Państwowa Medyczna Wyższa Szkoła Zawodowa w Opolu

ISSN 2080-2021

Prace oryginalne - Original papers

PROBLEMY ANONIMIZACJI DOKUMENTÓW MEDYCZNYCH. CZĘŚĆ 2. ANONIMIZACJA ZAAWANSOWANA ORAZ STEROWANA PRZEZ POSIADACZA DANYCH WRAŻLIWYCH

The issues connected with the anonymization of medical data. Part 2. Advanced anonymization and anonymization controlled by owner of protected sensitive data

ARKADIUSZ LIBER

Instytut Informatyki, Politechnika Wrocławska

A- przygotowanie projektu badania (study design), **B**- zbieranie danych (data collection), **C**- analiza statystyczna (statistical analysis), **D**- interpretacja danych (data interpretation), **E**- przygotowanie maszynopisu (manuscript preparation), **F**- opracowanie piśmiennictwa (literature search), **G**- pozyskanie funduszy (funds collection)

Streszczenie

Wstęp: Dokumentację medyczną powinno się udostępniać z zachowaniem jej integralności oraz ochrony danych osobowych. Jednym ze sposobów zabezpieczenia danych przed ujawnieniem jest anonimizacja. Współczesne metody zapewniają anonimizację bez uwzględnienia możliwości sterowania dostępem do danych wrażliwych. Wydaje się, że przyszłość systemów przetwarzania danych wrażliwych należy do metod personalizowanych. W części pierwszej omówiono metody *k*-anonimizacji, (*X*,*Y*)-anonimizacji, (α,*k*)-anonimizacji oraz (*k*,*e*)-anonimizacji. Metody te należy zaliczyć do dobrze znanych metod elementarnych, które są przedmiotem znacznej liczby publikacji. Jako materiały źródłowe do metod anonimizacji opisanych w części pierwszej podano prace Samaratiego, Sweeneya, Wanga, Wonga i Zhanga oraz innych. Wybór tych pozycji uzasadniony jest szerszymi badaniami przeglądowymi, prowadzonymi na przykład przez Funga, Wanga, Fu i Yu. Należy jednak zwrócić uwagę na fakt, iż metody anonimizacji danych wywodzą się z metod ochrony statystycznych baz danych, które sięgają lat siedemdziesiątych dwudziestego wieku. Ze względu na powiązaną treść oraz odnośniki literaturowe część pierwsza i druga stanowią integralną całość.

Cel pracy: Analiza metod anonimizacji, metod ochrony zanonimizowanych danych oraz opracowanie nowego typu zabezpieczenia prywatności umożliwiającego sterowanie udostępnianiem danych wrażliwych przez podmiot, którego te dane dotyczą.

Materiał i metody: Metody analityczne.

Wyniki: Dostarczenie materiału wspomagającego wybór i analizę sposobów anonimizacji danych medycznych, opracowanie nowego typu zabezpieczenia umożliwiające kontrolę danych wrażliwych przez podmioty, których dane te dotyczą.

Wnioski: W pracy przeprowadzono analizę rozwiązań w zakresie anonimizacji danych pod kątem zastosowania ich do ochrony prywatności w zbiorach danych medycznych. Przeprowadzono analizę takich metod, jak: *k*-anonimizacji, (*X*,*Y*)-anonimizacji, (*α*,*k*)-anonimizacji, (*k*,*e*)-anonimizacji, *l*-dywersyfikacji, (*X*,*Y*)-dołączalności, (*X*,*Y*)-prywatności, LKC-prywatności, *t*-bliskości, ograniczonego zaufania oraz personalizowanej prywatności. Szczególnej analizie poddano problem możliwości personalizacji sterowania prywatnością danych wrażliwych przez podmiot, którego dane te dotyczą. Oprócz samych metod anonimizacji przeprowadzono analizę metod ochrony zanonimizowanych danych. W szczególności zaś metod: *δ*-obecności, prywatności *e*-różnicowej, (*d*,*γ*)-prywatności, prywatności (*α*,*β*)-dystrybucyjnej oraz ochrony przed (*c*,*t*)-izolacją. W pracy zaproponowano nowe rozwiązanie w zakresie kontrolowanej ochrony prywatności. Rozwiązanie oparte jest na wydzieleniu chronionych pól i wielokluczowym szyfrowaniu i deszyfrowaniu danych wrażliwych. Zaproponowano sposób wydzielenia pól zgodny z obowiązującym standardem XML. Do szyfrowania wybrany został schemat szyfrowania posiadający *n* różnych kluczy. Do deszyfrowania zawartości wystarczające jest *p* spośród wszystkich możliwych. Umożliwia to tworzenie zupełnie nowych systemów dostępu do danych wrażliwych, dając wszystkim stronom możliwość zapewnienia pełnego poszanowania i kontroli ich prywatności.

Słowa kluczowe: anonimizacja danych, dokumentacja medyczna, prywatność w ochronie zdrowia, kontrola właściciela w udostępnianiu danych medycznych, kryptografia wielokluczowa

Summary
Introduction: Medical documentation ought to be accessible with the preservation of its integrity as well as the protection of personal data. One of the manners of its protection against disclosure is anonymization. Contemporary methods ensure anonymity without the possibility of sensitive data access control. It seems that the future of sensitive data processing systems belongs to the personalized method. In the first part of the paper k-Anonymity, (X,Y)- Anonymity, (α,k)- Anonymity, and (k,e)-Anonymity methods were discussed. These methods belong to well - known elementary methods which are the subject of a significant number of publications. As the source papers to this part, Samarati, Sweeney, Wang, Wong and Zhang's works were accredited. The selection of these publications is justified by their wider research review work led, for instance, by Fung, Wang, Fu and Y. However, it should be noted that the methods of anonymization derive from the methods of statistical databases protection from the 70s of 20th century. Due to the interrelated content and literature references the first and the second part of this article constitute the integral whole.
Aim of the study: The analysis of the methods of anonymization, the analysis of the methods of protection of anonymized data, the study of a new security type of privacy enabling device to control disclosing sensitive data by the entity which this data concerns.
Material and methods: Analytical methods, algebraic methods.
Results: Delivering material supporting the choice and analysis of the ways of anonymization of medical data, developing a new privacy protection solution enabling the control of sensitive data by entities which this data concerns.
Conclusions: In the paper the analysis of solutions for data anonymization, to ensure privacy protection in medical data sets, was conducted. The methods of: k-Anonymity, (X,Y)- Anonymity, (α,k)- Anonymity, (k,e)-Anonymity, (X,Y)-Privacy, LKC-Privacy, l-Diversity, (X,Y)-Linkability, t-Closeness, Confidence Bounding and Personalized Privacy were described, explained and analyzed. The analysis of solutions of controlling sensitive data by their owner was also conducted. Apart from the existing methods of the anonymization, the analysis of methods of the protection of anonymized data was included. In particular, the methods of: δ-Presence, e-Differential Privacy, (d, γ)-Privacy, (α,β)-Distributing Privacy and protections against (c,t)-Isolation were analyzed. Moreover, the author introduced a new solution of the controlled protection of privacy. The solution is based on marking a protected field and the multi-key encryption of sensitive value. The suggested way of marking the fields is in accordance with XML standard. For the encryption, (n,p) different keys cipher was selected. To decipher the content the p keys of n were used. The proposed solution enables to apply brand new methods to control privacy of disclosing sensitive data.
Keywords: data anonymization, health documents, privacy in health care, owner controlled access to medical data, multi key cryptography

1. Wprowadzenie

W pierwszej części artykułu [1] przedstawiono analizę metod zabezpieczeń wrażliwych danych medycznych przy wykorzystaniu metod: *k*-anonimizacji, (*X*,*Y*)-anonimizacji, (*α*,*k*)-anonimizacji oraz (*k*,*e*)-anonimizacji. Metody te należy zaliczyć do dobrze znanych metod elementarnych, które są przedmiotem znacznej liczby publikacji. Przegląd tych metod można znaleźć na przykład w pracy Funga, Wanga, Fu i Yu [2]. W opublikowanych do tej pory pracach brak jest pozycji, które zawierają kompletną analizę metod anonimizacji danych w odniesieniu do specyfiki ochrony danych medycznych, w świetle obowiązujących przepisów prawnych oraz specyficznych rozwiązań technicznych stosowanych w służbie zdrowia. W literaturze polskojęzycznej na szczególną uwagę zasługuje praca Boruckiego [3], w której autor podejmuje próbę opisu stanu prawnego oraz opisu prostych metod stosowanych do ochrony prywatności danych medycznych. W pracy [3], jako metodę ochrony wrażliwych danych medycznych, przedstawiono opis prostego modelu separacji słownikowej rekordów.

