Państwowa Medyczna Wyższa Szkoła Zawodowa w Opolu

ISSN 2080-2021

Prace oryginalne - Original papers

PROBLEMY ANONIMIZACJI DOKUMENTÓW MEDYCZNYCH. CZĘŚĆ 1. WPROWADZENIE DO ANONIMIZACJI DANYCH MEDYCZNYCH. ZAPEWNIENIE OCHRONY DANYCH WRAŻLIWYCH METODAMI F(A)- I F(A,B)-ANONIMIZACJI

The issues connected with the anonymization of medical data. Part 1. The introduction to the anonymization of medical data. Ensuring the protection of sensitive information with the use of such methods as f(a) and f(a,b)

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)

Summary

Introduction: Medical documentation must be protected against damage or loss, in compliance with its integrity and credibility and the opportunity to a permanent access by the authorized staff and, finally, protected against the access of unauthorized persons. Anonymization is one of the methods to safeguard the data against the disclosure.

Aim of the study: The study aims at the analysis of methods of anonymization, the analysis of methods of the protection of anonymized data and the study of a new security type of privacy enabling to control sensitive data by the entity which the data concerns.

Material and methods: The analytical and algebraic methods were used.

Results: The study ought to deliver the materials supporting the choice and analysis of the ways of the anonymization of medical data, and develop a new privacy protection solution enabling the control of sensitive data by entities whom this data concerns.

Conclusions: In the paper, the analysis of solutions of data anonymizing used for medical data privacy protection was conducted. The methods, such as k-Anonymity, (X,Y)- Anonymity, (a,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 to control sensitive data by their owners was also conducted. Apart from the existing methods of the anonymization, the analysis of methods of the anonimized data protection was conducted, in particular the methods of: d-Presence, e-Differential Privacy, (d,g)-Privacy, (a,b)-Distributing Privacy and protections against (c,t)-Isolation were analyzed. The author introduced a new solution of the controlled protection of privacy. The solution is based on marking a protected field and multi-key encryption of the sensitive value. The suggested way of fields marking is in accordance to the XML standard. For the encryption (n,p) different key cipher was selected. To decipher the content the p keys of n is used. The proposed solution enables to apply brand new methods for the control of privacy of disclosing sensitive data.

Keywords: data anonymization, health documents, privacy in health care, owner controlled access to medical data, multi key cryptography

Streszczenie
Wstęp: Dokumentacja medyczna powinna być zabezpieczona przed uszkodzeniami lub utratą. Sposób zabezpieczenia musi uwzględniać zachowanie integralności i wiarygodności oraz zapewniać stały dostęp do dokumentacji osobom uprawnionym, a także uniemożliwiać dostęp osobom nieuprawnionym. Dokumentację medyczną powinno się udostępniać z zachowaniem jej integralności oraz ochrony danych osobowych. Jednym ze sposobów zabezpieczenia danych przed ujawnieniem jest anonimizacja.
Cel badań: 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 umożliwiającego 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 oraz ich udostępniania.
Słowa kluczowe: anonimizacja danych, dokumentacja medyczna, prywatność w ochronie zdrowia, kontrola właściciela w udostępnianiu danych medycznych, kryptografia wielokluczowa

1. Wprowadzenie

W związku z rozwojem elektronicznych mediów komunikacyjnych oraz systemów przetwarzania informacji obserwuje się coraz większe zainteresowanie gromadzeniem i przetwarzaniem danych o charakterze osobowym. Gromadzeniem takich danych zainteresowane są zarówno instytucje państwowe, jak i przedsiębiorstwa prywatne. Zgodnie z uregulowaniami prawnymi wiele danych jest nie tylko gromadzonych, ale również udostępnianych publicznie. Ze względu na funkcjonowanie nowoczesnego państwa oraz zapewnienie prawidłowego obrotu gospodarczego zaprzestanie gromadzenia danych osobowych jest praktycznie niemożliwe. Sytuacja taka prowadzi do szybkiego ograniczenia prywatności lub do całkowitej jej utraty. Należy tu podkreślić fakt, iż szybkie narastanie problemu związane jest w zasadzie z przetwarzaniem materiałów w postaci elektronicznej. Dokumenty klasyczne, w postaci zapisów na papierze lub innym materiale posiadającym określone rozmiary, masę i kształt były dosyć trudne do rozpowszechniania a czas na to potrzebny był bardzo długi. Istotnie ograniczało to problemy związane z naruszaniem prywatności jednostki, nie tylko ze względu na trudności techniczne i koszty rozpowszechniania informacji, lecz również z powodu długiego czasu, jaki upływał pomiędzy zebraniem informacji a jej rozpowszechnieniem. Opóźnienie takie stanowiło dodatkowe zabezpieczenie prywatności jednostki, dając jej realną możliwość zabezpieczenia się przed skutkami niechcianego lub wrogiego zachowania. Problem naruszania prywatności wydaje się wymykać spod kontroli. Oprócz danych gromadzonych przez instytucje zobligowane do przestrzegania uregulowań prawnych, coraz więcej jest danych pozyskiwanych, przetwarzanych i publikowanych w sposób niekontrolowany. Należą do nich: dane gromadzone w serwisach społecznościowych, na blogach i forach, w ogólnodostępnych repozytoriach pamięciowych czy nawet w indywidualnych serwisach internetowych. Dane tam gromadzone wprowadzane są z prywatnych aparatów cyfrowych, z kamer monitoringu, z samochodowych rejestratorów wideo. Uzupełniane są one o czas rejestracji, dane geolokalizacyjne, opisy i wpisy osób rejestrujących zdarzenia oraz osób, które rozpoznają na fotografiach swoich znajomych, znajome miejsca czy znajome przedmioty. Raz opublikowane w Internecie dane są trudne lub niemożliwe do całkowitego usunięcia. Wynika to z faktu, iż zawartość Internetu jest w sposób ciągły przeszukiwana i indeksowana przez specjalnie do tego celu skonstruowane maszyny i programy. Indeksowana zawartość jest zapamiętywana i można się do niej dostać na wiele sposobów, na przykład za pomocą waybackmachine (http://archive.org/web/). W niniejszej pracy przedstawiono problemy związane z zachowaniem prywatności w odniesieniu do dokumentacji medycznej. Wiadomości zawarte w takiej dokumentacji to nie tylko dane identyfikujące pacjenta, lecz również informacje o jego dolegliwościach, sposobach leczenia, miejscach świadczenia usług zdrowotnych, danych personelu medycznego. Jednym ze sposobów ochrony prywatności jest anonimizacja dokumentów. Anonimizacja taka staje się coraz powszechniejsza nie tylko w odniesieniu do dokumentacji medycznej, ale również innych ważnych dokumentów, takich jak na przykład decyzje urzędowe czy sentencje wyroków i akta postępowań sądowych.

