national and international multicenter epidemiological studies: a cryptographic algorithm", *Epidemiol Sante Publique*, № 57 (1), pp. 33-39, 2009. doi: 10.1016/j.respe.2008.10.010.
[13] S.B. Johnson, "Using global unique identifiers to link autism collections", *Am Med Inform Assoc*., № 17 (6), pp. 689-695, 2010. doi: 10.1136/jamia.2009.002063.

YAROSLAV DOROHYI

SEARCHING TECHNOLOGY FOR QUESTIONABLE RECORDS WHEN CREATING THE UNIFIED REGISTRY OF UKRAINIAN INDIVIDUALS IDENTIFICATION

One of the most effective solutions for protecting personal data when building a Unified Identity Registry is to share end-to-end identifier and hash codes generated from combinations of personal data using one-sided hash functions. This is due to the fact that the stage of creating a unified personal identification register does not involve the use of open personal data and therefore no personal data is allowed on the server and only unique identifiers and hash codes are allowed. In accordance with the principles of creating the above registry, five required and fifteen optional types of personal data stored in the registry were analyzed and used to generate hash codes, as well as possible combinations of personal data fields (ten different combinations of personal data were used in the work) data) built on the types specified. The technology of end-to-end identification has been developed, which has the ability to track errors in the fields with personal data when entering new data and when searching the registry. For the evaluation of the proposed technology, 100,000 simulated individuals were selected with random errors in the appropriate fields that store personal data. These errors are randomly placed in the fields of the created registry database that store personal information of the required and optional types. The efficiency of the proposed technology has also been verified by registering new persons in the registry. The proposed technology has a high tolerance for errors and can correctly identify and associate an individual, even with errors in multiple fields of personal data. Correct personal data, especially in the fields of the database with mandatory personal data, is crucial to avoid erroneous entries in the created registry. In the context of one-sided hash transformation, a doubtful record with personal data can be identified by applying hash operators based on hash codes calculated according to certain combinations of personal data.

Keywords: data accuracy; personal data; privacy; end-to-end identifier; registry; confidentiality.

Дорогий Ярослав Юрійович, кандидат технічних наук, доцент, доцент кафедри автоматики і управління в технічних системах, Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна.
ORCID: 0000-0003-3848-9852.
E-mail: argusyk@gmail.com.

Dorohyi Yaroslav, candidate of technical sciences, associate professor, associate professor in department of Automation and Control in Technical Systems, National technical university of Ukraine "Igor Sikorsky Kyiv polytechnic institute", Kyiv, Ukraine.