Jako prace źródłowe dla metod anonimizacji opisanych w pierwszej części artykułu podano prace Samaratiego, Sweeneya, Wanga, Wonga i Zhanga oraz innych [4-10], uważane powszechnie za podstawowe w tym zakresie. Definicje metod opartych na *k*-anonimizacji ulegają jednak ciągłej ewolucji. Przykładem mogą tu być definicje proponowane w pracach [11,12], które nie doczekały się jeszcze głębszej analizy porównawczej z innymi metodami (na przykład z metodami zawartymi w pracy [2]). Przy analizie prac źródłowych dotyczących algorytmów anonimizacji należy pamiętać, iż wiele z nich swój początek i rozwój zawdzięcza intensywnie badanym w latach siedemdziesiątych XX wieku metodom ochrony statystycznych baz danych [13].

W dalszej części pracy przedstawione zostały kolejne metody anonimizacji oraz metody ochrony stanu anonimizacji danych medycznych. Metody te stanowią podstawę do implementacji kolejnych mechanizmów ochrony danych osobowych w aktualnie funkcjonujących systemach oraz służą do konstrukcji nowych typów zabezpieczeń baz danych.

Analizując dostępne rozwiązania oraz literaturę przedmiotu, można zauważyć, iż przy tworzeniu i wprowadzaniu nowych metod anonimizacji bardzo często pomija się funkcjonalności kontrolne sprawowane przez podmiot, którego dane wrażliwe dotyczą.

W pracy przedstawiono propozycję autora w zakresie rozwiązania umożliwiającego sterowanie udostępnianiem danych wrażliwych przez ich posiadacza lub inny upoważniony podmiot.

Ze względu na powiązaną treść, odnośniki do literatury oraz wyjaśnienia, część pierwsza i druga artykułu stanowią integralną całość.

2. Metody anonimizacji danych oparte na dywersyfikacji, ograniczeniu zaufania oraz inne złożone metody anonimizacji i ochrony

W dalszej części pracy przeprowadzono analizę takich metod, jak: *l*-dywersyfikacji, (*X*,*Y*)-dołączalności, (*X*,*Y*)-prywatności, LKC-prywatności, *t*-bliskości, ograniczonego zaufania oraz personalizowanej prywatności. Sama anonimizacja nie wystarczy do ochrony danych wrażliwych. Konieczne jest również prowadzenie ochrony danych zanonimizowanych. Do ochrony

takich danych służą metody: δ-obecności, prywatności *e*-różnicowej, (d,γ)-prywatności, prywatności (α,β)-dystrybucyjnej oraz ochrony przed (c,t)-izolacją oraz inne, których podstawą są własności statystyczne chronionej zawartości.

2.1. Zapewnienie anonimowości metodą *l*-dywersyfikacji

Metoda l-dywersyfikacji [14] związana jest bezpośrednio z atrybutem zawierającym dane wrażliwe. W metodzie *l*-dywersyfikacji przyjmuje się, iż dla każdej grupy wartości pseudoidentyfikatora powinno występować przynajmniej *l* „dobrze reprezentowanych" wartości danych wrażliwych. Pod pojęciem dobrej reprezentacji rozumie się odpowiednie dobranie wartości wrażliwego atrybutu, tak aby - z jednej strony - zapewnić ich odpowiednią liczność a - z drugiej - poprawność dziedzinową. W przypadku danych medycznych „dobra reprezentacja" wiąże się, na przykład z konstrukcją dobrze dobranych grup jednostek chorobowych, badań czy zabiegów. Metoda *l*-dywersyfikacji wprowadza ochronę danych zarówno przed dołączaniem do rekordów, jaki i dołączaniem do atrybutów.

Zastosowanie tej metody nie jest możliwe dla atrybutów separujących tabele na przykład w metodzie separacji słownikowej opisanej przez Boruckiego [3]. Podobnie jest dla przypadku, gdy atrybut separujący należy do zbioru atrybutów tworzących pseudoidentyfikator.

Metoda *l*-dywersyfikacji implikuje występowanie w zbiorze danych *k*-anonimizacji dla pseudoidentyfikatora. Wynika stąd, że współczynnik *k*-anonimizacji jest równy przynajmniej współczynnikowi *l*-dywersyfikacji. Dzięki temu w praktycznych rozwiązaniach implementowanych w systemach zabezpieczeń baz danych medycznych można implementować algorytmy *l*-dywersyfikacji, bez konieczności dodawania modułów realizujących *k*-anonimizację.

Zastosowanie l-dywersyfikacji zabezpiecza przed statystycznym ujawnieniem danych wrażliwych. W przypadku danych medycznych szczególnie dobrze do tworzenia zdywersyfikowanych grup nadają się jednostki chorobowe. Dosyć łatwo można tu kontrolować prawdopodobieństwo ujawnienia cechy, gdyż jest ono równe odwrotności poziomu dywersyfikacji i wynosi 1/l. Podobnie jak przy innych metodach anonimizacji, podczas opracowywaniu nowych typów algorytmów, jakość rozwiązania należy oceniać na podstawie wartości funkcji entropii ujawnianej informacji w grupie. Metoda l-dywersyfikacji jest stosunkowo ogólna i można ją zastosować do ochrony sieci powiązanych danych. Na przykład w sposób zaproponowany w pracy Prasada i innych [15]. Na uwagę zasługuje również fakt, iż metoda l-dywersyfikacji może być uważana przez niektórych badaczy jako podstawa do formalizacji prywatności [16].

2.2. Zapewnienie anonimowości metodą ograniczonego zaufania

Ograniczone zaufanie jest pojęciem dość ogólnym. W praktycznych rozwiązaniach związanych z przetwarzaniem danych zaufanie definiuje się jako miarę określoną na zbiorze danych. Granice zaufania można określić formalnie przez formalne zdefiniowanie poziomu dolnego i poziomu górnego [17].

Jako metodę zapewnienia anonimowości na drodze ograniczonego zaufania przyjmuje się najczęściej rozwiązanie zaproponowane w pracy [18]. Zaproponowane tam rozwiązanie ma postać szablonów odwzorowań wartości pseudoidentyfikatora *PID* w zbiór wartości atrybutu wrażliwego *W*, z zadanym parametrem zaufania $PID \rightarrow (W,t)$. Parametr zaufania *t* oznacza maksymalną wartość procentowego udziału wartości atrybutu *W* we wszystkich grupach wyznaczonych przez wartość pseudoidentyfikatora. Na rysunku 1 przedstawiono przykład wyliczenia wartości poziomu zaufania *t*. Przy anonimizacji danych medycznych metodą ograniczonego zaufania konieczne jest założenie wartości poziomu zaufania *t,* a następnie taki dobór wartości atrybutu wrażliwego, aby dla żadnej grupy danych poziom ten nie został przekroczony.

Płeć	Zawód	Województwo	Choroba
*	Techniczny	Małopolskie	AIDS
*	Techniczny	Małopolskie	AIDS
*	Techniczny	Małopolskie	Grypa
*	Techniczny	Małopolskie	AIDS
*	Artystyczny	Opolskie	Nowotwór
*	Artystyczny	Opolskie	Grypa
*	Artystyczny	Opolskie	Nowotwór
*	Artystyczny	Opolskie	AIDS
*	Artystyczny	Opolskie	AIDS

Obliczenie poziomu *t* dla wartości atrybutu wrażliwego Choroba

Choroba	t_1 dla *PID1*	t_2 dla *PID2*	t=max($t1$,$t2$)
AIDS	0,75	0,5	0,75
Grypa	0,25	0,25	0,25
Nowotwór	0	0,25	0,25

Rysunek 1. Przykład wyznaczenia poziomu zaufania *t* dla wartości zadanego atrybutu wrażliwego

2.3. Zapewnienie anonimowości metodą (*X*,*Y*)-dołączalności

Metoda (X,Y)-dołączalności [8] stanowi propozycję poprawienia metody (X,Y)-anonimowości. Ograniczeniem metody (*X*,*Y*)-anonimowości jest fakt, iż przy ograniczeniu zbioru danych do rekordów reprezentujących podzbiór osób, wartość prawdopodobieństwa ujawnienia danych wrażliwych może być większa niż 1/*k*. Aby poprawić ten stan rzeczy można wprowadzić wiele zbiorów Y_i, których wartości mogą być wstawiane zamiennie jako wartości danych dla atrybutu wrażliwego [2]. Dzięki wprowadzeniu różnowartościowego odwzorowania pomiędzy tymi zbiorami ($Y_i \leftrightarrow Y_j$ dla każdego i,j: $i \neq j$) możliwe jest wstawianie wartości atrybutów i ich zamienników z dowolnego

zbioru. Można w ten sposób uzyskać mniejsze prawdopodobieństwo występowania wartości atrybutów wrażliwych w tabeli wyjściowej. Tworzenie zbiorów Y_i dla atrybutów wrażliwych może być utrudnione w przypadku chorób i usług medycznych o szczególnym lub unikatowym charakterze, charakteryzujących się brakiem równoważnych odpowiedników. Na rysunku 2 przedstawiono przykład zastosowania metody (*X*,*Y*)-dołączalności.

