# Standard bezpieczeństwa

STANDARDY OTWARTOŚCI DANYCH

Fundusze Europejskie Polska Cyfrowa
Ministerstwo Cyfryzacji
OTWARTE DANE
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

# Spis treści

# Wstęp

Otwieranie danych publicznych zwiększa transparentność działań administracji oraz zakres kontroli nad działalnością państwa przez obywateli, a także umożliwia ich dalsze wykorzystywanie w produktach, aplikacjach czy usługach, w tym dla celów naukowych czy biznesowych. Rekomendowaną praktyką dla dysponentów danych jest organizowanie konsultacji z środowiskami (sektor badań i rozwoju, naukowcy, firmy komercyjne, organizacje pozarządowe) w celu określenia oczekiwanej zawartości informacyjnej zbiorów. Na podstawie takiego zapotrzebowania oraz źródłowego zakresu danych osobowych należy przeprowadzać procedurę otwierania danych publicznych.

Otwieranie danych publicznych do ponownego wykorzystywania[1] może jednocześnie rodzić ryzyko wkroczenia w sferę autonomii informacyjnej jednostek, dlatego istotne jest zapewnienie ochrony danych osobowych przez dysponentów informacji. Przez dane osobowe rozumie się informacje o zidentyfikowanej lub możliwej do zidentyfikowania osobie fizycznej („osobie, której dane dotyczą").

Możliwa do zidentyfikowania osoba fizyczna to osoba, którą można bezpośrednio lub pośrednio zidentyfikować, w szczególności na podstawie identyfikatora takiego jak: imię i nazwisko, numer identyfikacyjny, dane o lokalizacji, identyfikator internetowy lub jeden bądź kilka szczególnych czynników określających fizyczną, fizjologiczną, genetyczną, psychiczną, ekonomiczną, kulturową lub społeczną tożsamość osoby fizycznej[2].

Dobór odpowiednich technik – w kontekście udostępniania danych z zasobów publicznych – pozwalających na zapewnienie prywatności, ale umożliwiających zachowanie wartości informacyjnych danego zbioru i potencjału dla ponownego wykorzystywania, jest w tym wypadku kwestią kluczową.

---

[1] W rozumieniu art. 2 ust. 2 **ustawy z dnia 25 lutego 2016 r. o ponownym wykorzystywaniu informacji sektora publicznego**.

[2] W rozumieniu **Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych)**, dalej: RODO.

# 1.

# Minimalne czynności dotyczące udostępniania danych do ponownego wykorzystywania zgodnie ze standardem bezpieczeństwa

Opisane czynności dotyczą podstawowych zagadnień, jakie powinien wziąć pod uwagę dysponent danych publicznych (na gruncie przepisów o ochronie danych osobowych jest to administrator) udostępniający je do ponownego wykorzystywania.

Niezależnie od wymienionych w nim czynności, zadaniem dysponenta jest śledzenie osiągnięć w dziedzinie zabezpieczania systemów informatycznych i wdrażanie takich narzędzi oraz sposobów zarządzania danymi, które zapewnią bezpieczeństwo danych osobowych.

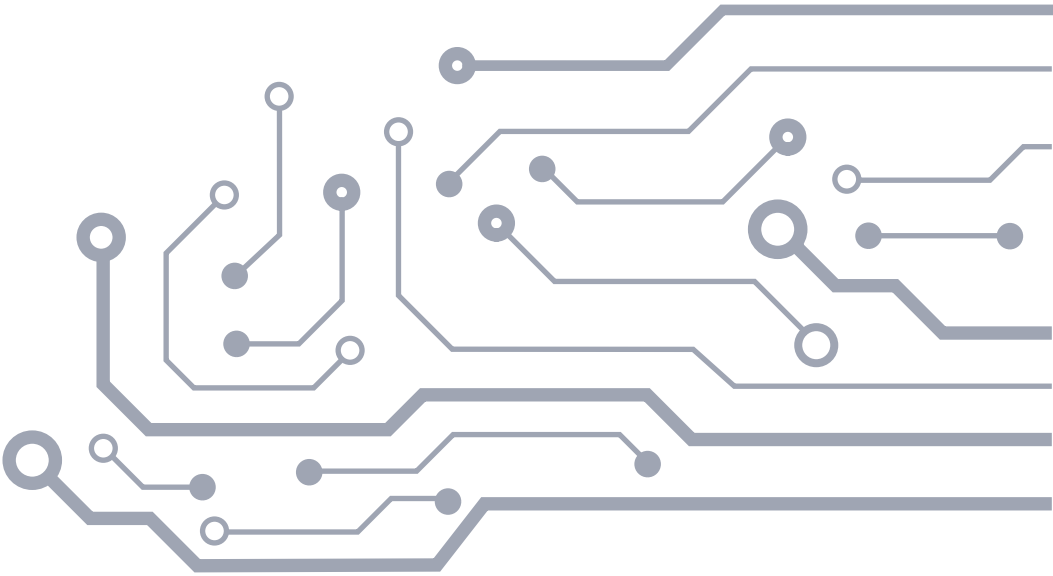
| Czynność | Możliwe wnioski |
|---|---|
| Analiza zawartości rejestru | • Nie zawiera danych osobowych<br>• Zawiera dane podlegające nieograniczonemu udostępnianiu – **zobacz cz. 2**<br>• Zawiera szczególne kategorie danych osobowych – **zobacz cz. 3**<br>• Zawiera dane osobowe – **zobacz cz. 3** |
| Analiza celu zbierania danych | • Przepisy prawa określają cele zbierania danych – porównaj z celem udostępniania<br>• Cel zbierania wynika z innych źródeł, niż przepisy prawa |
| Analiza związku z celem udostępniania | • Cel udostępniania jest tożsamy z celem zebrania<br>• Cel udostępniania jest inny niż cel przetwarzania danych, ale zachodzą okoliczności dopuszczające zmianę celu przetwarzania danych w stosunku do celu zebrania – **zobacz cz. 4** |
| Ocena kontekstu mającego wpływ na udostępnianie danych osobowych | • Dane nie wymagają anonimizacji lub pseudonimizacji, bo z ustawy wynika ich jawność<br>• Dane wymagają zastosowania anonimizacji lub pseudonimizacji – **zobacz cz. 5** i **Załącznik**<br>• Dane są udostępniane przez okres wskazany w przepisach prawa – **zobacz cz. 3** |

| Czynność | Możliwe wnioski |
|---|---|
| Ocena stosowanych środków techniczno-organizacyjnych służących zapewnieniu bezpieczeństwa przetwarzanych danych osobowych | Zastosowane środki:<br>• Pseudonimizacja i szyfrowanie danych osobowych<br>• Zdolność do ciągłego zapewnienia poufności, integralności, dostępności i odporności systemów i usług przetwarzania<br>• Zdolność do szybkiego przywrócenia dostępności danych osobowych i dostępu do nich w razie incydentu fizycznego lub technicznego<br>• Regularne testowanie, mierzenie i ocenianie skuteczności środków technicznych i organizacyjnych mających zapewnić bezpieczeństwo przetwarzania są odpowiednie do grup danych i wystarczające – **zobacz cz. 5** i **Załącznik** |
| Ocena ryzyka | Występuje:<br>• Możliwość szczątkowej identyfikacji<br>• Możliwość identyfikacji wskutek łączenia danych pochodzących z różnych źródeł<br>• Możliwość przetwarzania danych osobowych po ich usunięciu z zasobu publicznego<br>• Zagrożenie dla ochrony szczególnych kategorii danych<br>• Możliwość szerszego wykorzystywania danych osobowych, niż cele udostępniania<br>Opis ryzyk – **zobacz cz. 6**.<br>Postępowanie na wypadek ryzyka identyfikacji danych osobowych – **zobacz cz. 7**. |

# 2.

# Grupy danych podlegające nieograniczonemu udostępnianiu do ponownego wykorzystywania

1) Dane osób prawnych, w szczególności przedsiębiorstw będących osobami prawnymi, w tym danych o firmie i formie prawnej oraz danych kontaktowych osoby prawnej.
2) Dane wykorzystywane na podstawie rezygnacji osoby fizycznej lub przedsiębiorcy z przysługującego im prawa.
3) Dane osób publicznych w zakresie pełnienia przez nie funkcji, przy czym nie dotyczy to danych niemających związku z pełnioną funkcją (np. dane o miejscu zamieszkania, numer telefonu, dane o stanie rodzinnym itp.).
4) Publicznie dostępne dane statystyczne, pozyskiwane np. w oparciu o **ustawę o statystyce publicznej**.



# 3.

# Dane podlegające anonimizacji i pseudonimizacji oraz sposoby doboru technik

**1) Przykłady zastosowania technik dla grup danych**

**a) Dane osobowe**

Dane osobowe niedostępne publicznie zgodnie z obowiązującymi przepisami.

**TECHNIKA:** w celu udostępniania danych do ponownego wykorzystywania należy dokonać pełnej anonimizacji, dla zapewnienia odpowiedniego poziomu bezpieczeństwa wskazane jest jednoczesne zastosowanie technik randomizacji i uogólniania.

**b) Szczególne kategorie danych osobowych**

Dotyczy danych osobowych ujawniających pochodzenie rasowe lub etniczne, poglądy polityczne, przekonania religijne lub światopoglądowe, przynależność do związków zawodowych oraz dane genetyczne, dane biometryczne przetwarzane w celu jednoznacznego zidentyfikowania osoby fizycznej lub dane dotyczące zdrowia, seksualności lub orientacji seksualnej tej osoby[3].

Są to dane osobowe niedostępne powszechnie zgodnie z obowiązującymi przepisami. Ich udostępnianie w celu ponownego wykorzystywania jest zakazane.

**TECHNIKA:** w celu udostępniania danych do ponownego wykorzystywania należy dokonać pełnej anonimizacji, dla zapewnienia odpowiedniego poziomu bezpieczeństwa wskazane jest jednoczesne zastosowanie technik randomizacji i uogólniania.