2. Dokument, dokumentacja medyczna oraz ich ochrona

Dokumenty pełnią bardzo ważną rolę informacyjną i prawną we wszystkich nowoczesnych społeczeństwach. W ogólności dokument można zdefiniować jako rzeczowe świadectwo jakiegoś zjawiska sporządzone w formie właściwej dla danego miejsca i czasu.

Taka definicja dokumentu nie oddaje jednak jego roli społecznej oraz znaczenia prawnego.

2.1. Definicja dokumentu w kontekście jego roli prawnej

Dużo lepszą definicją dokumentu wydaje się być definicja stosowana w postępowaniu karnym oraz w kryminalistyce. Zgodnie z nią dokumentem jest każdy przedmiot lub inny zapisany nośnik informacji, z którym związane jest określone prawo, albo który - ze względu na zawartą w nim treść - stanowi dowód prawa, stosunku prawnego lub okoliczności mającej znaczenie prawne. Definicja ta obejmuje zarówno dokumenty w postaci klasycznej, jak i dokumenty sporządzone w formie elektronicznej. Dokumenty mogą zawierać dane o charakterze osobowym.

2.2. Dane osobowe, identyfikacja osoby

Formalnie za dane osobowe uznaje się dane charakteryzujące osobę zidentyfikowaną lub dane pozwalające na zidentyfikowanie osoby fizycznej. Za osobę możliwą do zidentyfikowania uznaje się taką, której tożsamość można określić bezpośrednio lub pośrednio. W szczególności uznaje się, iż osobę można zidentyfikować przez powołanie się na jej numer identyfikacyjny albo jeden lub kilka czynników specyficznych, określających jej cechy fizyczne, fizjologiczne, umysłowe, ekonomiczne, kulturowe lub społeczne. W rzeczywistych uregulowaniach prawnych przyjmuje się, że informacji nie uważa się za umożliwiającą określenie tożsamości osoby, jeżeli wymagałoby to nadmiernych kosztów.

Występują dwa rodzaje identyfikacji: identyfikacja bezpośrednia oraz identyfikacja pośrednia. Identyfikacja bezpośrednia zachodzi wówczas, gdy osoba jest identyfikowana jednoznacznie bezpośrednio na podstawie posiadanych danych. W przypadku identyfikacji pośredniej jednoznaczna identyfikacja następuje na podstawie relacji pomiędzy posiadanymi danymi lub z wykorzystaniem informacji zewnętrznych. Identyfikacja bezpośrednia następuje nie tylko przez podanie imienia, nazwiska i adresu zamieszkania. Osobę mogą bezpośrednio identyfikować daty (na przykład data urodzenia, data wstąpienia do organizacji itp.), charakterystyczne numery (np. ubezpieczenia, kont bankowych, pojazdów itp.). Identyfikacja pośrednia wymaga wnioskowania. W praktycznych rozwiązaniach wnioskowanie może zachodzić automatycznie na przykład w serwisach społecznościowych, w maszynach i robotach wyszukujących, na drodze obliczeniowej poprzez zadawanie sekwencji pytań do ogólnodostępnych baz danych. Bardzo ciekawymi danymi są obrazowe dane medyczne. Na ich podstawie może nastąpić nie tylko rekonstrukcja twarzy, ale pozwolą one identyfikować osoby pośrednio na podstawie charakterystycznych zmian obrazowych i wiązania ich z innymi posiadanymi już danymi.