Płeć	Zawód	Województwo	Choroba
*	Techniczny	Małopolskie	AIDS
*	Techniczny	Małopolskie	AIDS
*	Techniczny	Małopolskie	Grypa
*	Techniczny	Małopolskie	AIDS
*	Artystyczny	Opolskie	Nowotwór
*	Artystyczny	Opolskie	Grypa
*	Artystyczny	Opolskie	Nowotwór
*	Artystyczny	Opolskie	AIDS
*	Artystyczny	Opolskie	AIDS

Równoważne zbiory wartości atrybutów dla atrybutu wrażliwego Choroba

Y_1	Y_2	Y_3
AIDS	Test A pozytywny	Choroba A
Grypa	Choroba G	G zakażenie
Nowotwór	Choroba N	NN

Tabela po korekcie

Płeć	Zawód	Województwo	Choroba
*	Techniczny	Małopolskie	AIDS
*	Techniczny	Małopolskie	Test A pozytywny
*	Techniczny	Małopolskie	Grypa
*	Techniczny	Małopolskie	Choroba A
*	Artystyczny	Opolskie	Nowotwór
*	Artystyczny	Opolskie	Grypa
*	Artystyczny	Opolskie	Choroba N
*	Artystyczny	Opolskie	AIDS
*	Artystyczny	Opolskie	Choroba A

Rysunek 2. Przykład zastosowania metody (*X*,*Y*)-dołączalności. Uzyskano zmniejszenie prawdopodobieństwa wystąpienia wartości AIDS dla pierwszej grupy wyznaczonej przez *PID* z 0,75 do 0,25

Metody oparte na dołączalności są dosyć łatwe w implementacji i cieszą się niesłabnącym zainteresowaniem badawczym [19].

2.4. Zapewnienie anonimowości metodą (*X*,*Y*)-prywatności

Typowym sposobem podwyższania poziomu bezpieczeństwa jest składanie metod ochrony. Podobnie jest w przypadku metod anonimizacji. Przykładem może być tu metoda (*X*,*Y*)-prywatności [2,8], która jest połączeniem dwóch metod: (*X*,*Y*)-anonimowości i (*X*,*Y*)-dołączalności. Ogólną ideą jest tu zapewnienie, aby liczba elementów w grupach wyznaczanych przez pseudoidentyfikator *PID* była nie mniejsza niż *k* oraz jednocześnie spełniony była zasada ograniczonego zaufania. To jest warunek, aby częstość wystąpień wartości atrybutów ze zbioru *Y* w grupach nie przekraczała maksymalnej wartości *t*. Metoda ta jest podobna do (α,*k*)-anonimizacji, ale jest od niej ogólniejsza [20].

2.5. Zapewnienie anonimowości metodą LKC-prywatności

Anonimowość jest cechą pewnego elementu *e* posiadającego cechy należące do zbioru P_1 [1]. Przy przekształcaniu danych opisujących cechy tego elementu, do postaci spełniających warunki anonimizacji, najczęściej stosuje się metody generalizacji atrybutów i wartości. Otrzymuje się w wyniku tego zbiór cech zanonimizowanych P_{a1} identyfikujących więcej niż jeden element *e*. Pomija się często istotną rolę w ujawnieniu, jaką pełni wiedza dodatkowa o pacjencie, czyli cech, które należą do pewnego udostępnianego publicznie zbioru *P'*, mającego wspólne elementy ze zbiorem P_1 ($P' \cap P_1 \neq \emptyset$). Pozyskanie danych wrażliwych z reguły następuje na podstawie *L* atrybutów stanowiących podzbiór zbioru atrybutów tworzących pseudoidentyfikator. To spostrzeżenie jest podstawą konstrukcji metody LKC-prywatności [2,21]. Zakłada się w niej, że *L* jest maksymalną ilością atrybutów wartości pochodzących z zasobów zewnętrznych. Na podstawie wartości liczby *L* konstruuje się warunek na zbiór atrybutów pseudoidentyfikatora *PID*. Liczność tego zbioru nie może być większa niż *L*. Kolejnym warunkiem jest ograniczenie liczby rekordów dla wszystkich możliwych zbiorów pseudoidentyfikatorów tak, aby była nie mniejsza niż *K*. Ostatnim ograniczeniem jest wymaganie, aby wszystkie wartości dla atrybutu wrażliwego posiadały prawdopodobieństwo wystąpień w grupach nie większe niż *C*. Wartości *L*, *K*, *C* są ustalane przez właściciela danych.

Metoda LKC-prywatności gwarantuje ograniczenie prawdopodobieństwa dołączenia rekordu do wartości mniejszej lub równej 1/*k* oraz ograniczenie prawdopodobieństwa dołączenia atrybutów do wartości mniejszej lub równej *C*. Metoda LKC-prywatności nadaje się dobrze do anonimizacji danych wielowymiarowych. Przyglądając się metodzie LKC, widać iż stanowi ona próbę połączenia i uogólnienia innych metod anonimizacji. Wydaje się, iż już na etapie bieżących badań nad algorytmami anonimizacji należałoby wprowadzić metody opisu anonimizowanych cech w ogólnej postaci wektorowej.

2.6. Zapewnienie anonimowości metodą *t*-bliskości

Kolejną metodą zapewnienia anonimowości jest metoda t-bliskości. Pomimo że metoda ta nie daje pełnej ochrony prywatności w odniesieniu do zakresów danych [22], stanowi ważną metodę opartą na podobieństwie rozkładu cech.

Metoda *t*-bliskości [2,23] nakłada ograniczenia na rozkłady prawdopodobieństwa występowania wartości atrybutów wrażliwych w grupach identyfikowanych przez pseudoidentyfikatory *PID* oraz w całej tabeli. Dąży się do tego, aby oba rozkłady były bliskie. Odległość między rozkładami oblicza się stosując na przykład metrykę Wassersteina. Na rysunku 3 przedstawiono przykład tabeli spełniającej wymagania metody *t*-bliskości. Zgodność rozkładów w tej tabeli zapewniają jednakowe wartości prawdopodobieństwa wystąpienia każdej z chorób w grupach *PID* (*Płeć*, *Zawód*, *Województwo*).

Płeć	Zawód	Województwo	Choroba
M	Techniczny	Małopolskie	AIDS
M	Techniczny	Małopolskie	Nowotwór
M	Techniczny	Małopolskie	Grypa
K	Artystyczny	Dolnośląskie	AIDS
K	Artystyczny	Dolnośląskie	Nowotwór
K	Artystyczny	Dolnośląskie	Grypa
K	Artystyczny	Opolskie	Nowotwór
K	Artystyczny	Opolskie	Grypa
K	Artystyczny	Opolskie	AIDS

Rysunek 3. Tabela spełniająca warunki metody *t*-bliskości

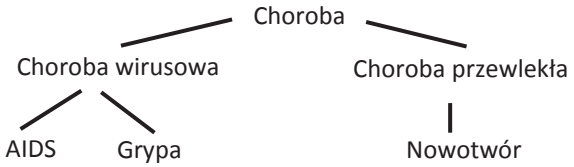
Ciekawym rozwiązaniem polepszającym jakość anonimizacji w stosunku do t-dywersyfikacji, jest dywersyfikacja zakresu zaproponowana w [23].