**c) Dane powstałe w wyniku agregacji lub anonimizacji danych osobowych**

W przypadku udostępniania do ponownego wykorzystywania zbiorów danych powstałych w wyniku anonimizacji danych osobowych konieczne jest każdorazowo przeprowadzenia oceny ryzyka ich deanonimizacji.

**TECHNIKA:** w celu udostępniania danych do ponownego wykorzystywania oraz w razie ustalenia wystąpienia ryzyka identyfikacji danych osobowych należy powtórzyć proces anonimizacji.

[3] Chodzi o szczególne kategorie danych osobowych w rozumieniu art. 9 i 10 (czyli tzw. dane wrażliwe), tj. dane osobowe: ujawniające pochodzenie rasowe lub etniczne, poglądy polityczne, przekonania religijne lub światopoglądowe, przynależność do związków zawodowych oraz przetwarzania danych genetycznych, danych biometrycznych w celu jednoznacznego zidentyfikowania osoby fizycznej lub danych dotyczących zdrowia, seksualności lub orientacji seksualnej tej osoby; dotyczące wyroków skazujących oraz naruszeń prawa lub powiązanych środków bezpieczeństwa.

**d) Dane osobowe dostępne jawnie w zasobach publicznych**

Udostępnianie w celu ponownego wykorzystywania nawet jawnych danych osobowych rodzi ryzyko naruszenia praw osoby, której dane dotyczą.

**TECHNIKA:** w celu udostępnienia jawnych danych osobowych do ponownego wykorzystywania należy dokonać ich pseudonimizacji, wskazane jest także stosowanie środków technicznych uniemożliwiających masowe pobieranie danych.

Dodatkowo w przypadku zbiorów danych, które mają być udostępniane jedynie przez określony czas (w przypadku niektórych danych osobowych zawartych w zasobach publicznych okres ich przetwarzania został wprost wskazany w przepisach, po czym prawo do ich przetwarzania wygasa), konieczne jest zastosowania ich pełnej anonimizacji lub agregacji. Zgodnie z zasadą minimalizacji danych i ograniczenia przechowywania zakazane jest przechowywanie danych ponad okres, w jakim jest to niezbędne do celu, w jakim zostały zebrane.

**2) Przykłady zastosowania technik dla typów danych**

Każdy typ danych może podlegać obowiązkowi anonimizacji, jeśli może on zostać powiązany z konkretną osobą fizyczną.

Konieczność anonimizacji może wynikać z tego, że typ danych stanowi sam w sobie daną osobową (np. numer PESEL) lub z tego, że zbiór atrybutów jednoznacznie wskazuje na daną osobę lub istnieje szczątkowe ryzyko identyfikacji dla osoby, których dane dotyczą. Przykładem może być tu wyznanie, które nie powinno być ujawnianie w kontekście konkretnej osoby, ale jeśli zbiór danych jest odpowiednio uogólniony, możemy zostawić takie pole informacyjne bez obawy, że dla jakiejś osoby fizycznej zostaną ujawnione konkretne dane.

**a) Imię i nazwisko**

**TECHNIKA:** pełna anonimizacja lub wymazanie (np. przez maskowanie, aby nie było możliwości odczytu). Pseudonimizacja w przypadku, jeśli może wystąpić konieczność identyfikacji na indywidualny, uzasadniony prawnie wniosek.

**b) Adres (miasto, ulica, adres domu/mieszkania)**

**TECHNIKA:** pełna anonimizacja lub wymazanie (np. przez maskowanie, aby nie było możliwości odczytu). W przypadku braku ryzyka deanonimizacji można stosować uogólnienie, poprzez ujawnienie przykładowo województwa.

**c) Kod pocztowy**

**TECHNIKA:** pełna anonimizacja lub wymazanie (np. przez maskowanie, aby nie było możliwości odczytu). W przypadku braku ryzyka deanonimizacji można stosować uogólnienie poprzez maskowanie kilku ostatnich cyfr kodu.

**d) Dane liczbowe (zarobki, waga, wzrost lub inne dane opisujące osobę**

**TECHNIKA:** jednoczesne zastosowanie metod randomizacji (np. dodawanie zakłóceń +/- 20%) oraz uogólniania (np. agregacja przez wprowadzenie przedziałów wartości, przy założeniu, że poszczególne przedziały będą wystarczająco licznie reprezentowane).

**e) Identyfikatory będące samodzielnie danymi osobowymi (np. REGON, NIP, PESEL)**

**TECHNIKA:** pełna anonimizacja lub wymazanie (np. przez maskowanie, aby nie było możliwości odczytu). Pseudonimizacja w przypadku, jeśli może wystąpić konieczność identyfikacji na indywidualny, uzasadniony prawnie wniosek. Jeśli nie występuje ryzyko szczątkowe identyfikacji dla osoby, której dane dotyczą, możliwe jest udostępnienie części takiego pola, np. dwóch pierwszych cyfr numeru PESEL, które oznaczają rok urodzenia. Jeśli po takim ograniczeniu zawartości pola istnieje ryzyko identyfikacji, należy zastosować jednoczesne zastosowanie metod randomizacji i uogólniania, tak jak to opisano powyżej.

# 4.

# Zmiana celu przetwarzania danych osobowych z zasobów publicznych

Ponowne wykorzystywanie co do zasady będzie związane ze zmianą celu przetwarzania danych osobowych w stosunku do celu, dla jakiego były one zebrane.

Z uwagi na charakter danych osobowych znajdujących się w zasobach publicznych należy uznać, że osoby, które przekazują do nich dane, nie wyrażały do tej pory zgody na ich ponowne wykorzystywanie.

Dopuszcza się zmianę celu przetwarzania danych osobowych w stosunku do celu, w jakim zostały one pierwotnie zebrane przy spełnieniu jednej z poniższych przesłanek[4]:

1) Kiedy osoba, której dane dotyczą, wyraziła na to zgodę.
2) Kiedy nowy cel przetwarzania danych jest zgodny z pierwotnym celem ich zebrania, przy czym dalsze przetwarzanie dla celów archiwalnych w interesie publicznym, do celów badań naukowych lub historycznych lub do celów statystycznych nie jest uznawane za niezgodne z pierwotnymi celami.

W celu ustalenia, czy ponowne wykorzystywanie danych osobowych jest dopuszczalne, konieczne jest przeprowadzenie testu zgodności ponownego wykorzystywania z pierwotnym celem zebrania danych. Przykładowe okoliczności, które dysponent danych musi wziąć pod uwagę:

1) Wszelkie związki między celami, w których zebrano dane osobowe, a celami zamierzonego dalszego przetwarzania.
2) Kontekst, w którym zebrano dane osobowe, w szczególności relację między osobami, których dane dotyczą, a dysponentem danych.
3) Ewentualne konsekwencje zamierzonego dalszego przetwarzania dla osób, których dane dotyczą.
4) Istnienie odpowiednich zabezpieczeń, w tym ewentualnie szyfrowanie lub pseudonimizacja.

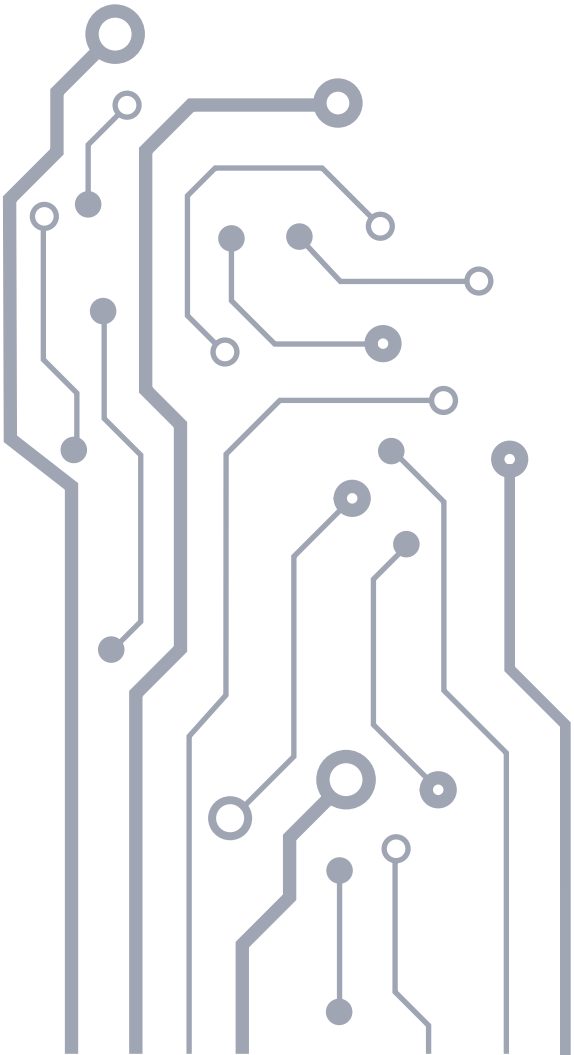
[4] Art. 23 ust. 1 RODO dopuszcza ponadto zmianę w następujących celach: a) ochrona bezpieczeństwa narodowego; b) zapewnienie obrony; c) zapewnienie bezpieczeństwa publicznego; d) zapobieganie przestępczości, prowadzeniu postępowań przygotowawczych, wykrywaniu lub ściganiu czynów zabronionych lub wykonywaniu kar, w tym ochronie przed zagrożeniami dla bezpieczeństwo publicznego i zapobieganiu takim zagrożeniom; e) inne ważne cele leżące w ogólnym interesie publicznym Unii lub państwa członkowskiego, w szczególności ważnemu interesowi gospodarczemu lub finansowemu Unii lub państwa członkowskiego, w tym kwestiom pieniężnym, budżetowym i podatkowym, zdrowiu publicznemu i zabezpieczeniu społecznemu; f) ochronie niezależności sądów i postępowania sądowego; g) zapobieganiu naruszeniom zasad etyki w zawodach regulowanych, prowadzeniu postępowań takich sprawach, ich wykrywaniu oraz ściganiu; h) cel kontrolny, inspekcyjny lub regulacyjny związany, nawet sporadycznie ze sprawowaniem władzy publicznej w przypadkach o których mowa w lit a-e) powyżej oraz g); i) ochronny wobec osoby, której dane dotyczą, lub praw i wolności tych osób oraz j) dla celów egzekucji roszczeń cywilnoprawnych. Analiza tych celów jasno wskazuje również, że przepis ten nie będzie mieć zastosowania do ponownego wykorzystywania danych publicznych.