2.3. Dokumentacja medyczna

Dokumentację medyczną stanowi zbiór dokumentów spełniających ściśle określone warunki techniczne i prawne. Mogą one mieć postać papierową, elektroniczną lub mieszaną. Dokumentacja może się odnosić do pojedynczego pacjenta korzystającego ze świadczeń medycznych (dokumentacja indywidualna) lub może się odnosić do zbioru pacjentów (minimum dwa elementy stanowią dokumentację zbiorczą). Dokumentacja medyczna indywidualna składa się z dwóch części: dokumentacji indywidualnej wewnętrznej oraz dokumentacji indywidualnej zewnętrznej. Dokumentacja indywidualna wewnętrzna jest przeznaczona na potrzeby podmiotu udzielającego świadczenia zdrowotnego. Dokumentacja indywidualna zewnętrzna - na potrzeby pacjenta korzystającego ze świadczeń zdrowotnych. Istotną cechą dokumentacji medycznej zewnętrznej jest fakt, iż stanowią ją między innymi dokumenty przekazywane za pośrednictwem pacjenta, innym instytucjom. Od strony bezpieczeństwa danych oraz zapewnienia ich prywatności istotne jest nie tylko właściwe przetwarzanie dokumentacji wewnętrznej, ale również zawartość i sposób przetwarzania dokumentacji zewnętrznej. Dokumentacja zewnętrzna z reguły jest przekazywana innym podmiotom. W związku z tym na bezpieczeństwo danych oraz zapewnienie prywatności wpływa działanie więcej niż jednego systemu przetwarzania oraz bezpieczeństwo kanału przesyłania dokumentacji medycznej zewnętrznej. Dokumentacja medyczna indywidualna zawiera: oznaczenie podmiotu, pacjenta, osoby udzielającej świadczeń, datę wpisu, informacje dotyczące stanu zdrowia i choroby oraz procesów, jakim poddawany był pacjent. Istotne jest tu to, iż w oznaczeniu pacjenta zawarte są istotne dane identyfikacyjne w postaci: imienia i nazwiska, daty urodzenia, płci, adresu zamieszkania, numeru identyfikacyjnego. Przy badaniu metod anonimizacji danych medycznych często zapomina się o danych osób udzielających świadczeń. Dane te mogą również być wykorzystane do ataku na zbiory poddane anonimizacji.

Dokumentacja medyczna jest bardzo ważnym przykładem dokumentacji zawierającej dane wrażliwe, stanowiące grupę szczególnie chronionych danych osobowych. Oprócz danych o pochodzeniu rasowym lub etnicznym, danych dotyczących poglądów politycznych, filozoficznych, przynależności wyznaniowej, partyjnej lub związkowej, danych o skazaniach i karach w postepowaniu administracyjnym i sądowym, danych o kodzie genetycznym, nałogach i życiu seksualnym, dane wrażliwe stanowią również dane medyczne.

2.4. Ochrona dokumentów i dokumentacji medycznej

Ze względu na swój charakter prawny dokumenty powinny być przetwarzane i przechowywane ze szczególną starannością. Fałszowanie i niszczenie dokumentów jest przestępstwem i podlega ściganiu. To samo dotyczy również zbioru dokumentów, jakim jest dokumentacja medyczna. W celu zabezpieczenia danych i dokumentów w systemach teleinformatycznych stosuje się typowe kryptograficzne i proceduralne systemy zabezpieczeń. Środki te z reguły nie rozwiązują problemu zabezpieczenia prywatności osób, których dane są przetwarzane. Do takich celów stosowane są dodatkowe rozwiązania, takie jak: depersonalizacja, anonimizacja czy pseudonimizacja.

Dokumentacja medyczna [1] musi być zabezpieczona przed uszkodzeniami lub utratą. Sposób zabezpieczenia musi uwzględniać zachowanie integralności i wiarygodności oraz zapewniać stały dostęp do dokumentacji osobom uprawnionym oraz uniemożliwiać dostęp osobom nieuprawnionym. Dokumentację medyczną powinno się udostępniać z zachowaniem jej integralności oraz ochrony danych osobowych. Niestety, przy danych osobowych następuje sprzeczność interesów pomiędzy interesami instytucji starającymi się pozyskać jak najszerszy dostęp do danych a interesami jednostek zmierzających do minimalizacji rozprzestrzeniania się danych oraz zachowania prywatności. Poszukiwane są rozwiązania zapewniające utrzymanie równowagi pomiędzy gromadzeniem i przetwarzaniem danych a zachowaniem prywatności. W chwili obecnej coraz więcej danych posiada postać elektroniczną. Postać ta jest dogodna do szybkiego przetwarzania maszynowego, umożliwiającego poszukiwanie w danych związków trudnych do znalezienia w przypadku dokumentacji tradycyjnej. Coraz większym zainteresowaniem cieszy się personalizacja informacji oraz wyszukiwanie związków o charakterze osobowym w ogólnodostępnych zasobach.