2.7. Zapewnienie anonimowości metodą personalizowanej prywatności

Analizując prawne oraz algorytmiczne metody ochrony dóbr osobistych wydaje się nieuniknione rozwijanie w przyszłości tych metod, w których możliwa będzie pełna kontrola udostępniania danych wrażliwych przez podmiot, którego te dane dotyczą. Personalizowania prywatność już teraz jest przedmiotem ciekawych propozycji rozwiązywania problemów identyfikacji osób pojawiających się w sieciach społecznościowych [24]. Podstawowa metoda personalizowanej prywatności, przedstawiona w pracy [25], pozwala na określenie poziomu prywatności rekordów przez ich właścicieli. W modelu tym [2,25] każdy atrybut posiada drzewo nazw a właściciel rekordu wskazuje węzeł ochrony z tego drzewa. Węzły ochrony pozwalają na kontrolę poziomu prywatności informacji związanej z pacjentem. Na rysunku 4 przedstawiono przykład drzewa nazw oraz tabeli z personalizowaną prywatnością.

2.8. Ochrona anonimowości metodą *δ*-obecności

Metoda *δ*-obecności [2,22,26] związana jest z zabezpieczeniem wrażliwych danych przed dołączaniem tabel. Poszukując danych wrażliwych w publicznie udostępnianych zbiorach danych pojawia się pytanie, czy w ogóle dane ofiary są tam zawarte. Aby uniemożliwić odpowiedź na to pytanie stosuje się metodę *δ*-obecności. Metoda polega na tym, iż publikowane dane spełniają warunek, że potencjalny rekord ofiary znajduje się tam z zadanym z góry prawdopodobieństwem p_0 należącym do przedziału $[p_{min}, p_{max}]$. Źródłowa tabela *T* zostaje przekształcona do nowej tabeli *T'* (na przykład metodą generalizacji) w taki sposób, aby prawdopodobieństwo występowania rekordów identyfikowanych przez zewnętrzną tablicę T_E spełniało zależność: $\forall$ t$\in$ TE: p0 = p0 (t$\in$ T' dla T $\subseteq$ TE) $\in$ [pmin, pmax]. Jako parametr δ przyjmuje się graniczne wartości prawdopodobieństwa z przedziału $[p_{min}, p_{max}]$ [2].



Właśc.	Płeć	Wiek	Miasto	Choroba	Węzeł bezpieczeństwa
Jan	M	25	Opole	AIDS	Choroba wirusowa
Tomasz	M	31	Kraków	Nowotwór	Choroba przewlekła
Adam	M	28	Brzeg	Grypa	Grypa
Iwona	K	43	Opole	AIDS	Choroba wirusowa
Ewa	K	68	Wrocław	Nowotwór	
Anna	K	22	Wrocław	Grypa	Choroba wirusowa
Teresa	K	53	Opole	Nowotwór	Choroba przewlekła
Inga	K	19	Opole	Grypa	φ

Rysunek 4. Drzewo nazw oraz tabela źródłowa z personalizowaną prywatnością

2.9. Ochrona anonimowości metodą zabezpieczenia przed (*c*,*t*)-izolacją

Dużą grupę metod anonimizacji stanowią rozwiązania oparte na modelach losowych. Przedstawicielami tej grupy mogą być na przykład metody: zabezpieczenia przed izolacją, prywatności *e*-różnicowej czy też (*d*,*γ*)-prywatności. W ogólnym przypadku metody te mogą być opisane przy użyciu automatów losowych i losowych π-obliczeń [27]. Metoda zabezpieczenia przed (*c*,*t*)-izolacją [27] należy do grupy metod ochrony przed zmianą wiedzy lub wyobrażeń o ofierze, które są już w posiadaniu atakującego. Baza danych jest tu reprezentowana jaklo *n* punktów w wielowymiarowej przestrzeni (wymiar tej przestrzeni jest równy liczbie atrybutów). Punkt x_0 w bazie danych (*c*,*t*)-izoluje punkt *x*, jeżeli w kuli o środku w punkcie x_0 i promieniu $r=c||x_0-x||$ (gdzie ||...|| oznacza normę w przestrzeni atrybutów) zawartych jest mniej niż *t* punktów. Zabezpieczenie polega na takim przekształceniu tabeli źródłowej, aby nie dochodziło do (*c*,*t*)-izolacji. Uzyskiwany tą metodą efekt jest podobny do zabezpieczenia przed dołączaniem rekordów.

2.10. Ochrona anonimowości metodą prywatności *e*-różnicowej

Metoda ta oparta jest na dążeniu do tego, aby dodanie lub usunięcie pojedynczego rekordu nie wpływało istotnie na rezultaty zewnętrznej analizy opublikowanych danych [2,28,29]. Metoda ta nie zabezpiecza bezpośrednio tabeli przed ujawnieniem danych wrażliwych metodami dołączenia rekordów lub dołączania atrybutów. Zapewnia raczej stałą jakość i ograniczenie ryzyka związanego z utrzymaniem bazy danych. Parametr e określa maksymalną dopuszczalną zmianę logarytmu ze stosunków prawdopodobieństw.

2.11. Ochrona anonimowości metodą (*d*,*γ*)-prywatności

Metoda (*d*,*γ*)-prywatności związana jest z różnicą prawdopodobieństw obecności rekordu ofiary w bazie danych przed przeprowadzeniem i po przeprowadzeniu badań [27,30]. Parametr *d* jest równy prawdopodobieństwa dostępu do rekordów z zewnątrz bazy danych (z reguły przed poznaniem jej zawartości). Po dokonaniu analizy rekordów znajdujących się w bazie danych prawdopodobieństwo to może ulec zmianie. Parametr *γ* ogranicza stosunek prawdopodobieństwa dostępu do rekordów przed analizą do prawdopodobieństwa dostępu do rekordów po analizie. Metoda ta podobnie jak metoda prywatności *e*-różnicowej stosowana jest do zapewnienia odpowiedniej jakości danych i ochrony przed dostarczaniem dodatkowej wiedzy o ofierze w stosunku do już posiadanej ze źródeł zewnętrznych.

2.12. Ochrona anonimowości metodą prywatności (*α*,*β*)-dystrybucyjnej

Metoda ochrony anonimowości metodą prywatności dystrybucyjnej związana jest z mechanizmem udostępniania danych. Następuje to na przykład w momencie, kiedy udostępniane są dane o identycznych chorobach pochodzące z różnych szpitali. Pojawia się wtedy pytanie, czy można udostępniać zanonimizowane dane z jednego szpitala bez brania pod uwagę danych pochodzących z innych szpitali. W przypadku takim konieczne jest nie tylko skonstruowanie odpowiednich metod anonimizacji danych, lecz również zweryfikowanie samego mechanizmu udostępniania. Od strony teoretycznej system publicznego udostępniania danych można przedstawić jako połączenie bazy danych *D* z mechanizmem udostępniania *A*. System taki spełnia kryteria prywatności (*α*,*β*)-dystrybucyjnej [31] jeżeli z prawdopodobieństwem 1-*β*, dwie *n*-elementowe bazy danych D_1, D_2 otrzymane z *D* dla dowolnego zapytania i dowolnej wartości wyjściowej spełniają warunek $p_1 \leq e^{\alpha} p_2$ [2,31] (gdzie p_1, p_2 są prawdopodobieństwami uzyskania takiej samej odpowiedzi dla takich samych zapytań, odpowiednio dla bazy D_1 oraz bazy D_2. W przypadku danych medycznych zachowanie prywatności (*α*,*β*)-dystrybucyjnej gwarantuje na przykład udostępnianie danych medycznych przez kilka szpitali w regionie bez istotnego ujawnienia, z którego szpitala dane pochodzą.

3. Operacje na danych medycznych prowadzące do ich anonimizacji

Osiągnięcie określonego stanu anonimizacji wymaga przeprowadzenia transformacji danych. Do podstawowych transformacji należą:

- tworzenie tabel pośrednich i klas równoważności atrybutów;
- uogólnianie i ograniczanie;
- anatomizacja;
- permutacja;
- zaburzenia losowe.