5) Konsekwencje i zagrożenia wynikające z utraty kontroli nad udostępnionymi danymi po ich udostępnieniu.

6) Szczególną kategorię danych.

Jeśli w wyniku przeprowadzonego testu zgodności ponownego wykorzystywania z pierwotnym celem, dla którego dane zebrano, uzyskano odpowiedź negatywną dla udostępnienia danych osobowych zawartych w danych publicznych, konieczna będzie ich depersonalizacja z wykorzystaniem technik opisanych w **Załączniku**.

.



# 5.

# Środki techniczno--organizacyjne służące zapewnieniu bezpieczeństwa przetwarzanych danych osobowych

Dysponent danych i podmiot przetwarzający dane osobowe wdrażają odpowiednie środki techniczne i organizacyjne, aby zapewnić stopień bezpieczeństwa odpowiadający danemu ryzyku, w tym między innymi[5]:

1) Pseudonimizację i szyfrowanie danych osobowych.
2) Zdolność do ciągłego zapewnienia poufności, integralności, dostępności i odporności systemów i usług przetwarzania.
3) Zdolność do szybkiego przywrócenia dostępności danych osobowych i dostępu do nich w razie incydentu fizycznego lub technicznego.
4) Regularne testowanie, mierzenie i ocenianie skuteczności środków technicznych i organizacyjnych mających zapewnić bezpieczeństwo przetwarzania.

Jeżeli dany rodzaj przetwarzania – w szczególności z użyciem nowych technologii – ze względu na swój charakter, zakres, kontekst i cele z dużym prawdopodobieństwem może powodować wysokie ryzyko naruszenia praw lub wolności osób fizycznych, dysponent danych przed rozpoczęciem przetwarzania powinien dokonać oceny skutków planowanych operacji przetwarzania dla ochrony danych osobowych[6].

Ocena skutków dla ochrony danych zawiera co najmniej:

1) Systematyczny opis planowanych operacji przetwarzania i celów przetwarzania, w tym, gdy ma to zastosowanie – prawnie uzasadnionych interesów realizowanych przez dysponenta danych.
2) Ocenę, czy operacje przetwarzania są niezbędne oraz proporcjonalne w stosunku do celów.
3) Ocenę ryzyka naruszenia praw lub wolności osób, których dane dotyczą.
4) Środki planowane w celu zaradzenia ryzyku, w tym zabezpieczenia oraz środki i mechanizmy bezpieczeństwa mające zapewnić ochronę danych osobowych i wykazać przestrzeganie niniejszego rozporządzenia, z uwzględnieniem praw i prawnie uzasadnionych interesów osób, których dane dotyczą, i innych osób, których sprawa dotyczy.

Dysponenci danych powinni w tym zakresie posiłkować się dostępnymi wskazówkami (**Grupa Robocza Art. 29, Wytyczne dotyczące oceny skutków dla ochrony danych oraz ustalenia czy przetwarzanie „z dużym prawdopodobieństwem może powodować wysokie ryzyko”**), dobrymi praktykami i normami, jak np. norma ISO/IEC 29134:2017 (Information technology -- Security techniques -- Guidelines

[5] Zgodnie z artykułem 32 RODO.

[6] Art. 35 RODO, ang. Data Protection Impact Assessment (DPIA).

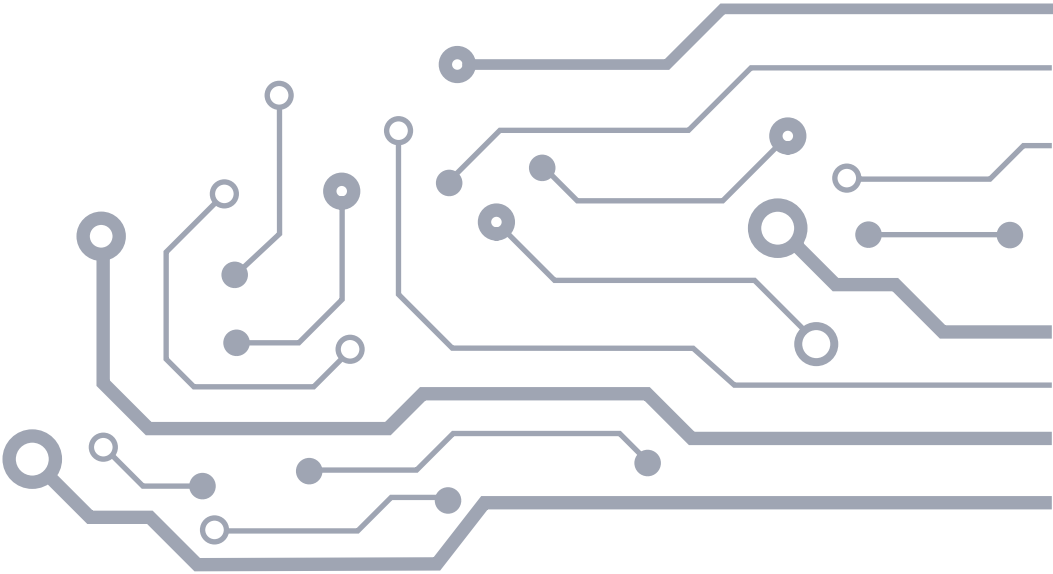
for privacy impact assessment). Warto także wykorzystywać dostępne narzędzia informatyczne (np. **http://arx.deidentifier.org/**) pozwalające na wykonanie anonimizacji za pomocą różnych technik oraz ocenę ryzyka naruszenia prywatności zanonimizowanego zbioru.

Przy analizie ryzyka na potrzeby DPIA można posiłkować się normą PN-ISO/IEC 27005 (Technika informatyczna -- Techniki bezpieczeństwa -- Zarządzanie ryzykiem w bezpieczeństwie informacji) oraz normą PN-ISO 31000 (Zarządzanie ryzykiem - Zasady i wytyczne). Analiza ryzyka powinna zostać udokumentowana oraz regularnie przeglądana.

Zgodnie z zasadami ochrony danych (odnoszących się do fazy projektowania oraz domyślnej ochrony danych), przy rozważaniu udostępnienia danych publicznych należy jak najwcześniej uwzględnić fakt, że niektóre dane publiczne mogą zawierać dane osobowe.

Przy udostępnieniu danych z zasobów publicznych dobór odpowiednich technik, pozwalających na zapewnienie prywatności, ale umożliwiających zachowanie wartości informacyjnej zbioru danych, jest kwestią kluczową. Technikami, które pozwalają na utrzymywanie korzyści z danych oraz minimalizują ryzyko w związku z utratą prywatności, jest anonimizacja lub pseudonimizacja.

Techniki anonimizacji i pseudonimizacji zostały opisane w **Załączniku**.



# 6.

# Ryzyka dla ochrony danych osobowych zawartych w danych publicznych

Z przetwarzania danych osobowych zawartych w zasobach publicznych, udostępnianych przez systemy teleinformatyczne do ponownego wykorzystywania, może wynikać ryzyko naruszenia praw lub wolności osób fizycznych o różnym prawdopodobieństwie i wadze zagrożenia.

Można wymienić następujące ryzyka związane z udostępnianiem danych:

1) **Możliwość szczątkowej identyfikacji z uwagi na zbyt powierzchowną anonimizację**
Użytkownicy postronni mogą dokonać identyfikacji danych konkretnej osoby, które w ocenie dysponenta danych zostały w pełni zanonimizowane bądź zagregowane. Problem ten dotyczy w szczególności danych statystycznych, które w wyniku zbyt dużej szczegółowości oraz w połączeniu ze zbyt małą próbą powodują, że osoby wchodzące w skład danej wspólnoty mogą przy pomocy powszechnie znanych im informacji o jej członkach dokonać odkodowania anonimowych informacji. Ryzyko identyfikacji danych może nastąpić także poprzez połączenie zanonimizowanego zbioru danych z innymi zbiorami, co w konsekwencji także pozwala na odkodowanie anonimowych informacji.

2) **Łączenie danych pochodzących z różnych źródeł**
W powiązaniu z danymi publicznie dostępnymi w innych zbiorach, dane mogą posłużyć do stworzenia kompleksowych zbiorów danych osobowych osób, których dane dotyczą, co może prowadzić do naruszenia ich prywatności.

3) **Przetwarzanie danych osobowych po ich usunięciu z zasobu publicznego (wykreśleniu z zasobu)**
Dane przetwarzane w zasobach publicznych często są w nich ujawniane na zadany okres, po czym podlegają wykreśleniu. W efekcie, po wykreśleniu danych osoba, której dane dotyczą, ma zagwarantowane zaprzestanie ich udostępniania po upływie okresu, przez jaki miały być one zgodnie z prawem przetwarzane. Po udostępnieniu danych do ponownego wykorzystywania dysponent danych traci wpływ na sposoby jego wykorzystywania, ponowne udostępnienie danych osobowych rodzi ryzyko ich przetwarzania w okresie, w jakim zostały one już wykreślone z zasobów publicznych.

4) **Zagrożenie dla ochrony szczególnych kategorii danych**
W przypadku udostępniania szczególnych kategorii danych zawartych w zbiorach publicznych istnieje ryzyko upublicznienia tych danych do nieograniczonego kręgu adresatów. Zasadniczo ich ponowne wykorzystywanie w celach innych, niż cele dla których zostały zebrane jest niedozwolone.