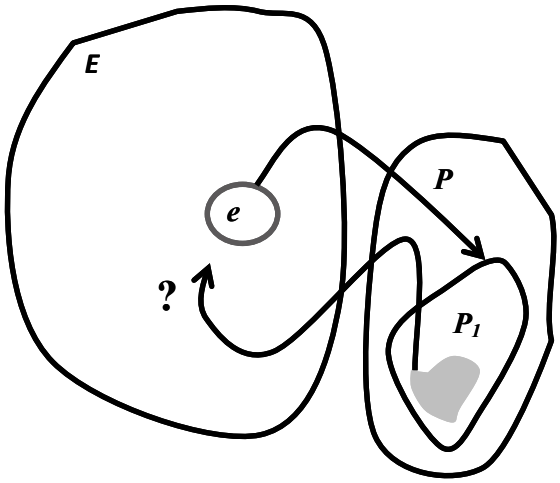
3. Ochrona prywatności

Zgodnie z obowiązującymi przepisami każda osoba ma prawo do ochrony życia prywatnego, rodzinnego, czci i dobrego imienia oraz decydowania o swoim życiu osobistym. Zapisy tego typu są charakterystyczne dla aktów prawnych o charakterze ustawy zasadniczej. W praktyce kształt ochrony prywatności kreowany jest przez akty wykonawcze, które dziedzinowo ograniczają ogólne prawo do prywatności. Ich ewolucja zmierza w kierunku zapisów o charakterze fakultatywnym, nie dając bezwarunkowej ochrony prywatności.

Coraz większego znaczenia nabiera proces udostępniania i zarządzania informacjami o charakterze osobistym. Raz wprowadzona do systemu informatycznego informacja może być bardzo trudna lub wręcz niemożliwa do usunięcia. Jednym ze sposobów zabezpieczenia danych przed ujawnieniem jest anonimizacja. Anonimizacja jest procesem, który pozwala osiągnąć stan anonimowości.

3.1. Anonimowość

Z formalnego punktu widzenia anonimowość jest cechą elementu *e* pewnego zbioru *E*. Elementy tego zbioru posiadają pewne cechy należące do zbioru cech *P*. Załóżmy, że element *e* posiada zbiór cech P_1. Jeżeli na podstawie podzbioru cech P_1 ze zbioru cech *P* nie można wskazać elementu *e,* to element ten nazywamy elementem niezidentyfikowanym i często mówimy, iż jest to element anonimowy.



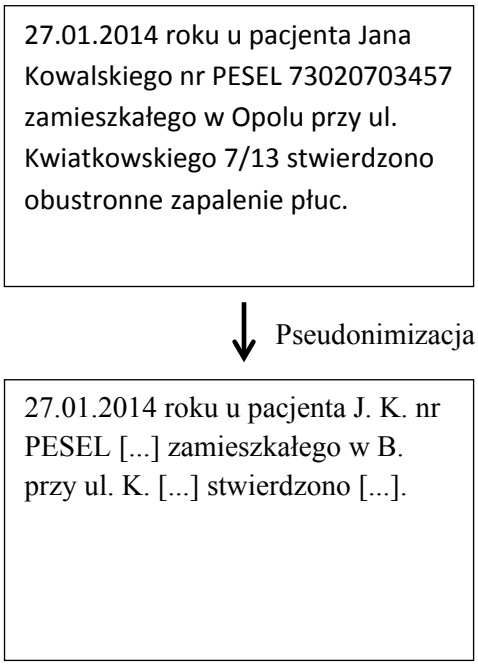
Ryc. 1. Ilustracja problemu anonimowości elementu *e* posiadający cechy należące do zbioru P_1. Oznaczenia: *E* – zbiór elementów *e* dla których rozpatrujemy cechy poddawane anonimizacji, *P* – zbiór wszystkich cech, P_1 – podzbiór cech które posiada element *e*. Część cech z podzbioru P_1 ulega anonimizacji

3.2. Anonimizacja

Anonimizacja elementu *e* jest tak naprawdę operacją na zbiorze jego cech, które są udostępniane. Jeżeli na podstawie ujawnionych cech nie można wskazać dokładnie *e,* to element ten pozostaje anonimowy. Oprócz anonimizacji przy ochronie prywatności występują także takie pojęcia, jak: deidentyfikacja, depersonalizacja, pseudonimizacja, ukrywanie, uniejednoznacznianie. Nie są one jednak równoważne anonimizacji. Jako deidentyfikację rozumie się najczęściej proces usuwania danych identyfikacyjnych.

Deidentyfikacja nie musi prowadzić do anonimizacji. Na przykład pozbawienie mieszkańca budynku tabliczki z nazwiskiem na drzwiach do mieszkania, dokumentu tożsamości, numerów identyfikacyjnych itp. nie spowoduje tego, iż sąsiedzi przestaną go rozpoznawać w zbiorze innych mieszkańców budynku.

Pseudonimizacja jest procesem zastępowania danych identyfikacyjnych pseudonimami. Jest ona obecnie stosowana jako metoda anonimizacji dokumentów medycznych, dokumentów sądowych oraz innych dokumentów udostępnianych publicznie. Zasady pseudonimizacji, to jest reguły zastępowania,z reguły powtarzalne. Oznacza to, że różne dokumenty dotyczące tej samej osoby posiadają jednakowe pseudonimy, co umożliwia łatwą agregację i rozróżnialność. Zaletą pseudonimizacji jest jej prostota i możliwość stosowania w każdych warunkach. Pseudonimizacja jest jednak dosyć słabą metodą anonimizacji.
Na rycinie 2 przedstawiono przykład pseudonimizacji.