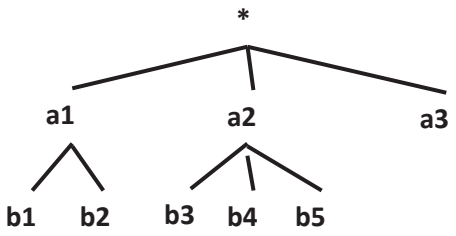
3.1. Tworzenie tabel pośrednich i klas równoważności danych

Tworzenie tabel pośrednich jest najprostszą metodą prowadzącą do odseparowania wartości przyjmowanych przez atrybuty pseudoidentyfikatora od wartości atrybutu wrażliwego. Załóżmy, że tablica *T* posiada pseudoidentyfikator *PID* oraz atrybut wrażliwy *W*. Tablicę taką można podzielić na dwie części. Część zawierającą *PID* oraz część zawierającą *W*. Aby nie doszło do utraty danych każda z części musi zawierać tyle samo rekordów co tablica *T*. Dla tablicy *T*, która zawiera klucz główny *ID* (w znormalizowanej bazie danych jest to typowo spełnione), *T*(*ID*, *PID*, *W*) najwygodniej jest dokonać podziału na T_1(*ID*, *PID*), T_2(*ID*, *W*). Stąd już bardzo prosta droga do utworzenia tablicy pośredniej T_3 łączącej ze sobą T_1 i T_2 za pośrednictwem wartości *ID*. Przyporządkowując wzajemnie jednoznacznie wartościom atrybutu ID wartości nowego atrybutu *IDP* otrzymujemy: T_1(*ID*, *PID*), T_3(*ID*, *IDP*) T_2(*IDP*, *PID*). Tabela T_2(*IDP*, *PID*) zawiera po przekształceniu zupełnie nowe wartości atrybutu kluczowego. Utrzymując w tajemnicy tabelę T_3 oraz stosując permutacyjne przekształcenie $ID \rightarrow_{PERM} IDP$, można uzyskać efekt ochrony danych wrażliwych.

Niewątpliwą wadą takiego rozwiązania jest znaczna utrata jakości informacyjnej i analitycznej w stosunku do innych metod prowadzących do anonimizacji, w tym na przykład do dosyć podobnej metody: anatomizacji.

3.2. Uogólnianie i ograniczanie danych

Uogólnienie jest operacją, która ma na celu ukrycie pewnych szczegółów w obszarze wartości pseudoidentyfikatora. Najczęściej odbywa się to niezależnie dla poszczególnych wartości atrybutów składowych *PID*. Najprostszą reprezentacją dla podstawień jest drzewo nazw. W przypadku wartości liczbowych uogólnieniem są zakresy wartości. Dla najwyższego poziomu ogólności stosuje się symbol gwiazdki *. Używając drzewa nazw wartości z danego poziomu drzewa zastępuje się wartościami przypisanymi do rodziców. Na rysunku 5 przedstawiono przykładowe trzypoziomowe drzewo nazw. Na drzewie tym widoczne jest, iż na przykład wartości atrybutów *b1* i *b2* są uogólniane wartością *a1* znajdującą się o poziom wyżej w hierarchii. Wartości z przedostatniego poziomu, na przykład *a1*, *a2*, *a3* są uogólniane przez symbol *.



Rysunek 5. Przykładowe drzewo nazw stosowane do uogólniania wartości

W przypadku uogólniania i ograniczania wartości należy zachować szczególną uwagę, gdyż przy niewłaściwycm doborze algorytmów bardzo szybko może dojść do zastąpienia wartości pierwotnych wartościami z najwyższego poziomu ogolności (n.p *). Dla danych tekstowych można tu z powodzeniem stosować metody LCS, oparte na wyznaczaniu najdłuższych wspólnych podciągów znaków.

3.3. Anatomizacja

Anatomizacja polega na podziale tabeli źródłowej *T* na dwie tabele składowe *PIDT* oraz *WT*. Tabela *PIDT* zawiera wartości pseudoidentyfikatora a tabela *WT* zawiera wartości dla atrybutu wrażliwego. Obie tabele posiadają wspólny atrybut *IG* (identyfikator grupy). Tabela *WT* zawiera wartości wrażliwego atrybutu przypisane do grup. Istotną cechą tabeli *WT* jest podanie ilości wystąpień wartości wrażliwych w grupach a nie ich pozycji, tak jak występuje to w tabeli źródłowej. Atrybut *IG* łączy ze sobą tabele. Na rysunku 6 przedstawiono przykład anatomizacji danych. Tabela *T* podzielona została na tabele *PIDT* ora *WT*. Na podstawie zanonimizowanej tabeli *T'* wyznaczono dwie grupy. Dla tych dwóch grup dokonano podziału tabeli *T*. W tabeli *PIDT* pozostały wartości pseudoidentyfikatora z tabeli *T* oraz wartości przyporządkowania do grupy. Natomiast w tabeli *WT* zawarta została ilość wystąpień wartości atrybutu wrażliwego dla każdej z grup. Warto zauważyć, iż w wyniku przeprowadzonego procesu zmniejszona została destrukcja pierwotnych wartości w porównaniu z 2-anonimizacją przeprowadzoną metodą uogólniania danych.

T

Płeć	Zawód	Miasto	Choroba
M	Inżynier	Kraków	AIDS
M	Inżynier	Kraków	AIDS
M	Inżynier	Kraków	Grypa
M	Inżynier	Kraków	AIDS
M	Malarz	Opole	Nowotwór
M	Malarz	Opole	Grypa
K	Śpiewak	Brzeg	Nowotwór
K	Tancerz	Brzeg	AIDS
M	Muzyk	Brzeg	AIDS

T' ↓ tabela 4-zanonimizowana

Płeć	Zawód	Województwo	Choroba
*	Techniczny	Małopolskie	AIDS
*	Techniczny	Małopolskie	AIDS
*	Techniczny	Małopolskie	Grypa
*	Techniczny	Małopolskie	AIDS
*	Artystyczny	Opolskie	Nowotwór
*	Artystyczny	Opolskie	Grypa
*	Artystyczny	Opolskie	Nowotwór
*	Artystyczny	Opolskie	AIDS
*	Artystyczny	Opolskie	AIDS

↓

PIDT

Płeć	Zawód	Miasto	IG
M	Inżynier	Kraków	1
M	Inżynier	Kraków	1
M	Inżynier	Kraków	1
M	Inżynier	Kraków	1
M	Malarz	Opole	2
M	Malarz	Opole	2
K	Śpiewak	Brzeg	2
K	Tancerz	Brzeg	2
M	Muzyk	Brzeg	2

WT

IG	Choroba	Ilość
1	AIDS	3
1	Grypa	1
2	AIDS	2
2	Grypa	1
2	Nowotwór	2

Rysunek 6. Ilustracja procesu anatomizacji tabeli $T \longrightarrow PIDT + WT$

3.4. Permutacja

Permutacja, podobnie jak anatomizacja, jest alternatywą wobec metod uogólniania i ograniczania danych. Klasyczna permutacja danych w tabeli zawierającej atrybut wrażliwy [10] polega na wykonaniu przestawień w obrębie bloków wartości pseudoidentyfikatora. Wartości wrażliwe pozostawia się bez zmian. Wykonywanie permutacji atrybutów pseudoidentyfikatora nie powoduje destrukcji wartości. Może jednak powodować zaburzenia w odpowiedziach na zapytania, które wiążą wartości atrybutów składowych pseudoidentyfikatora. Na rysunku 7 przedstawiono przykład praktycznego zastosowania permutacji w grupach wartości pseudoidentyfikatora.

	Płeć	Zawód	Miasto	Choroba
	M	Inżynier	Kraków	AIDS
PID_1	M	Inżynier	Kraków	AIDS
	M	Inżynier	Kraków	Grypa
	M	Inżynier	Kraków	AIDS
	M	Malarz	Opole	Nowotwór
PID_2	M	Malarz	Opole	Grypa
PID_3	K	Inżynier	Brzeg	Nowotwór
PID_4	K	Tancerz	Brzeg	AIDS
PID_5	M	Muzyk	Brzeg	AIDS

Dodatkowe źródło wiedzy

Płeć	Zawód	Miasto	Nazwisko
M	Inżynier	Kalisz	Kowalski
K	Tancerz	Brzeg	Nowak
K	Malarz	Kraków	Jagiel
M	Inżynier	Kraków	Osowski

↓

Możliwe ujawnienie danych wrażliwych

Płeć	Zawód	Miasto	Nazwisko	Choroba
K	Tancerz	Brzeg	Nowak	AIDS

Rysunek 7. Ilustracja procesu permutacji tabeli *T* względem 2 grup wartości pseudoidentyfikatora z tabeli *T'*

3.5. Zaburzenia losowe danych

Ideą wprowadzania zaburzeń losowych jest zamiana wartości atrybutów danymi wygenerowanymi syntetycznie. Nowe wartości tworzone są na podstawie wartości źródłowych. W zależności od stopnia losowości dane mogą w różnym zakresie odzwierciedlać lub przybliżać dane pochodzące z rzeczywistości. Zaburzeń losowych można dokonać przy wykorzystaniu trzech podstawowych technik: dodanie szumu, zamianę danych oraz całkowicie syntetycznego wytworzenia danych.