5) **Możliwość szerszego wykorzystywania danych**
Przetworzenie danych do formatu odczytu maszynowego powoduje możliwość szerszego wykorzystywania danych osobowych np. stosowanie profilowania, marketing bezpośredni, tworzenie baz adresowych oraz danych komunikacyjnych.

# 7.

# Postępowanie na wypadek ryzyka identyfikacji danych osobowych

Techniki anonimizacji wiążą się z konkretnymi ograniczeniami. Dysponenci danych muszą rozważyć te ograniczenia, zanim zastosują daną technikę w celu przeprowadzenia procesu anonimizacji. Muszą zwrócić uwagę na cele, jakie należy osiągnąć przez anonimizację – takie jak ochrona prywatności osób fizycznych przy publikowaniu zbioru danych lub dopuszczaniu wyszukania części informacji ze zbioru danych.

Nawet po zastosowaniu anonimizacji nie można wykluczyć innych zagrożeń, takich jak możliwość wyodrębnienia określonej osoby fizycznej, możliwość tworzenia powiązań między zapisami dotyczącymi określonej osoby, czy możliwość wnioskowania w odniesieniu do określonej osoby. Niektóre z tych zagrożeń można jednak wyeliminować lub ograniczyć za pomocą konkretnej techniki, przy czym zalecane jest łączenie różnych technik, co zwiększa szanse na osiągnięcie skutecznej ochrony danych osobowych.

W celu ograniczenia ryzyka identyfikacji danych osobowych należy dodatkowo uwzględnić następujące kwestie:

1) Należy jasno określić cele, jakie planuje się osiągnąć poprzez zanonimizowanie zbioru danych, ponieważ odgrywają one ważną rolę w określaniu ryzyka identyfikacji.

2) Ze względu na ryzyko szczątkowej identyfikacji dysponenci danych powinni:

   a) identyfikować nowe rodzaje ryzyka i regularnie przeprowadzać ponowne oceny ryzyka szczątkowego,

   b) ocenić, czy kontrole w odniesieniu do zidentyfikowanego ryzyka są wystarczające i odpowiednio dostosowywane,

   c) monitorować i kontrolować ryzyko.

3) W ramach ryzyka szczątkowego należy ocenić, czy istnieje możliwość identyfikacji w niezanonimizowanej części zbioru (jeżeli taka istnieje), w szczególności jeżeli została ona połączona z częścią zanonimizowaną, oraz możliwe korelacje między atrybutami, np. między danymi dotyczącymi lokalizacji geograficznej a danymi dotyczącymi poziomu zamożności.

4) Jednocześnie należy uwzględnić wszystkie odpowiednie elementy kontekstowe, np. charakter danych pierwotnych, istniejące mechanizmy kontroli (w tym środki bezpieczeństwa służące ograniczeniu dostępu do zbiorów danych), liczebność próby (cechy ilościowe), dostępność zasobów informacji publicznych (na których mogą opierać się odbiorcy), przewidziane udostępnienie danych osobom trzecim (ograniczone, nieograniczone np. w Internecie itp.).

5) Należy zwrócić uwagę na potencjalnych atakujących, uwzględniając atrakcyjność danych z perspektywy ukierunkowanych ataków (w tym kontekście ponownie najważniejszymi czynnikami będą szczególna ochrona informacji i charakter danych).

6) Ze zbioru danych należy usunąć oczywiste (np. rzadkie) atrybuty / quasi-identyfikatory.

7) Jeżeli stosuje się technikę dodawania zakłóceń (w randomizacji), poziom zakłóceń dodanych do zapisów należy określić jako funkcję wartości atrybutu (tj. nie należy dodawać żadnych zakłóceń wykraczających poza skalę), wpływu atrybutów, które mają podlegać ochronie, na osoby, których dane dotyczą, lub rozproszenie zbioru danych.

8) W przypadku opierania się na prywatności różnicowej (w randomizacji) należy uwzględnić konieczność monitorowania zapytań, aby wykrywać zapytania naruszające prywatność, ponieważ naruszenia w ramach zapytań mają charakter kumulacyjny.

9) Jeżeli wdrożono techniki uogólniania, bardzo istotne jest, aby dysponent danych nie ograniczał się do jednego kryterium uogólniania nawet w odniesieniu do tego samego atrybutu. Oznacza to, że należy wybierać różne poziomy szczegółowości lub różne przedziały czasowe. Wybór kryteriów, które należy stosować, musi zależeć od dystrybucji wartości atrybutu w danej populacji. Nie wszystkie dystrybucje nadają się do uogólniania, tj. w przypadku uogólniania nie można zastosować podejścia uniwersalnego. Należy zapewnić zmienność w ramach klas równoważności: na przykład wybrać określony próg na podstawie „elementów kontekstowych", o których mowa powyżej (liczebność próby itp.), i jeżeli próg ten nie zostanie osiągnięty, wówczas należy odrzucić określoną próbę (lub należy określić inne kryterium uogólniania).

10) Należy ujawnić technikę anonimizacji / połączenie technik, które zastosowano do udostępnionego zbioru danych.

# Zakończenie

Z przetwarzania danych osobowych zawartych w zasobach publicznych, udostępnianych poprzez systemy teleinformatyczne do ponownego wykorzystywania, może wynikać ryzyko naruszenia praw lub wolności osób fizycznych. Ryzyko to może mieć różne prawdopodobieństwo oraz wagę. Dysponent danych i podmiot przetwarzający dane osobowe powinien wdrożyć odpowiednie środki techniczne i organizacyjne, aby zapewnić stopień bezpieczeństwa odpowiadający temu ryzyku. Dobór odpowiednich technik pozwalających na ochronę prywatności, ale umożliwiających zachowanie wartości informacyjnych danego zbioru, w wypadku udostępniania danych do ponownego wykorzystywania jest bardzo ważną kwestią. Technikami, które pozwalają na uzyskiwanie korzyści z danych oraz minimalizują ryzyko w związku z utratą prywatności, jest anonimizacja lub pseudonimizacja.

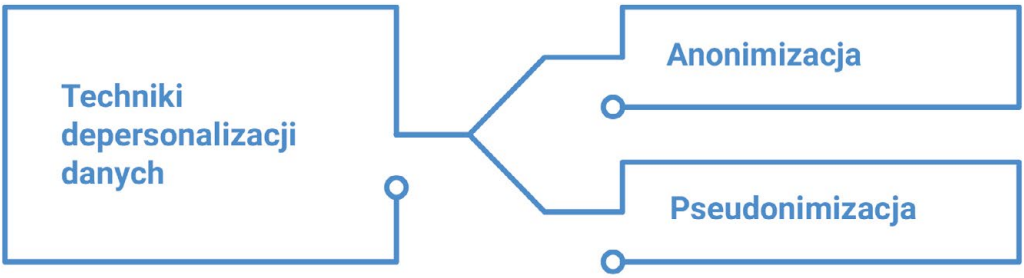
Należy podkreślić, że konieczność zapewnienia ochrony danych osobowych nie powinna stać na przeszkodzie otwieraniu danych publicznych. Zastosowanie odpowiedniej techniki depersonalizacji danych pozwala na pogodzenie obu wartości, tj. prywatności osób fizycznych i prawa do ponownego wykorzystywania danych publicznych.

# Załącznik

## Techniki depersonalizacji danych

Istnieją dwie główne techniki depersonalizacji danych osobowych. Jest to anonimizacja oraz pseudonimizacja. Z technicznego punktu widzenia największa różnica między nimi polega na tym, że pseudonimizacja jest procesem odwracalnym a anonimizacja już nie.

*Rysunek 1. Techniki depersonalizacji danych*




Techniki depersonalizacji danych posiadają zróżnicowany stopień odporności na czynniki ryzyka. Dla każdej z technik poniżej wskazane zostały podatności na trzy główne ryzyka zagrażające prywatności:

**Wyodrębnienie** jest to możliwość wyizolowania niektórych lub wszystkich rekordów, które identyfikują daną osobę w zbiorze danych.

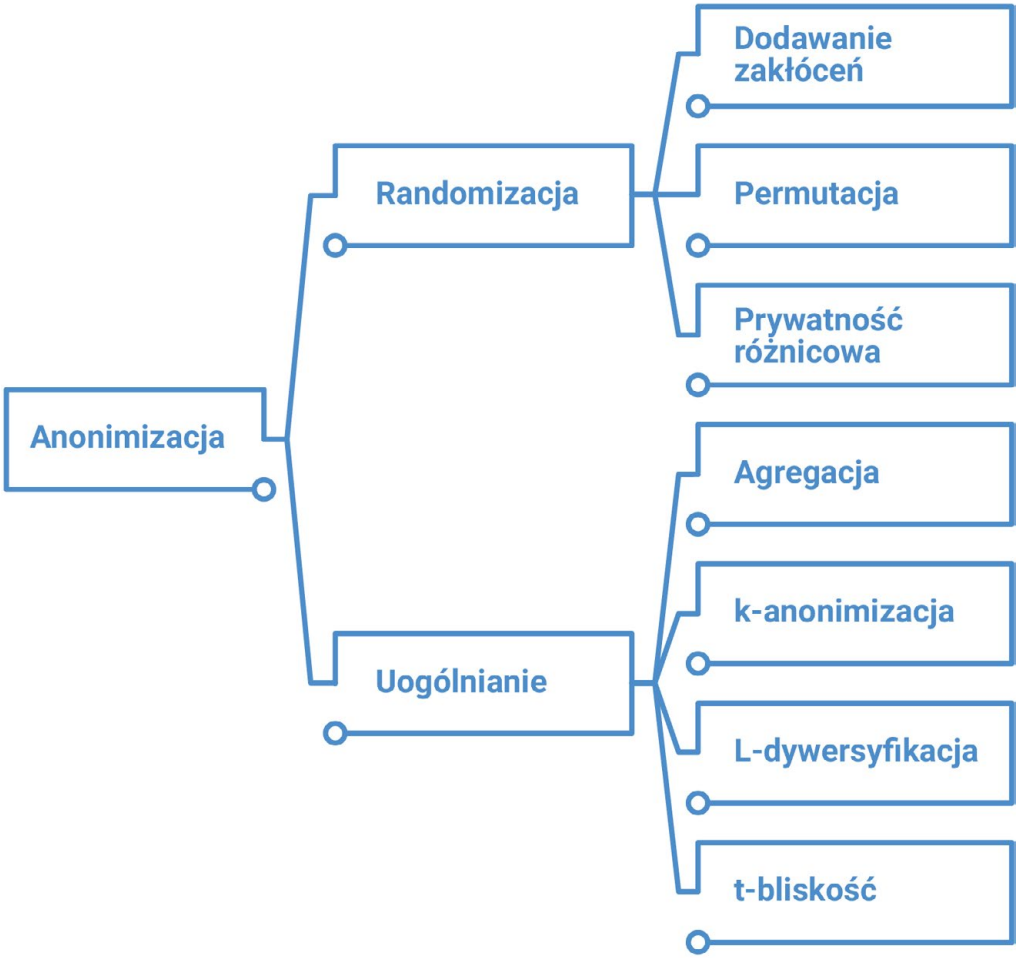
**Tworzenie powiązań** jest to możliwość powiązania co najmniej dwóch rekordów dotyczących tego samego podmiotu danych (w tej samej bazie danych lub w dwóch różnych bazach danych).