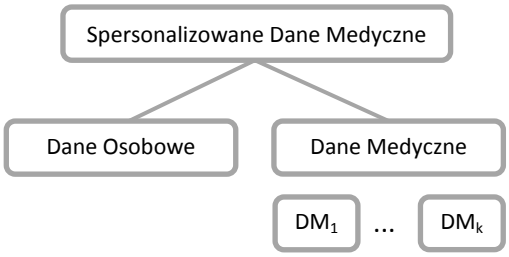
Ryc. 2. Ilustracja przykładowych efektów pseudonimizacji dokonanej zgodnie z zarządzeniem Pierwszego Prezesa Sądu Najwyższego RP [2]

Ukrywanie jest procesem ogólniejszym niż anonimizacja. Ukryciu mogą podlegać nie tylko cechy, ale również wzajemne zależności pomiędzy cechami mogące prowadzić do wskazania anonimizowanego elementu. Ukrywanie stosowane jest również w celach identyfikacyjnych, na przykład do identyfikacji praw autorskich, wykonawcy badania obrazowego, numerów identyfikacyjnych urządzenia rejestrującego itp.

Uniejednoznacznienie jest procesem, którego celem jest zaburzenie relacji pomiędzy cechami, cechami i elementami oraz relacji wzajemnych między elementami. Niektóre metody pseudonimizacji wprowadzają niejednoznaczność danych (na przykład zastępowanie nazw własnych ich pierwszymi literami).

4. Dane medyczne i ich przetwarzanie

Jako dane medyczne (DM) najczęściej przyjmuje się wszelkie informacje o stanie zdrowia pacjenta. Dane te powstają i są przetwarzane z reguły łącznie z danymi osobowymi pacjenta w postaci tak zwanych spersonalizowanych danych medycznych (SDM). W nowoczesnych systemach medycznych dane te mają postać elektroniczną i stanowią tak zwane elektroniczne spersonalizowane dane medyczne (eSDM).



Ryc. 3. Zależności pomiędzy danymi personalnymi a danymi medycznymi

4.1. Prywatność, poufność i bezpieczeństwo danych medycznych

Prywatność związana jest wprost z prawem jednostki do zachowania prywatności. Prawo to stanowi ochronę jednostki przed ingerencją osób trzecich lub instytucji w sferę życia osobistego i pozwala na ochronę przed niechcianymi lub nieprzyjaznymi ingerencjami. Prywatność jest pewnym szczególnym stanem bezpieczeństwa odnoszącym się do osoby. Bezpieczeństwo definiuje się formalnie jako stan braku zagrożeń. Konstrukcja zbioru zagrożeń determinuje rodzaj bezpieczeństwa. Podobnie jest z prywatnością. Przy analizie matematycznej naruszania stanu prywatności najczęściej zbiór zagrożeń redukuje się do ujawnienia tożsamości oraz ujawnienia i przyporządkowania wartości atrybutów wrażliwych.

Z bezpieczeństwem, a co za tym idzie również z prywatnością, związane jest pojęcie poufności. Poufność z kolei związana jest z przekazywaniem danych pomiędzy stronami. Jej istotą jest zapewnienie takich rozwiązań, aby dostęp do informacji posiadały tylko zdefiniowane precyzyjnie strony. Coraz częściej publikowane są bardzo szczegółowe zasady deidentyfikacji, takie jak na przykład opublikowane w USA HIPAA [3,4]. Problem anonimizacji został dostrzeżony przez Światową Organizację Zdrowia WHO i jest przedmiotem szerzej prowadzonych badań [5].

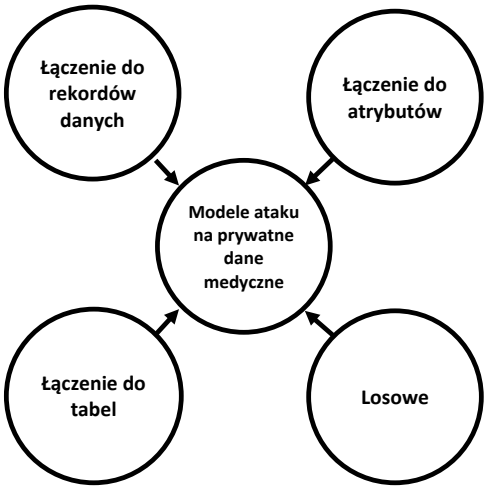
4.2. Publiczne udostępnianie danych medycznych a ochrona prywatności

Jedną z metod zapewnienia prywatności jest anonimizacja. W przypadkach danych medycznych anonimizacja prowadzona jest dopiero po ich zgromadzeniu i przetworzeniu wynikającym z potrzeb medycznych oraz uregulowań prawnych. Niezanonimizowane dane występują w postaci różnych kopii i są przetwarzane przez długi czas w jawnej postaci. Ochrona danych w takiej postaci powinna być prowadzona w sposób zapewniający ich integralność i poufność. Pojęcia integralności i poufności danych związane są z procesem ich przekazywania. Jeżeli sposób przekazywania danych zapewnia, iż dane, jakie otrzymuje odbiorca są identyczne z danymi wysyłanymi przez nadawcę. Istotne jest tu zachowanie zarówno wartości danych jak i ich struktury. Integralność danych można zapewnić na drodze kryptograficznej. Na przykład przez wyliczenie oraz przesłanie wraz z danymi wartości jednokierunkowej funkcji skrótu. Poufność zachowana jest wówczas, gdy z danymi wysłanymi przez nadawcę może zapoznać się wyłącznie odbiorca, dla którego te dane są przeznaczone. Podobnie jak integralność, również poufność można zapewnić na drodze kryptograficznej. Można do tego celu wykorzystać na przykład szyfrowanie danych przesyłanych pomiędzy nadawcą i odbiorcą. Sytuacja związana z ochroną prywatności zmienia się istotnie, gdy dane udostępniane są publicznie. Publiczne udostepnienie danych podyktowane jest interesem gospodarczym, naukowym oraz informacyjnym. Tu właśnie pojawia się pytanie, w jakiej formie i w jaki sposób udostępniać dane z zachowaniem prawa do prywatności osób, których dane zostały zgromadzone. Poszukuje się takich rozwiązań, które bez nadmiernych