Najprostszą techniką jest dodanie szumu losowego. Chociaż dodawanie szumu losowego jest najprostsze w przypadku danych o charakterze liczbowym, to metodę tę można stosować też do innych postaci danych. W przypadku danych medycznych, modyfikacja za pomocą szumu losowego może dotyczyć danych laboratoryjnych, obrazów graficznych i niektórych tekstów. Odpowiednio generowany szum dosyć skutecznie zmienia wartości indywidualne pozostawiając cechy statystyczne dokumentacji.

Zamiana danych jest dokonywana pomiędzy rekordami. Najczęściej dotyczy wartości atrybutów wrażliwych. Zamiana danych jest łatwa do przeprowadzenia, jednak może dość znacznie zaburzyć wyniki statyczne. Aby ograniczyć zaburzenia, zamiana ta może być ograniczona przez wprowadzenie stopnia lub prawdopodobieństwa zmiany. Zamiana w danych medycznych, przeznaczonych do dalszych badań, powinna być stosowana ostrożnie i najlepiej z uwzględnieniem zmian zachodzących w lokalnych grupach danych po zamianie.

Całkowicie syntetyczne wytwarzanie danych umożliwia zarówno dokonywanie zamiany rzeczywistych wartości z wartościami syntetycznymi, jak również generowanie kompletnych rekordów dodatkowych. W oczywisty sposób dodawanie danych syntetycznych podnosi poziom zabezpieczenia indywidualnych danych wrażliwych. Jest też dobrym sposobem przygotowania danych testowych posiadających oczekiwany rozmiar. Generację danych zaczyna się od budowy modelu statystycznego na podstawie danych rzeczywistych, a następnie wytworzeniu punktów pasujących do tego modelu. Tak wytworzone punkty wprowadza się do tabel w miejsce danych oryginalnych lub dodaje się dodatkowe sztuczne rekordy. Alternatywną metodą w stosunku do budowy modelu jest kondensacja. Polega ona na skondensowaniu rekordów do charakterystycznych statystycznie grup, a następnie - po wyliczeniu - parametrów statystycznych, utworzeniu nowych zestawów danych zachowujących lokalne charakterystyki grup.

4. Anonimizacja dokumentacji medycznej kontrolowana przez właściciela danych

Większość metod anonimizacji nie wiąże się z zagadnieniem fundamentalnym, to jest realizacją prawa do zachowania prywatności. Wydaje się, iż to właśnie osoba, której dane dotyczą, powinna w pełni decydować o ich publicznym udostępnianiu. Próba personalizacji udostępniania danych podjęta została z wykorzystaniem metody personalizowanej prywatności w pracy [25]. W metodzie tej nie ma jednak zaimplementowanych rozwiązań związanych ze zgodą na udostępnianie danych, ale jedynie z możliwością specyfikacji wartości alternatywnych dla atrybutów wrażliwych.

Nasza propozycja rozwiązania tego problemu polega na wprowadzeniu kontroli dostępu do wszystkich spersonalizowanych danych medycznych przez pacjenta, którego dane te dotyczą. Załóżmy, że dane te mają postać jednego rekordu t_p w bazie danych. W rekordzie takim wystąpią dane osobowe oraz dane wrażliwe związane z chorobą i jej leczeniem. Dla uproszczenia niech tabela *T* zawiera atrybuty {*Id*, *Nazwisko*, *Zawód*, *Miasto*, *Choroba*}. Maksymalna

informacja o wyróżnionym pacjencie będzie wtedy, gdy wszystkie pola będą zawierać rzeczywiste wartości atrybutów. Minimalna informacja o pacjencie będzie wtedy, gdy jako wartości we wszystkich polach wystąpią wartości ogólne *. Gdy dane są rzeczywiste, pacjent w zasadzie nie kontroluje swojej prywatności, natomiast gdy dane są całkowicie uogólnione są one bezużyteczne dla odbiorcy publicznego. Najprostszą metodą pozostawienia danych w bazie i sprawowanie nad nimi kontroli jest zastosowanie metod kryptograficznych, na przykład szyfrowania. Na rysunku 8 przedstawiono przykładową tabelę zawierającą dane dla kilku pacjentów, w tym dla pacjenta wyróżnionego $Id=t_p$. Poniżej tej tabeli pokazano tabelę zawierającą: rekord pacjenta z całkowicie ogólnymi wartościami i rekord pacjenta z wartościami zaszyfrowanymi.

Id	Nazwisko	Zawód	Miasto	Choroba
1	Kowalski	Inżynier	Kraków	AIDS
2	Nowak	Malarz	Opole	Nowotwór
3	Kwiatkowski	Inżynier	Brzeg	Nowotwór
4	Pyc	Malarz	Opole	Grypa
5	Kociniak	Inżynier	Kraków	AIDS
6	Wielgosz	Inżynier	Kraków	Grypa
7	Aster	Tancerz	Brzeg	AIDS
8	Lewkowicz	Muzyk	Brzeg	AIDS
9	Spytek	Inżynier	Kraków	AIDS

Id	Nazwisko	Zawód	Miasto	Choroba
1	Kowalski	Inżynier	Kraków	AIDS
2	Nowak	Malarz	Opole	Nowotwór
3	Kwiatkowski	Inżynier	Brzeg	Nowotwór
4	Pyc	Malarz	Opole	Grypa
5	*	*	*	*
6	Wielgosz	Inżynier	Kraków	Grypa
7	Aster	Tancerz	Brzeg	AIDS
8	Lewkowicz	Muzyk	Brzeg	AIDS
9	Spytek	Inżynier	Kraków	AIDS

Id	Nazwisko	Zawód	Miasto	Choroba
1	Kowalski	Inżynier	Kraków	AIDS
2	Nowak	Malarz	Opole	Nowotwór
3	Kwiatkowski	Inżynier	Brzeg	Nowotwór
4	Pyc	Malarz	Opole	Grypa
5	AAFGHA	BCCHFF	DFSCXX	MBJKLO
6	Wielgosz	Inżynier	Kraków	Grypa
7	Aster	Tancerz	Brzeg	AIDS
8	Lewkowicz	Muzyk	Brzeg	AIDS
9	Spytek	Inżynier	Kraków	AIDS

Rysunek 8. Przykładowe tabele z wyróżnionym pacjentem tp=5 zawierające kolejno: rekord z danymi rzeczywistymi, rekord z danymi maksymalnie uogólnionymi, rekord zawierający dane rzeczywiste zaszyfrowane

To oczywiste iż, aby kontrola mogła być sprawowana przez pacjenta parametry szyfrowania powinny być przez niego ustalane. Wprowadzenie prostego rozwiązania z szyfrowaniem wartości jest nie do przyjęcia w praktyce. Obarczone jest bowiem szeregiem wad:

1) zaszyfrowane wartości są trudne do odróżnienia od wartości rzeczywistych i wartości wprowadzanych podczas procesu anonimizacji, co może prowadzić do zaburzeń w przetwarzaniu danych oraz zaburzeń wyników analizy danych zawierających szyfrowane rekordy;

2) brak niezależnego od pacjenta dostępu do danych rzeczywistych przez lekarza i przez osoby oraz instytucje prawnie umocowane do takiego dostępu;

3) problemy związane z utratą kluczy szyfrowania bądź ich wymianą;

4) problemy związane z wygasaniem uprawnień do korzystania z danych pacjenta przez instytucje i lekarzy.

Rozwiązanie wyszczególnionych wyżej problemów wymaga przyjęcia wygodnej notacji, pozwalającej na filtrowanie wartości zaszyfrowanych oraz odpowiedniego algorytmu szyfrowania danych i wymiany klucza.

Do selekcji wartości atrybutów zaszyfrowanych proponuje się zastosować znaczniki XML wyróżniające część zaszyfrowaną. Rozwiązanie to może mieć postać: <UCRYPT> zaszyfrowana_wartość_atrybutu </UCRYPT>, gdzie <UCRYPT> oraz </UCRYPT> są odpowiednio elementami rozpoczynającymi i kończącymi sekcję zaszyfrowanej wartości. Element UCRYPT może mieć dodatkowe argumenty na przykład do wskazania rodzaju zastosowanego szyfru. Może to wyglądać np. w sposób: <CRYPT cipher="szyfr1">zaszyfrowana_wartość_atrybutu</UCRYPT>. Proponowane rozwiązanie jest zgodne z obowiązującymi tendencjami opisu danych w języku XML. Przykładowa tabela po zastosowaniu proponowanego rozwiązania będzie wyglądać jak na rysunku 9.