**Wnioskowanie** pozwala z dużym prawdopodobieństwem wydedukować wartość atrybutu z wartości zbioru innych atrybutów.

## Anonimizacja

Anonimizacja jest to proces, w którym dane osobowe są trwale i nieodwracalnie przekształcone. Technika ta uniemożliwia (w rozsądnym wymiarze czasowym i finansowym) przyporządkowanie informacji o określonej lub możliwej do zidentyfikowania osobie fizycznej.

*Rysunek 2. Podział technik anonimizacji*



1) **Randomizacja** jest to rodzina technik, które zmieniają prawdziwość danych w celu wyeliminowania silnego związku między danymi a konkretną osobą. Dane charakteryzują się wystarczającą niepewnością i nie można ich już odnieść do konkretnej osoby. Randomizacja jednak sama w sobie nie zmniejsza osobliwości każdego rekordu, ponieważ każdy rekord będzie pochodził od pojedynczego podmiotu danych, ale może natomiast chronić przed zagrożeniami wynikającymi z wnioskowania. W celu zapewnienia większej ochrony zaleca się połączenie randomizacji z innymi technikami np. generalizacji.
2) **Uogólnianie** stanowi drugą rodzinę technik anonimizacji. Polega ono na uogólnianiu lub osłabieniu atrybutów dla pomiotów danych, których dane dotyczą, poprzez modyfikację skali lub rzędu wielkości (np.: podanie miast zamiast dzielnic, tygodnia zamiast dnia). Mimo, że uogólnianie może być skuteczne, aby zapobiec wyodrębnieniu, nie pozwala ono na skuteczną anonimizację we wszystkich przypadkach. Wymaga specyficznych i wyrafinowanych podejść ilościowych, aby zapobiec tworzeniom powiązań i wnioskowaniu.

## Techniki randomizacji

1) **Dodawanie zakłóceń**

   Polega na modyfikowaniu atrybutów w zbiorze danych tak, aby były mniej dokładne przy zachowaniu ogólnej dystrybucji. Technika ta jest szczególnie użyteczna, gdy atrybuty mogą wywierać niekorzystny wpływ na poszczególne osoby. Zakłada się, że wartości danych będą prawdziwe (w znaczeniu wartości oryginalnych), ale tylko do pewnego zakresu. Przykładowo, jeżeli waga danej osoby została zmierzona z dokładnością do 0,5 kg, zbiór danych może zawierać wagę z dokładnością do +/-5 kg. Skuteczne zastosowanie tej techniki nie pozwoli osobie trzeciej na zidentyfikowanie konkretnej osoby. Osoba trzecia także nie będzie w stanie naprawić danych lub w inny sposób wykryć, w jaki sposób dane zostały zmodyfikowane.

   Technika ta zwyczajowo jest łączna z innymi technikami anonimizacji takimi jak usunięcie oczywistych atrybutów oraz quasi-identyfikatorów. Poziom zakłóceń powinien zależeć od poziomu wymaganych informacji i wpływu na prywatność osób w wyniku ujawnienia chronionych atrybutów.

*Tabela 1. Podatność na ryzyka metody dodawania zakłóceń*

| Wyodrębnienie | Tworzenie powiązań | Wnioskowanie |
|---|---|---|
| Możliwe jest wyodrębnianie zapisów danej osoby nawet, jeśli zapisy są mniej wiarygodne. | Można łączyć zapisy tej samej osoby, ale zapisy są mniej wiarygodne, a zatem prawdziwy zapis może być powiązany ze sztucznie dodanym (to znaczy z tzw. „szumem”).<br>W niektórych przypadkach błędne przypisanie może narazić osobę, której dane dotyczą na znaczący, a nawet wyższy poziom ryzyka niż właściwy. | Wnioskowanie jest możliwe, ale wskaźnik sukcesu będzie niższy, a niektóre fałszywe trafienia (i fałszywe negatywy) są wiarygodne. |

W technice tej często popełniane są dwa błędy. Pierwszym z nich jest dodawanie niespójnego zakłócenia. Jeżeli zakłócenie jest wykonane w sposób niesemantyczny (np. jest poza skalą lub jest nielogiczne) osoba trzecia będzie w stanie odfiltrować szum, a w niektórych przypadkach uzupełnić brakujące wpisy. Drugim błędem jest założenie, że samo dodanie zakłóceń jest wystarczające do zabezpieczenia danych. Technikę tę należy łączyć z innymi technikami anonimizacji, by w pełni zapewnić ochronę danych.

Tabela 2 przedstawia przykład wprowadzenia metody zakłóceń do zbioru danych. Kolumny: wiek, waga, wzrost zostały poddane zakłóceniom, przez co określenie konkretnej osoby jest znacznie utrudnione.

*Tabela 2. Zastosowanie metody dodawania zakłóceń*

Tabela pierwotna

| Wiek | Waga | Wzrost | Miasto |
|---|---|---|---|
| 25 | 62 | 180 | Lublin |
| 20 | 80 | 160 | Poznań |
| 35 | 75 | 176 | Opole |
| 45 | 49 | 156 | Kraków |
| 25 | 82 | 162 | Ełk |

Tabela zmodyfikowana

| Wiek (+/-5) | Waga (+/-5) | Wzrost (+/-5) | Miasto |
|---|---|---|---|
| 25 | 66 | 178 | Lublin |
| 22 | 79 | 165 | Poznań |
| 31 | 72 | 178 | Opole |
| 47 | 45 | 161 | Kraków |
| 24 | 79 | 162 | Ełk |

### 2) Permutacja

Permutacja polega na przetasowaniu wartości atrybutów w tabeli tak, aby wartości dla jednego podmiotu danych były sztucznie przypisane do innego. Dane w zbiorze pozostają niezmienione, zmieniona jest natomiast korelacja między wartościami a poszczególnym podmiotem. Technika ta jest szczególnie przydatna, gdy ważne jest zachowanie dokładnego rozkładu atrybutów w zestawie danych.

Permutacja jest to szczególna forma dodawania zakłóceń. W klasycznej technice dodawania zakłóceń atrybuty są modyfikowane przy pomocy losowych wartości. Generowanie spójnego szumu może być trudnym zadaniem, gdyż modyfikowanie wartości atrybutów w nieznaczny sposób może nie zapewniać odpowiedniej prywatności. Alternatywnie, techniki permutacji zmieniają wartości wewnątrz zbioru danych, zamieniając je pomiędzy rekordami. Takie zamiany zapewniają, że zakres i rozkład wartości pozostaną niezmienione, a korelacje między wartościami a podmiotami nie będą występować. Należy jednak pamiętać, że jeśli dwa lub więcej atrybutów ma związek logiczny lub posiadają korelację statystyczną i są one permutowane niezależnie, relacja ta zostanie zerwana. Dlatego ważne jest, aby permutować zestaw powiązanych atrybutów, aby nie zerwać zależności logicznej. W przeciwnym razie osoba atakująca mogłaby zidentyfikować permutowane atrybuty i odwrócić permutację.

*Tabela 3. Podatność na ryzyka metody permutacji*

| Wyodrębnienie | Tworzenie powiązań | Wnioskowanie |
|---|---|---|
| Możliwe jest wyodrębnienie zapisów danej osoby, nawet jeśli zapisy są mniej wiarygodne. | Może uniemożliwić "poprawne" powiązanie atrybutów zarówno wewnętrznie, jak i zewnętrznie z zestawem danych, ale nadal pozwala na "niepoprawną" powiązalność, ponieważ prawdziwy wpis może być powiązany z innym podmiotem danych. | Wnioskowanie jest możliwe, szczególnie jeśli atrybuty są skorelowane lub mają silne zależności logiczne; jednak nie wiedząc, które atrybuty zostały zmienione, atakujący musi wziąć pod uwagę, że jego wnioskowanie opiera się na błędnej hipotezie, a zatem możliwe jest jedynie wnioskowanie probabilistyczne. |

Najczęściej popełniane błędy w metodzie permutacji to:

- **Wybór niewłaściwego atrybutu:** polega na wyborze do permutacji atrybutów niewrażliwych lub nieobarczonych ryzykiem, przez co nie poprawia się ochrony danych osobowych;
- **Osobna permutacja atrybutów:** jeśli atrybuty są silnie skorelowane, należy je permutować razem, wykonanie osobnej permutacji takich parametrów pozwala na ich identyfikację i odwrócenie permutacji;
- **Założenie, że sama permutacja jest wystarczająca:** permutacja sama w sobie nie zapewnia anonimowości i powinna być łączona z innymi technikami, takimi jak usuwanie oczywistych atrybutów.