kosztów i bez nadmiernych strat informacyjnych pozwalają pogodzić interesy związane z udostępnieniem i z zachowaniem prywatności. Problem ten nie dotyczy tylko danych medycznych, ale również wszelkich innych danych udostępnianych publicznie. Dane medyczne, ze względu na bardzo wrażliwe informacje w nich zawarte, traktowane są w sposób szczególny, wymagający wyjątkowo skutecznych zabezpieczeń. Powszechnie poszukuje się tu rozwiązań wykorzystujących anonimizację. Proste metody anonimizacji, takie jak pseudonimizacja działają lokalnie bez uwzględnienia wszystkich danych zgromadzonych w systemie oraz bez uwzględnienia dodatkowej wiedzy, jaką może posiadać podmiot zmierzający do uzyskania danych prywatnych.

5. Metody anonimizacji danych i ich zastosowanie do ochrony prywatności danych medycznych

Aby skutecznie zaprojektować mechanizmy ochrony prywatności konieczne jest precyzyjne zdefiniowanie zagrożeń. W literaturze wymienia się cztery podstawowe grupy metod ataku na prywatne dane medyczne: metody łączenia do rekordów danych, metody łączenia do atrybutów, metody łączenia do tabel oraz metody probabilistyczne (rycina 4). W pierwszych trzech grupach metod zakłada się, że atakujący zna atrybuty wchodzące w skład pseudoidentyfikatora PID. W metodach łączenia do rekordów i do atrybutów zakłada się, że znana jest wartość pseudoidentyfikatora dla ofiary i poszukuje się rekordu zawierającego dane wrażliwe ofiary. W przypadku łączenia do tabel atakujący dąży do ustalenia czy w opublikowanej tabeli znajduje się rekord ofiary. W przypadku metod probabilistycznych istotą jest zwiększenie wiedzy atakującego o danych rzeczywistych, na podstawie danych zanonimizowanych. W dalszej części przedstawiono metody anonimizacji oznaczone symbolicznie jako f(a)-anonimizacji oraz f(a,b)-anonimizacja. Jako f(a) oznaczono metody, które w najczęściej spotykanych źródłach oznaczane są jednym parametrem. Przykładem może tutaj być k-anonimizacja (f(a)=k). Symbolem f(a,b) oznaczono metody, które w źródłach literaturowych oznaczane są dwoma parametrami. W dalszej części przedstawiono metody: (X,Y)-anonimizacji, (α,k)-anonimizacji oraz (k,e)-anonimizacji.



Ryc. 4. Metody ataków na prywatne dane medyczne

5.1. Zapewnienie anonimowości metodą *k*-anonimizacji

Metoda *k*-anonimizacji [6-9] jest jedną z najczęściej opisywanych metod zapewnienia anonimowości. Związana jest ona z ochroną przed naruszeniem prywatności na drodze łączenia do rekordów w tabelach danych, np. danych pacjenta. Niech będzie dana tabela *T* zawierająca udostępniane dane pacjenta. W tabeli tej można wydzielić zbiór atrybutów stanowiących pseudoidentyfikator *PID*. Poszczególne wartości pseudoidentyfikatora wyznaczają z reguły mało liczne grupy rekordów w publikowanej tablicy *T*. Mała liczba rekordów w grupie daje również małą liczbę wartości, które można przypisać do pacjenta, którego dane chce się pozyskać. Mając dostęp do dodatkowej wiedzy o pacjencie, np. wiedzy pochodzącej z innych źródeł, można wyizolować poszukiwany rekord z grupy wskazywanej przez *PID*. Metoda *k*-anonimizacji jest najprostszym sposobem, aby zabezpieczyć publikowane dane przed dołączeniem rekordów zewnętrznych i jednoznacznym zidentyfikowaniem danych konkretnego pacjenta, i polega na zapewnieniu, iż liczność rekordów w grupach identyfikowanych przez wartości pseudoidentyfikatora *PID* jest nie mniejsza niż przyjęta z góry wartość *k*. W praktycznych zastosowaniach spełnienie tego warunku wymaga korekty danych dla atrybutów tworzących *PID*. Na rycinie 5 przedstawiona została przykładowa tabela *T*. Dla uproszczenia zawiera ona czteroelementowy zbiór atrybutów *A*={*Płeć*, *Zawód*, *Miasto*, *Choroba*}. Pseudifdentyfikator tworzą trzy z nich *PID*=*PID*(*Płeć*, *Zawód*, *Miasto*). Pseudoidentyfikator posiada tutaj trzy możliwe wartości, wyznaczając jednocześnie cztery grupy rekordów. Grupy te przedstawiono na rycinie 6. Niektóre wartości pseudoidentyfikatora mogą być mało liczne. Posiadając dodatkowe informacje, które mogą pochodzić ze źródeł ogólnodostępnych, możliwe staje się ujawnienie danych wrażliwych przez połączenie ze sobą rekordów o jednakowej wartości pseudoidentyfikatora. Przykład takiego ujawnienia przedstawiono na rycinie 7.