Id	Nazwisko	Zawód	Miasto	Choroba
1	Kowalski	Inżynier	Kraków	AIDS
2	Nowak	Malarz	Opole	Nowotwór
3	Kwiatkowski	Inżynier	Brzeg	Nowotwór
4	Pyc	Malarz	Opole	Grypa
5	<UCRYPT> AAFGHA </UCRYPT>	<UCRYPT> BCCHFF </UCRYPT>	<UCRYPT> DFSCXX </UCRYPT>	<UCRYPT> MBJKLO </UCRYPT>
6	Wielgosz	Inżynier	Kraków	Grypa
7	Aster	Tancerz	Brzeg	AIDS
8	Lewkowicz	Muzyk	Brzeg	AIDS
9	Spytek	Inżynier	Kraków	AIDS

Rysunek 9. Widok zawartości tabeli z wartościami kontrolowanymi przez użytkownika oznaczone znacznikami <UCRYPT>

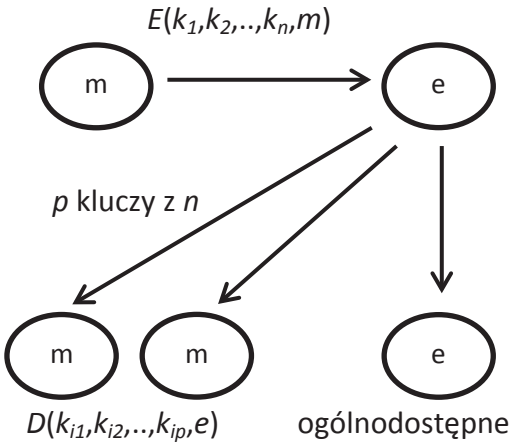
Zaproponowane rozwiązanie pozwala na łatwe wydzielenie rekordów zastrzeżonych przez pacjenta na drodze zapytań do bazy danych. Wystarczające jest ustawienie filtru eliminującego pozycje z <UCRYPT>. Poza tym jest to zgodne ze standardem XML i nie wymaga dodatkowej obsługi przy eksporcie i imporcie

bazy danych do formatu XML (na przykład dla prostego zapewnienia interoperacyjności pomiędzy różnymi systemami, wykonania kopii zapasowej danych wraz ze strukturą). Oczywiście nazwa elementu <UCRYPT> może być inna i dostosowana do potrzeb systemu baz danych lub systemu przetwarzania dokumentów.

Trudniejsza sytuacja występuje przy konstrukcji odpowiedniego schematu szyfrowania. Wydaje się, iż przyjęcie najprostszego rozwiązania w postaci schematu szyfrowania z kluczem prywatnym jest nieodpowiednie. Wynika to z faktu, iż do szyfrowania i deszyfrowania stosowany jest jeden klucz po udostępnieniu, a pacjent w zasadzie traci możliwość zarządzania udostępnianiem swoich danych. Klucz prywatny można kopiować, a każdy jego posiadacz jest nierozróżnialny. Prosty schemat z kluczem prywatnym nie pozwala na zmianę listy uprawnionych podmiotów, bez konieczności zmiany klucza i rozesłania wszystkim klucza zmienionego. Schemat szyfrowania z kluczem publicznym zawiera dwa klucze. Z reguły jeden z nich jest używany do szyfrowania a drugi do deszyfrowania. To rozwiązanie jest trochę lepsze. Pacjent może bowiem zaszyfrować dane jednym kluczem, a do odczytu danych - rozesłać drugi klucz. Wszyscy dopuszczeni użytkownicy będą wtedy posiadać jednakowy klucz i przez to nie można ich będzie rozróżnić, w przypadku chociażby wybiórczego pozbawinia jednego z użytkowników dostępu do danych rzeczywistych (na przykład, gdy pacjent zmieni swojego lekarza lub zmieni się osoba upoważniona do kontroli danych i procedur medycznych).

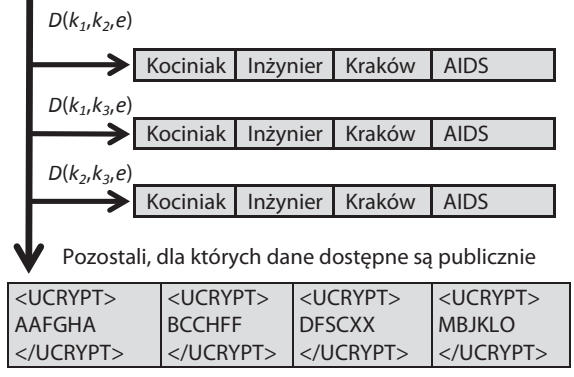
Proponuje się zastosowanie kryptograficznego rozwiązania, w którym schemat szyfrowania oparty jest na n kluczach. W tym przypadku, aby odszyfrować dane medyczne konieczne będzie posiadanie p dowolnych kluczy z n możliwych. Wyobraźmy sobie, iż schemat szyfrowania posiada n=3 klucze, z których wystarczą p=2 do odszyfrowania wartości. Przy takim schemacie klucz k_1 może posiadać pacjent, klucz k_2 - lekarz, klucz k_3 -płatnik świadczenia zdrowotnego. Odczyt danych będzie możliwy dla trzech konfiguracji kluczy. Dla pary pacjent-lekarz (k_1,k_2), pacjent-płatnik (k_1,k_3) oraz lekarz-płatnik (k_2,k_3). Widoczne jest, iż zastosowanie takiego rozwiązania może zadowolić wszystkie strony uprawnione do korzystania z danych. Pacjent zabezpieczony jest przed udostępnieniem danych wrażliwych publicznie, lekarz oraz płatnik współdziałając razem mogą zapoznać się z danymi rzeczywistymi, mimo odmowy lub śmierci pacjenta. Zwiększenie liczby kluczy n pozwala na dołączenie innych podmiotów z uprawnieniami odczytu danych rzeczywistych. Szczegóły matematyczne proponowanego rozwiązania wykraczają poza ramy niniejszej pracy.

Na rysunku 10 przedstawiono propozycję opisywanego schematu szyfrowania. Na rysunku 11 przedstawiono ilustrację procesu udostępniania zabezpieczonego rekordu.



Rysunek 10. Ilustracja działania wielokluczowego systemu proponowanego do zarządzania udostępnianiem danych wrażliwych, m - szyfrowana wartość atrybutu wrażliwego, k_1, k_2, k_n, k_{i1}, k_{i2}, k_{ip} – klucze, e – zaszyfrowana wartość

Id	Nazwisko	Zawód	Miasto	Choroba
1	Kowalski	Inżynier	Kraków	AIDS
2	Nowak	Malarz	Opole	Nowotwór
3	Kwiatkowski	Inżynier	Brzeg	Nowotwór
4	Pyc	Malarz	Opole	Grypa
5	<UCRYPT> AAFGHA </UCRYPT>	<UCRYPT> BCCHFF </UCRYPT>	<UCRYPT> DFSCXX </UCRYPT>	<UCRYPT> MBJKLO </UCRYPT>
6	Wielgosz	Inżynier	Kraków	Grypa
7	Aster	Tancerz	Brzeg	AIDS
8	Lewkowicz	Muzyk	Brzeg	AIDS
9	Spytek	Inżynier	Kraków	AIDS



Rysunek 11. Ilustracja procesu udostępniania rekordu zabezpieczającego dane zabezpieczone przed ujawnieniem. Dostęp mają tylko pary użytkowników posiadających p=2 kluczy z n=3 możliwych

5. Wnioski

W pracy przeprowadzono analizę rozwiązań w zakresie anonimizacji danych pod kątem zastosowania ich do ochrony prywatności w zbiorach danych medycznych. Przeprowadzono analizę takich metod, jak: k-anonimizacji, (X,Y)-anonimizacji, (α,k)-anonimizacji, (k,e)-anonimizacji, (X,Y)-dołączalności, (X,Y)-prywatności, LKC-prywatności, l-dywersyfikacji, t-bliskości, ograniczonego zaufania oraz personalizowanej prywatności.