Przykład zastosowania metody permutacji przedstawia Tabela 4. Kolumny zawierające wiek i płeć zostały poddane permutacji (osobno). Warto zauważyć, że w przypadku przeprowadzenia permutacji na atrybutach dotyczących zawodu i zarobków należy wykonać permutację razem, by uniknąć sprzeczności logicznych (np. sprzedawca zarabiałby więcej od menadżera).

*Tabela 4. Zastosowanie metody dodawania zakłóceń*

Tabela pierwotna

| Zawód | Zarobki (tyś/ rok) | Rok ur. | Płeć |
|---|---|---|---|
| Sprzedawca | 24 | 1990 | K |
| Sprzedawca | 48 | 1985 | M |
| Kierowca | 60 | 1992 | K |
| Robotnik | 72 | 1986 | M |
| Menadżer | 120 | 1975 | M |

Tabela pierwotna

| Zawód | Zarobki (tyś/ rok) | Rok ur. | Płeć |
|---|---|---|---|
| Sprzedawca | 24 | 1980 | M |
| Sprzedawca | 48 | 1990 | M |
| Kierowca | 60 | 1986 | K |
| Robotnik | 72 | 1985 | M |
| Menadżer | 120 | 1980 | K |

### 3) Prywatność różnicowa

Prywatność różnicowa jest to technika randomizacji, która opiera się na innym podejściu niż wcześniej opisywane metody. Bazuje ona na wyborze odpowiedniego zanonimizowanego widoku danych przez dysponenta danych, przy czym zachowany jest oryginalny zbiór danych. Tak przygotowane zanonimizowane widoki danych przygotowywane są na podstawie zestawu zapytań generowanych przez osobę

trzecią. Zbiór taki posiada celowo dodane zakłócenia po procesie anonimizacji. Metoda prywatności różnicowej pozwala dysponentowi danych na dodawanie zakłóceń do zbioru w odpowiedniej ilości i formie, który zapewnienia adekwatny poziom ochrony prywatności. Niezwykle istotne jest, by stale monitorować możliwości identyfikacji podmiotu danych w zbiorach, które powstają na podstawie zapytań. Należy jednak podkreślić, że metoda prywatności różnicowej zachowuje dane oryginalne w postaci niezmienionej i dysponent danych może zidentyfikować poszczególne osoby fizyczne.

Niewątpliwą korzyścią wynikająca z zastosowania prywatności różnicowej jest dostarczenie osobom trzecim zbioru danych na podstawie konkretnego zapytania, a nie poprzez udostepnienie całego zbioru danych. Dysponent danych może gromadzić takie zapytania, kontrolując tym samym dostęp osób trzecich do danych, do których nie mają upoważnienia. Samo zapytanie można również poddać anonimizacji, zwiększając tym samym poziom ochrony prywatności.

Nie należy udostępniać baz danych wykorzystujących metodę prywatności różnicowej w systemach wyszukiwania, które nie zapewniają identyfikacji podmiotów wprowadzających zapytanie. Wykorzystanie bardzo wielu zapytań może umożliwić identyfikację konkretnego podmiotu danych poprzez wnioskowanie lub tworzenie powiązań.

*Tabela 5. Podatność na ryzyka metody prywatności różnicowej*

| Wyodrębnienie | Tworzenie powiązań | Wnioskowanie |
|---|---|---|
| W przypadku, gdy wynikiem zapytania są tylko statystyki, wyodrębnienie konkretnego podmiotu danych nie jest możliwe. | Wykorzystywanie dużej liczby zapytań może prowadzić do powiązania konkretnego podmiotu danych dwiema odpowiedziami. | Wykorzystywanie dużej liczby zapytań daje możliwość wnioskowania o konkretnym podmiocie danych. |

Najczęściej popełniany błąd w metodzie prywatności różnicowej to:

- **Wprowadzenie niewystarczającego zakłócenia:** w celu uniemożliwienia identyfikacji konkretnego podmiotu lub grupy podmiotów za pomocą dużej liczby zapytań, należy prawdziwe odpowiedzi uzupełnić o odpowiednią ilość zakłóceń.

**4) Agregacja i k-anonimizacja**

Techniki agregacji i k-anonimizacji polegają na grupowaniu ze sobą danych o podmiotach. W tym celu wartości atrybutów są uogólnianie do takiego stopnia, że dla kilku (umownie k) podmiotów przypisana jest ta sama wartość. Można to osiągnąć poprzez np. zmniejszenie szczegółowości lokalizacji z gminy do województwa. Dla danych liczbowych (takich jak zarobki, wzrost, parametry medyczne) generalizację można wykonać poprzez podanie wartości przedziałowych (np. wzrost 180–190 cm). Techniki te mogą być stosowane w przypadku, gdy korelacja wartości punktowych atrybutów może utworzyć quasi identyfikatory.

*Tabela 6. Podatność na ryzyka metody k-anonimizacji*

| Wyodrębnienie | Tworzenie powiązań | Wnioskowanie |
|---|---|---|
| Wyodrębnienie osoby z grupy k użytkowników nie powinno być już dłużej możliwe, ponieważ te same atrybuty są teraz udostępniane przez k użytkowników. | Pomimo ograniczenia tworzenia powiązań, możliwe jest łączenie rekordów według grup wpisów użytkowników. Prawdopodobieństwo, że dwa rekordy odpowiadają temu samemu pseudo-identyfikatorowi wynosi 1/k. | Wnioskowanie jest możliwe, zwłaszcza w przypadku gdy wiemy, że dana osoba należy do danej grupy. |

Najczęściej popełniane błędy w metodzie k-anonimizacji to:

- **Brak niektórych quasi-identyfikatorów:** najważniejszym parametrem w przypadku techniki k-anonimizacji jest parametr k czyli wielkość określająca ile podmiotów danych powinno być w grupie. Im większa wartość parametru k tym silniejsze są gwarancje zachowania prywatności. Często spotyka się sytuację, że w celu zwiększenia wartości parametru k ogranicza się ilość quasi-identyfikatorów, co pozwala na budowanie większych klastrów k-użytkowników. Taki zabieg może prowadzić do tego, że istnieje możliwość identyfikacji jednostki w zbiorze poprzez wykorzystywanie innych dostępnych informacji dotyczących danego podmiotu.
- **Zbyt mała wartość k:** w przypadku zmniejszenia wartości parametru k waga każdego podmiotu danych w danej grupie jest znacząca. Istnieje wtedy podwyższone ryzyko identyfikacji podmiotu danych na podstawie wnioskowania.

- **Pomijanie grupowania podmiotów:** w przypadku grupowania zbioru danych o nierównej dystrybucji wpływ rekordu na daną grupę będzie się różnił w zależności od jego reprezentacji w danej grupie.
- **Grupowanie zestawu osób o nierównomiernym rozmieszczeniu atrybutów:** wpływ rekordu danej osoby na zbiór danych będzie się różnić: niektóre będą stanowić znaczną część wpisów, podczas gdy wkład innych osób pozostanie dość nieznaczny. Dlatego ważne jest, aby upewnić się, że k jest wystarczająco wysokie, aby żadne osoby nie stanowiły zbyt ważnej części wpisów w klastrze.

W przykładzie poniżej przedstawiono zastosowanie metody k-anonimizacji dla wartości k=2 (Tabela 7). Zbiór został przekształcony w taki sposób, że dla minimum dwóch rekordów dane posiadają te same informacje. Uniemożliwia dokładną identyfikację konkretnej osoby.

*Tabela 7. Zastosowanie metody k-anonimizacji (k=2)*

Tabela pierwotna

| Nazwisko | Wiek | Płeć | Powiat | Choroba |
|---|---|---|---|---|
| Nowak | 29 | K | Miński | Rak |
| Kowal | 24 | K | Otwocki | Psychiczna |
| Gajda | 28 | K | Miński | Astma |
| Halicki | 27 | M | Płocki | Brak |
| Górski | 24 | K | Otwocki | Serca |
| Golec | 23 | M | Płocki | Astma |
| Dłuski | 19 | M | Otwocki | Rak |
| Bogucki | 29 | M | Płocki | Serca |
| Leja | 17 | M | Otwocki | Serca |
| Nowak | 19 | M | Otwocki | Psychiczna |

Tabela zmodyfikowana

| Nazwisko | Wiek | Płeć | Powiat | Choroba |
|---|---|---|---|---|
| * | 20 < Wiek | K | Miński | Rak |
| * | 20 < Wiek | K | Otwocki | Psychiczna |
| * | 20 < Wiek | K | Miński | Astma |
| * | 20 < Wiek | M | Płocki | Brak |
| * | 20 < Wiek | K | Otwocki | Serca |
| * | 20 < Wiek | M | Płocki | Astma |
| * | Wiek ≤ 20 | M | Otwocki | Rak |
| * | 20 < Wiek | M | Płocki | Serca |
| * | Wiek ≤ 20 | M | Otwocki | Serca |
| * | Wiek ≤ 20 | M | Otwocki | Psychiczna |

Stosowane są także bardziej rozbudowane metody k-anonimizacji takie jak (X,Y) – anonimizacji, (α,k)-anonimizacji, (k,e)-anonimizacji[7].

### 5) l-dywersyfikacja i t-bliskość

Metoda L-dywersyfikacji to rozszerzenie metody k-anonimizacji. Przyjmuje się w niej, że dla każdej grupy wartości powinno występować co najmniej L różnych wartości. Rozumie się przez to odpowiednie dobranie wartości atrybutu tak, aby zapewnić odpowiednią liczność, a z drugiej strony poprawność dziedzinową.

Dzięki takiemu podejściu technika ta jest odporna na ataki wnioskowania. W celu poprawnego zastosowania, w metodzie tej należy ograniczyć klasy o niskiej zmienności atrybutów. Dzięki temu osoba atakująca ma zawsze znaczną niepewność odnośnie do konkretnego podmiotu danych, równą odwrotności dywersyfikacji wynoszącą 1/L.