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

Ryc. 5. Przykładowa tabela o czterech atrybutach *T*(*Płeć*, *Zawód*, *Miasto*, *Choroba*)

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

Ryc. 6. Grupy rekordów wyznaczone przez pięć różnych wartości pseudidentyfikatora

	Płeć	Zawód	Miasto	Choroba
PID_1	M	Inżynier	Kraków	AIDS
	M	Inżynier	Kraków	AIDS
	M	Inżynier	Kraków	Grypa
	M	Inżynier	Kraków	AIDS
PID_2	M	Malarz	Opole	Nowotwór
	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

Ryc. 7. Ilustracja procesu możliwego ujawnienia informacji wrażliwych na podstawie dodatkowej wiedzy (na przykład danych pochodzących z innych źródeł ogólnodostępnych)

Ujawnieniu można zapobiec ustalając na przykład minimalną liczność jednakowych wartości przyjmowanych przez pseudoidentyfikator. Proces przekształcania publikowanej tabeli, tak aby zawierała minimum *k* jednakowych wartości pseudoidentyfikatora nazywa się *k*-anonimizacją. Na rycinie 8 przedstawiono przykład 4-anomimizacji. Anonimizacja została przeprowadzona metodą redukcji zbioru wartości dla pierwszej kolumny oraz metodą generalizacji atrybutów dla drugiej i trzeciej kolumny. Należy zwrócić uwagę na fakt, iż *k*-anonimizacja nie zabezpiecza w pełni przed ujawnieniem danych wrażliwych metodami statystycznymi.

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

↓

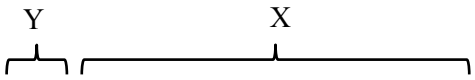
Płeć	Zawód	Województwo	Choroba
*	Techniczny	Małopolskie	AIDS
*	Techniczny	Małopolskie	AIDS
*	Techniczny	Małopolskie	Grypa
*	Techniczny	Małopolskie	AIDS
*	Artystyczny	Opolskie	Nowotwór
*	Artystyczny	Opolskie	Grypa
*	Artystyczny	Opolskie	Nowotwór
*	Artystyczny	Opolskie	AIDS
*	Artystyczny	Opolskie	AIDS

Ryc. 8. Ilustracja procesu 4-anonimizacji ze względu na zbiór atrybutów pseudoidentyfikatora *PID=PID* (*Płeć*, *Zawód*, *Miasto*)

Może dojść do ujawnienia statystycznego danych wrażliwych, na przykład dla inżyniera Osowskiego z Krakowa, wyszczególnionego w tabeli stanowiącej dodatkowe źródło wiedzy (rycina 7). W zanomizowanej tabeli opublikowanych danych medycznych, przedstawionej na rycinie 4, inżyniera Osowskiego można przyporządkować do grupy osób z województwa małopolskiego, legitymujących się zawodem technicznym. W grupie tej prawdopodobieństwo posiadania AIDS wynosi 0,75. Z takim też wysokim prawdopodobieństwem chorobę tę posiada rozpatrywana osoba.

5.2. Zapewnienie anonimowości metodą (*X,Y*)-anonimizacji

Metoda (*X,Y*)-anonimizacji jest uogólnieniem metody *k*-anonimizacji [10]. Istotą tej metody jest podział na dwa rozłączne zbiory atrybutów *X* oraz *Y*. Wymaga się tu, aby dla każdej wartości atrybutu *X* występowało przynajmniej *k* różnych wartości atrybutu *Y*. Należy zauważyć, iż *k*-anonimizacja jest szczególnym przypadkiem (*X,Y*)-anonimizacji. W najprostszy sposób jest to widoczne w tabelach, w których zbiór *Y* reprezentuje klucz główny. Na rycinie 9 przedstawiono przykład 4-(*X,Y*)-anonimizacji, w którym *Y*={*ID*}, *X*={*Płeć*, *Zawód*, *Województwo*}.



ID	Płeć	Zawód	Województwo
1	*	Techniczny	Małopolskie
2	*	Techniczny	Małopolskie
3	*	Techniczny	Małopolskie
4	*	Techniczny	Małopolskie
5	*	Artystyczny	Opolskie
6	*	Artystyczny	Opolskie
7	*	Techniczny	Opolskie
8	*	Artystyczny	Opolskie
9	*	Artystyczny	Opolskie

Ryc. 9. Przykład tabeli danych spełniających warunek 4-(*X*,*Y*)-anonimizacji

5.3. Zapewnienie anonimowości metodą (α,*k*)-anonimizacji

Metoda ta (α,*k*)-anonimizacji [11] oparta jest na *k*-anonimizacji oraz α-deasocjacji. Warunki *k*-anonimizacji i α-deasocjacji muszą być spełnione równocześnie. Pierwszy warunek nakłada ograniczenie na minimalną liczność rekordów w grupach odpowiadających poszczególnym wartościom pseudoidentyfikatorów. Liczność ta musi być większa lub równa *k*. Spełnienie warunku α-deasocjacji polega na tym, iż dla zadanej wartości wrażliwej *w* prawdopodobieństwo wystąpienia jest mniejsze lub równe α we wszystkich klasach równoważności. Na rycinie 10 przedstawiono przykład tabeli spełniającej warunki (0.5, 2)-anonimizacji.