Wszystkie przedstawione metody nadają się do anonimizacji danych medycznych. W porównaniu do prostych metod opartych na pseudonimizacji umożliwiają one ograniczenie destrukcji danych i lepszą kontrolę nad powiązaniami pomiędzy wartościami atrybutów. Na szczególną uwagę zasługuje metoda personalizowanej prywatności. W metodzie tej możliwy jest wpływ pacjenta na wartość przyjmowaną przez atrybut wrażliwy. Wpływ ten jest jednak ograniczony do podania alternatywnych wartości tylko dla chronionego pola. Oprócz samej anonimizacji w procesie ochrony prywatności istotną rolę odgrywają metody ochrony zanonimizowanych danych. Większość tych metod oparta jest na wyznaczaniu charakterystyk probabilistycznych. Przeprowadzono analizę metod ochrony zanonimizowanych danych: δ-obecności, prywatności e-różnicowej, (d,γ)-prywatności, prywatności (α,β)-dystrybucyjnej oraz ochrony przed (c,t)-izolacją. Metody te stanowią dodatkową ochronę danych wrażliwych. Część z nich w istotnym stopniu ogranicza możliwości pozyskiwania dodatkowej wiedzy przez obserwatora. Może to być bardzo istotne w systemach, w których często dochodzi do dołączania lub kasowania rekordów.

W pracy zaproponowano nowe rozwiązanie w zakresie kontrolowanej ochrony prywatności. Rozwiązanie to oparte jest na wydzieleniu chronionych pól i wielokluczowym szyfrowaniu i deszyfrowaniu danych wrażliwych. Wydzielenie pól zaproponowano w postaci sekwencji dodania elementu <CRYPT> wskazującego na zabezpieczaną wartość. Sposób wydzielenia wartości jest zgodny ze standardem XML. Pola poddawane indywidualnej ochronie są szyfrowane. Do szyfrowania wybrany został schemat szyfrowania posiadający n różnych kluczy. Do deszyfrowania zawartości wystarczające jest p spośród wszystkich możliwych. Umożliwia to zbudowanie systemu dostępu do danych wrażliwych, w którym ujawnienie następuje dla grupy posiadającej p kluczy. Przedstawione zabezpieczenie umożliwia zastosowanie dotychczasowych metod anonimizacji w przypadku, gdy wrażliwe pola są udostępniane lub następuje wykorzystanie zupełnie nowego modelu dostępu. Przy $n > 2$ i $p=2$ możliwe jest, na przykład korzystanie z danych wrażliwych w parach lekarz-pacjent, pacjent-płatnik świadczenia, lekarz-płatnik świadczenia. Dzięki nadmiarowości powstającej w przypadku, gdy $n>p$ istnieje możliwość wyłączania i dołączania uprawnionych użytkowników. Może być to bardzo użyteczne na przykład po zmianie przez pacjenta lekarza prowadzącego. System zabezpiecza również uprawniony dostęp po śmierci pacjenta przez p zaufanych użytkowników posiadających ważne klucze. Obecnie autorzy prowadzą badania nad konstrukcją specjalizowanych metod anonimizacji dobrze przystosowanych do zaproponowanego rozwiązania.

Piśmiennictwo

1. Liber A. Problemy anonimizacji dokumentów medycznych. Część 1. Wprowadzenie do anonimizacji danych medycznych. Zapewnienie ochrony danych wrażliwych metodami f(a)- i f(a,b)-anonimizacji. *Puls Uczelni* 2014; 1: 13-21.
2. Fung BCM, Wang Ke, Fu Wai-Chee A, Yu PS. *Introduction to Privacy-Preserving Data Publishing*. New York: CRS PRESS; 2011.
3. Borucki B. Metodyka ochrony poufności i bezpieczeństwa medycznych danych osobowych. *Ultrasonografia* 2009; 36: 9-20.
4. Samarati P. Protecting respondents identities in microdata release. *IEEE TKDE* 2001; 13(6): 1010-1027.
5. Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. Proceedings of the 17th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems. Seattle; 1998: 188-202.
6. Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report. SRI International; 1998: 1-19.
7. Sweeney L. k-Anonymity: A model for protecting privacy. *Int J Uncertain Fuzz* 2002; 10(5): 557-570.
8. Wang K, Fung BCM. Anonymizing sequential releases. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia; 2006: 414-423.
9. Wong R, Chi-Wing Li, Jiuyong Fu, Ada Wai-Chee, Wang K. (α, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing. SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia; 2006: 754-759.
10. Zhang Q, Koudas N, Srivastava D, Yu T. Aggregate query answering on anonymized tables. InProc. of the 23rd IEEE International Conference on Data Engineering (ICDE), April 2007.
11. Truta TM, Campan A. Avoiding Attribute Disclosure with the (Extended) p-Sensitive k-Anonymity Model. Data Mining. *AoIS* 2010: 353-373.
12. Friedman A, Wolff R, Schuster A. Providing k-Anonymity in Data Mining. *The VLDB Journal* 2008; 17(4): 789-804.
13. Denning DE. Secure Statistical Databases with Random Sample Queries. Purdue University. CSD-TR-302; 1979: 291-315.
14. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. Proceedings of the 22-nd IEEE International Conference on Data Engineering. Atlanta; 2006: Art. 1-52.
15. Prasad A, Panda GK, Mitra A, Singh A, Gour D. Applying l-Diversity in anonymizing collaborative social network. *IJCSIS* 2010; 8: 324-329.
16. Kern M. Anonymity: A Formalization of Privacy – l-Diversity. Proceedings of the Seminars Future Internet, Innovative Internet Technologies and Mobile Communications (IITM) and Autonomous Communication Networks (ACN), volume NET-2013-08-1 of Network Architectures and Services. Munich; 2013: 49-56.
17. Auer P. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *JMLR* 2002; 3: 397-422.
18. Wang K, Fung BCM, Yu PS. Handicapping attacker's confidence: An alternative to k-anonymization. *KAIS* 2007; 11(3): 345-368.
19. Veeningen M, De Weger B, Zannone N. Symbolic Privacy Analysis through Linkability and Detectability. Trust Management VII, *IFIP AICT* 2013 ; 401: 1-16.
20. Shabtai A. i wsp. A Survey of Data Leakage Detection and Prevention Solutions. *Springer Briefs in Com Sci* 2012: 47-68.
21. Mohammed N, Fung BCM, Hung PCK, Lee CK. Anonymizing healthcare data: A case study on the blood transfusion service. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris; 2009: 1285-1293.
22. Loukides G, Shao J. Preventing range disclosure in k-anonymised data. *Expert Systems Appl* 2011; 38(4): 4559-4574.
23. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. Proceedings of the 21st IEEE International Conference on Data Engineering. Istanbul; 2007: 106-115.

24. Yuan M, Chen L, Yu PS. Personalized Privacy Protection in Social Networks. Proceedings of the VLDB Endowment. 37th International Conference on Very Large Data Bases. 4(2). Seattle; 2010: 141-150.
25. Ercan Nergiz M, Atzori M, Clifton CW. Hiding the presence of individuals from shared databases. Proceedings of ACM International Conference on Management of Data (SIGMOD). Vancouver; 2007: 665-676.
26. Chawla S, Dwork C, McSherry F, Smith A, Wee H. Toward privacy in public databases. Proceedings of Theory of Cryptography Conference. Cambridge; 2005: 363-385.
27. Beauxis R, Palamidessi C. Probabilistic and nondeterministic aspects of anonymity. *Theor Com Sci* 2009; 410: 4006-4025.
28. Dwork C. Differential privacy. Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP). Venice; 2006: LNCS 4052, 1-12.
29. Dwork C. Differential privacy: A survey of results. In Proceedings of the 5th International Conference on Theory and Applications of Models of Computation (TAMC). Xian; 2008: LNCS 4978, 1-19.
30. Rastogi V, Suciu D, Hong S. The boundary between privacy and utility in data publishing. Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB). Vienna; 2007: 531-542.
31. Blum A, Ligett K, Roth A. A learning theory approach to non-interactive database privacy. Proceedings of the 40th annual ACM Symposium on Theory of Computing (STOC). Victoria; 2008: Art. 12, 1-25.

Adres do korespondencji
dr inż. Arkadiusz Liber
Politechnika Wrocławska
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
Tel. +48 713 203 207
E-mail: arkadiusz.liber@pwr.wroc.pl

Praca wpłynęła do redakcji: 21.02.2014
Po recenzji: 02.03.2014
Zaakceptowana do druku: 03.03.2014