[7] Liber, A. (2014). Problemy anonimizacji dokumentów medycznych. Część 1. Wprowadzenie do anonimizacji danych medycznych. Zapewnienie ochrony danych wrażliwych metodami f (a)-if (a, b)-anonimizacji. Higher School's Pulse, 8(1). Liber, A. (2014). Problemy anonimizacji dokumentów medycznych. Część 2 Anonimizacja zaawansowana oraz sterowana przez posiadacza danych wrażliwych.

Metoda L-dywersyfikacji skutecznie chroni przed atakami polegającymi na wnioskowaniu dla zbiorów o stosunkowo dobrze rozłożonych wartościach atrybutów. W przypadku zbioru danych o niewielkim zakresie lub o nierównym rozmieszczeniu metoda ta nie jest już odporna na ataki wnioskowania.

Metoda t-bliskości jest udoskonaleniem metody L-dywersyfikacji. Ma ona za zadanie stworzenie równoważnych klas, które przypominają początkowy rozkład atrybutów w tabeli. Nakłada ona ograniczenia na rozkłady prawdopodobieństwa występowania wartości atrybutów w grupach oraz w całej tabeli. Dąży się do tego, aby oba rozkłady były jak najbliższe oryginałowi.

*Tabela 8. Podatność na ryzyka metody L-dywersyfikacji i t-bliskości*

| Wyodrębnienie | Tworzenie powiązań | Wnioskowanie |
|---|---|---|
| Wyodrębnienie konkretnego podmiotu danych nie powinno być już możliwe. | Pomimo ograniczenia tworzenia powiązań, możliwe jest łączenie rekordów według grup wpisów użytkowników. | Nie jest możliwe ze 100% pewnością określenie danego podmiotu danych na podstawie wnioskowania. |

Najczęściej popełniany błąd w metodzie L-dywersyfikacji i t-bliskości to:

- **Zabezpieczenie szczególnie chronionych wartości atrybutów poprzez zmieszanie ich z innymi atrybutami szczególnie chronionymi:** dwie wartości atrybutu w klastrze nie wystarczają do zapewnienia prywatności. W praktyce dystrybucja wartości szczególnie chronionych w każdym klastrze powinna przypominać dystrybucję tych wartości w całej populacji lub przynajmniej powinna być ona jednakowa w całym klastrze.

Tabela 9 przedstawia przykład metody L-dywersyfikacji trzeciego poziomu (L = 3). Oznacza to, że dla każdej grupy wartości pseudo-identyfikatora występują przynajmniej trzy „dobrze reprezentowane" rekordy danych źródłowych (tu konkretnie informacje o dochodzie i schorzeniu). Lewa strona tabeli jest oryginalną tabelą, prawa strona pokazuje wersję zanonimizowaną spełniającą 3-dywersyfikację. Za wrażliwe atrybuty przyjęto tutaj informacje dotyczące wynagrodzenia i choroby.

*Tabela 9. Zastosowanie metody L-dywersyfikacji (L=3)*

Tabela pierwotna

| | Kod pocztowy | Wiek | Dochód | Schorzenie |
|---|---|---|---|---|
| 1 | 47677 | 29 | 3 000 | wrzody |
| 2 | 47602 | 22 | 4 000 | nieżyt żołądka |
| 3 | 47678 | 27 | 5 000 | rak żołądka |
| 4 | 47905 | 43 | 6 000 | nieżyt żołądka |
| 5 | 47909 | 49 | 11 000 | grypa |
| 6 | 47906 | 47 | 8 000 | zapalenie oskrzeli |
| 7 | 47605 | 30 | 7 000 | zapalenie oskrzeli |
| 8 | 47673 | 36 | 9 000 | zapalenie płuc |
| 9 | 47607 | 32 | 10 000 | rak żołądka |
| 10 | 47909 | 52 | 9 000 | nieżyt żołądka |
| 11 | 47905 | 59 | 12 000 | grypa |
| 12 | 47908 | 61 | 10 000 | rak żołądka |

Tabela zmodyfikowana

| Kod pocztowy | Wiek | Dochód | Schorzenie |
|---|---|---|---|
| 476** | 2* | 3 000 | wrzody |
| 476** | 2* | 4 000 | nieżyt żołądka |
| 476** | 2* | 5 000 | rak żołądka |
| 4790* | 40-50 | 6 000 | nieżyt żołądka |
| 4790* | 40-50 | 11 000 | grypa |
| 4790* | 40-50 | 8 000 | zapalenie oskrzeli |
| 476** | 3* | 7 000 | zapalenie oskrzeli |
| 476** | 3* | 9 000 | zapalenie płuc |

| Kod pocztowy | Wiek | Dochód | Schorzenie |
|---|---|---|---|
| 476** | 3* | 10 000 | rak żołądka |
| 4790* | ≥50 | 9 000 | nieżyt żołądka |
| 4790* | ≥50 | 12 000 | grypa |
| 4790* | ≥50 | 10 000 | rak żołądka |

Źródło: Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on (pp. 106-115). IEEE. https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf

Tabela 10 przedstawia przykład metody t-bliskości. Lewa strona tabeli jest oryginalną tabelą, prawa strona pokazuje wersję zanonimizowaną spełniającą odpowiednie założenia dotyczące rozkładu wynoszące dla dochodu t=0,167, a dla schorzenia t=0,278.

*Tabela 10. Zastosowanie metody L-dywersyfikacji (L=3)*

Tabela pierwotna

| | Kod pocztowy | Wiek | Dochód | Schorzenie |
|---|---|---|---|---|
| 1 | 47677 | 29 | 3 000 | Wrzody |
| 2 | 47602 | 22 | 4 000 | nieżyt żołądka |
| 3 | 47678 | 27 | 5 000 | rak żołądka |
| 4 | 47905 | 43 | 6 000 | nieżyt żołądka |
| 5 | 47909 | 52 | 11 000 | Grypa |
| 6 | 47906 | 47 | 8 000 | zapalenie oskrzeli |
| 7 | 47605 | 30 | 7 000 | zapalenie oskrzeli |
| 8 | 47673 | 36 | 9 000 | zapalenie płuc |
| 9 | 47607 | 32 | 10 000 | rak żołądka |

Tabela zmodyfikowana

| | Kod pocztowy | Wiek | Dochód | Schorzenie |
|---|---|---|---|---|
| 1 | 476** | ≤40 | 3000 | wrzody |
| 3 | 476** | ≤40 | 5 000 | rak żołądka |
| 8 | 476** | ≤40 | 9 000 | zapalenie płuc |
| 4 | 4790* | ≥40 | 6 000 | nieżyt żołądka |
| 5 | 4790* | ≥40 | 11 000 | grypa |
| 6 | 4790* | ≥40 | 8 000 | zapalenie oskrzeli |
| 2 | 476** | ≤40 | 4 000 | nieżyt żołądka |
| 7 | 476** | ≤40 | 7 000 | zapalenie oskrzeli |
| 9 | 476** | ≤40 | 10 000 | rak żołądka |

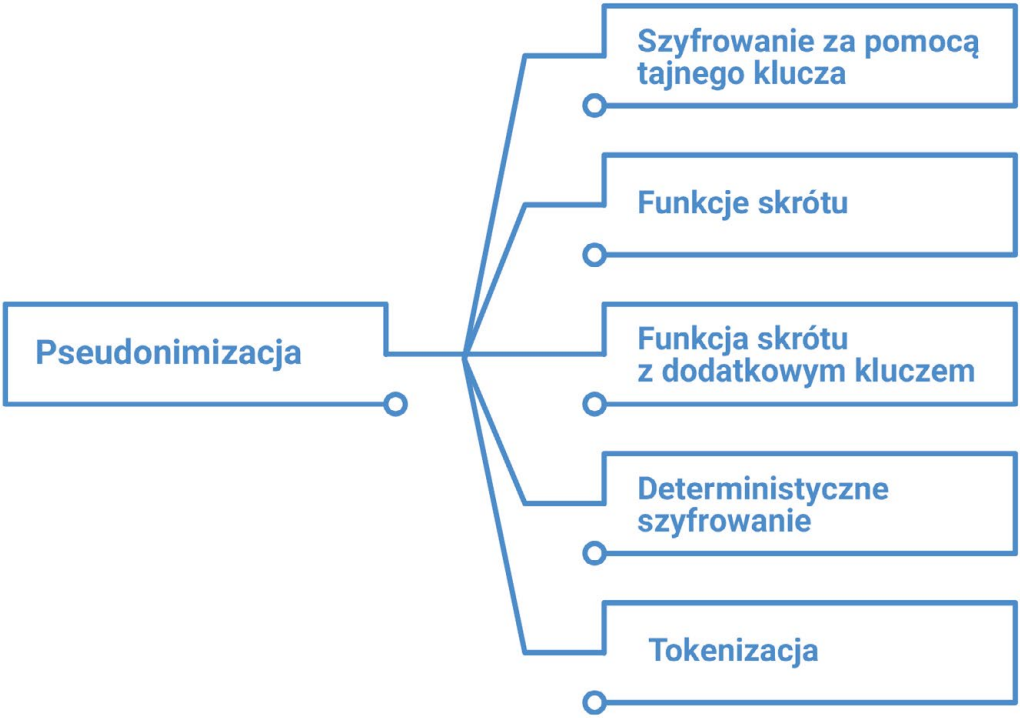
Źródło: Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on (pp. 106-115). IEEE. https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf

## Pseudonimizacja

Pseudonimizacja jest to odwracalny proces, polegający na zastąpieniu danej rzeczywistej nazwą przybraną, czyli nadanie jej tak zwanego pseudonimu. Pseudonimizacja utrudnia identyfikację, natomiast umożliwia przypisanie różnych czynności tej samej osobie (bez znajomości jej danych osobowych) oraz łączenie rożnych zbiorów danych między sobą. Pseudonimizacja skutecznie podwyższa bezpieczeństwo przetwarzania danych, ale nie jest równoznaczna anonimizacji, w związku z czym dane poddane pseudonimizacji dalej podlegają pełnej ochronie.