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

Ryc. 10. Przykładowa tabela spełniająca warunek (0.5, 2)-anonimizacji dla wartości *w*=AIDS atrybutu Choroba

5.4. Zapewnienie anonimowości metodą (*k*,*e*)-anonimizacji

Dane medyczne oprócz danych tekstowych mogą zawierać dane liczbowe, obrazy oraz sekwencje wideo itp. Metoda (*k*,*e*)-anonimizacji [12] przeznaczona jest do ochrony danych wrażliwych mających postać liczbową. Taką postać ma wiele danych znajdujących się w dokumentacji medycznej. Do najpopularniejszych należą tu wyniki badań krwi i moczu oraz innych badań laboratoryjnych. Idea ta jest bardzo prosta. Wymaga się, aby rekordy były podzielone na grupy zawierające przynajmniej *k* różnych wartości wrażliwych przy maksymalnej różnicy wartości w grupie wynoszącej przynajmniej *e*. Na rycinie 11 przedstawiono przykład tabeli spełniającej warunki (8,170)-anonimizacji.

Płeć	Zawód	Glukoza
*	Techniczny	90
*	Techniczny	62
*	Techniczny	79
*	Techniczny	90
*	Techniczny	230
*	Techniczny	60
*	Techniczny	110
*	Techniczny	99

Ryc. 11. Tabela spełniająca warunki (*k*,*e*)-anonimizacji dla *k*=8 i *e*=170

6. Podsumowanie

Przedstawiony wyżej materiał stanowi pierwszą część z dwuczęściowej pracy dotyczącej problemów anonimizacji danych medycznych. W części tej przedstawiono definicje pojęć związanych z anonimizacją danych meydcznych oraz przeprowadzono analizę grup metod annimizacji oznaczonych symbolicznie jako: *f*(*a*)-anonimizacji i *f*(*a*,*b*)-anonimizacji. Do grupy *f*(*a*)-anonimizacji należy *k*-anonimizacja. Metoda ta, dzięki zwiększeniu liczności wartości pseudidentyfikatora, stanowi zabezpieczenie przed naruszeniem prywatności przez dołączanie rekordów. Do grupy metod *f*(*a*,*b*)-anonimizacji opisanych w pracy należą: (*X*,*Y*)-anonimizacja, (α,*k*)-anonimizacja oraz (*k*,*e*)-anonimizacja. Pierwsza z nich stanowi uogólnienie k-anonimizacji i pozwala na zwiększenie ochrony prywatności przez nadanie więzów na dwa rozłączne zbiory atrybutów. Druga z wymienionych metod oparta jest na *k*-anonimizacji oraz α-deasocjacji. Metoda ta, podobnie jak metoda (X,Y)-prywatności omówiona w drugiej części, może prowadzić do znacznego zniekształcenia danych źródłowych. Metoda (*k*,*e*)-anonimizacji przystosowana jest do ochrony danych wrażliwych o charakterze liczbowym. W przypadku danych medycznych mogą to być na przykład: wyniki badań laboratoryjnych czy dane o charakterze finansowym. W drugiej części pracy zostaną opisane kolejne, bardziej rozbudowane metody anonimizacji, oraz przedstawione zostanie rozwiązanie umożliwiające sterowanie anonimizacją przez posiadacza danych wrażliwych.

Piśmiennictwo

1. Rozporządzenie Ministra Zdrowia z dnia 21 grudnia 2010 r. w sprawie rodzajów i zakresu dokumentacji medycznej oraz sposobu jej przetwarzania. Dz.U. 2010 nr 252 poz. 1697.
2. Zarządzenie nr 11/2012 Pierwszego Prezesa Sądu Najwyższego z dnia 10.04.2012 r. w sprawie anonimizacji i udostępniania orzeczeń Sądu Najwyższego oraz informacji o sprawach sądowych w Sądzie Najwyższym [online][cyt. 1.02.2014]. Dostępny na URL: http://www.sn.pl/Aktualnosci/SiteAssets/Lists/Aktualnosci/NewForm/Zarz_PP_SN_11_2012.pdf

3. Health Insurance Portability and Accountability Act, P.L. 104-191, 110 Stat. 2023, 1996.
4. The Privacy Rule, 67 Federal Register 53182, 14 August 2002, codified at 45 CFR par. 160-164.
5. Management of patient information. Trends and challenges in Member States. Global Observatory for eHealth series – volume 6. WHO.
6. Samarati P. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering. *TKDE* 2001; 13(6): 1010-1027.
7. Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. InProc. of the 17th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems (PODS), page 188, Seattle, WA, June 1998.
8. Samarati P, Sweeney L. Protecting privacy when disclosing information:k-anonymity and its enforcement through generalization and suppression. Technical report,SRI International, March 1998.
9. Sweeney L. k-Anonymity: A model for protecting privacy. *Int J Uncertain Fuzz* 2002; 10(5): 557-570.
10. Wang K, Fung B. C. M. Anonymizing sequential releases. InProc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 414-423, Philadelphia, PA,August 2006.
11. 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 (KDD'06), 20-23 Aug 2006, Philadelphia, USA.
12. 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.

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

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