Poniższy rysunek prezentuje podział pseudonimizacji na pięć głównych kategorii.

*Rysunek 3. Podział technik pseudonimizacji*




### 1) Szyfrowanie za pomocą tajnego klucza

Dane osobowe są nadal przechowywane w zbiorze danych, ale w formie zaszyfrowanej. Posiadanie klucza szyfrującego pozwala na pełen dostęp do danych osobowych. Używając szyfrowania, które zachowuje aktualne standardy bezpieczeństwa, możliwość odszyfrowania danych jest możliwa, ale tylko z użyciem klucza szyfrującego.

### 2) Funkcje skrótu

Polega na skróceniu każdej wielkości w danej liście atrybutów do stałej określonej wartości. Funkcji tej nie można odwrócić, tak jak w przypadku szyfrowania. Jakkolwiek, znając zakres wartości, jakie zostały poddane skracaniu oraz w jaki sposób zostało to wykonane, możliwe jest odtworzenie funkcji skrótu i uzyskanie prawidłowego zapisu poprzez tzw. atak siłowy (wypróbowanie wszystkich możliwych kombinacji w celu utworzenia tabel korelacji).

Funkcje skrótu można podzielić ze względu na wielkości bloku wyjściowego (ilość bitów). Obecne zalecenia amerykańskiej agencji NIST[8] dotyczące stosowania poszczególnych funkcji skrótu mówią, że do nowych aplikacji zalecane są funkcje skrótu z rodziny SHA-2[9], a w przyszłości funkcja SHA-3[10]. Do niedawna stosowane były funkcje skrótu MD5[11] oraz SHA-1[12], zostały on jednak wycofane ze względu na niewystarczający poziom zabezpieczeń. Tabela 4-11 prezentuje przykłady bloków wyjściowych dla różnych funkcji skrótu. Istnieją różnego rodzaju darmowe oprogramowania generujące funkcje skrótu: np. Gperf[13], CCP-Crypto[14].

*Tabela 11. Funkcje skrótu dla wybranej frazy*

| Funkcja skrótu | Blok wyjściowy (wejściowa fraza: Jan Kowalski) |
|---|---|
| SHA-2* | e70207d004bdffb1702b9c50f5a8c70a49f6f8a-68404ca59993d3c39142162ba1bac5cc2c396460c205b0a-79676c736ae345257aae26c61444f3efc1efa822c0 |
| SHA-3* | 88f2a580725f980b96829210bf3e7a159a5ce23a13f289d-077c3e45ef0c5a3876a81618f70c6641f14cd4fc58a5cf80d-928909187c772108e9342051e632fdbb |

*Wersja 512 bitowa
Źródło: Opracowanie własne na podstawie **https://emn178.github.io/online-tools/**

[8] National Institute of Standards and Technology.

[9] SHA-2 składa się z zestawu czterech funkcji dających skróty wielkości 224, 256, 384 lub 512 bitów.

[10] SHA-3 funkcja oparta o Algorytm Keccak charakteryzuje się wyższą wydajnością niż SHA-2 zarówno w implementacjach sprzętowych jak i programowych.

[11] MD5 – generuje z ciągu danych o dowolnej długości128-bitowy skrót.

[12] SHA-1 tworzą 160-bitowy skrót z wiadomości o maksymalnym rozmiarze 264 bitów i jest oparty na podobnych zasadach co MD5.

[13] **https://www.gnu.org/software/gperf/**

[14] **https://github.com/torvalds/linux/blob/master/drivers/crypto/ccp/ccp-crypto-main.c**

### 3) Funkcja skrótu z dodatkowym kluczem

Jest to rozbudowana funkcja skrótu, która dodatkowo wykorzystuje do ochrony prywatności tajny klucz. Możliwość odszyfrowania danych nie jest praktycznie możliwa przez osoby trzecie, ze względy na bardzo dużą liczbę powstałych kombinacji do deszyfracji. Dysponent danych może natomiast bez problemu powrócić do pierwotnej formy danych wykorzystując tajny klucz. Najpowszechniejszej wykorzystywanymi funkcjami są HMAC i UMAC[15].

*Tabela 12. Funkcje skrótu z dodatkowym kluczem dla wybranej frazy*

| Funkcja skrótu z kluczem | Blok wyjściowy (wejściowa fraza: Jan Kowalski z kluczem Cyfryzacja) |
|---|---|
| HMAC (SHA1) | cf7449fcdeea860bbfc503410520d6623dd40a1e |
| HMAC (SHA512) | 2fd3e276606e45b474e07fcca1b605e3127f-03ce9fb11aec09a5f3cda4f315b41a-6ecdfd94016edf49ab35fb1478f13247c7b-23c7e5e509e68238ad9088bec70 |

Źródło: Opracowanie własne na podstawie **https://www.freeformatter.com/hmac-generator.html**

### 4) Deterministyczne szyfrowanie

Technika ta polega na wygenerowaniu pseudonimu w postaci losowej liczby dla każdego atrybutu, a następnie usunięciu tabeli powiązań. Dzięki temu zabiegowi znacznie ograniczone jest ryzyko identyfikacji poszczególnego podmiotu danych na podstawie tworzenia powiązań z innymi tabelami, gdzie jest on wymieniony z innym pseudonimem.

### 5) Tokenizacja

Metoda ta polega na wykorzystaniu jednokierunkowych mechanizmów szyfrujących opartych na przypisaniu identyfikatora (indeksu, sekwencji lub losowo wygenerowanej liczby) w żaden sposób niezwiązanego z pierwotnymi danymi. Technika ta jest często spotykana w sektorze finansowym do autoryzacji operacji bankowych.

[15] HMAC / UMAC- kod MAC z wmieszanym kluczem tajnym zapewniający zarówno ochronę integralności jak i autentyczności danych

*Tabela 13. Podatność na ryzyka metod pseudonimizacji*

| Wyodrębnienie | Tworzenie powiązań | Wnioskowanie |
|---|---|---|
| Możliwe jest wyodrębnienie konkretnego podmiotu danych, ponieważ jest on identyfikowany przez unikalny pseudonimowany atrybut. | Tworzenie powiązań jest możliwe i proste na podstawie innych niepseudonimizowanych atrybutów. Ryzyko jest wykluczone tylko jeśli żaden inny atrybut w zestawie danych nie może zostać wykorzystany do zidentyfikowania podmiotu danych i jeśli wyeliminowano wszystkie powiązania między pierwotnym atrybutem a atrybutem pseudonimicznym (w tym przez usunięcie oryginalnych danych). | Wnioskowanie o konkretnym podmiocie danych jest możliwe, jeśli w jednym zbiorze lub w kilku zbiorach do pseudonimizacji wykorzystywane są te same atrybuty, lub jeśli pseudonimy nie ukrywają tożsamości w wystarczający sposób. |

Najczęściej popełniane błędy w technikach pseudonimizacji to:

- **Uznanie pseudonimizacji za anonimizację:** powszechnie panuje opinia, że usunięcie jednego lub kilku atrybutów danych lub zastąpienie ich pseudonimizowanymi danymi jest wystarczające do ochrony prywatności danej osoby. Należy podjąć dodatkowe działania mające na celu zwiększenie anonimizacji zbioru danych poprzez uogólnianie lub agregację atrybutów w takim stopniu, by niemożliwa była identyfikacja danego podmiotu danych.
- **Użycie tego samego klucza w różnych bazach:** zastosowanie za każdym razem innego klucza do pseudonimizacji danych pozwala w znacznym stopniu ograniczyć możliwości tworzenia powiązań między bazami, a tym samym zmniejsza ryzyko identyfikacji podmiotu danych. Stosowanie tego samego klucza w różnych bazach znacznie podwyższa ryzyko naruszenia prywatności.
- **Używanie różnych kluczy w bazie:** zastosowanie w jednej bazie niejednolitego klucza (np. w przypadku dodania lub zmiany wpisów dla jednego z podmiotów danych) może spowodować powstawanie unikalnych wzorców, co skutkuje dodatkową informacją na temat podmiotu, która może umożliwić jego identyfikację.

**Przechowywanie klucza:** w przypadku przechowywania w jednym miejscu klucza wraz z danymi źródłowymi, ewentualne naruszenie bezpieczeństwa (np. wykradzenie bazy wraz z kluczem) skutkuje bardzo szybkim powiązaniem danych i ujawnieniem tożsamości podmiotów danych, których dane dotyczą. Należy także pamiętać, że przechowywanie klucza w sposób niezapewniający wysokiej ochrony może wywołać ten sam skutek.

## Podsumowanie technik

Obecnie techniki depersonalizacji danych osobowych są intensywnie rozwijane. Powyżej opisane metody są to najczęściej wykorzystywane i rekomendowane do ochrony prywatności. Należy jednak pamiętać, że każdy zbiór danych należy rozpatrywać osobno, a po analizie ryzyka dostosować wybraną technikę do zakresu danych i celu, w jakim dane mają być udostępnione.

*Tabela 14. Podatność na ryzyka metod anonimizacji i pseudonimizacji danych*

| Technika depersonalizacji danych | Ryzyko wyodrębnia? | Ryzyko tworzenia powiązań? | Ryzyko związane z wnioskowaniem? |
|---|---|---|---|
| Dodawanie zakłóceń | Tak | Być może nie | Być może nie |
| Permutacja | Tak | Tak | Być może nie |
| Agregacja lub k-anonimizacja | Nie | Tak | Tak |
| L-dywersyfikacja | Nie | Tak | Być może nie |
| Prywatność różnicowa | Być może nie | Być może nie | Być może nie |
| Tokenizacja | Tak | Tak | Być może nie |
| Pseudonimizacja | Tak | Tak | Tak |