Operačný program
Efektívna
verejná správa

Európska únia
Európsky sociálny fond

MINISTERSTVO
INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA
A INFORMATIZÁCIE
SLOVENSKEJ REPUBLIKY

Výstup č. 5.2.1

Nastavenie pravidiel pre anonymizáciu osobných údajov v datasetoch

Zmluva o dielo č. RZ 445/2022

Projekt:

Zlepšenie využívania údajov vo verejnej správe

ITMS kód projektu:

314011S979

KPMG

Tento projekt je podporený z Európskeho sociálneho fondu.

Preskúmanie a schválenie dokumentu

História verzií

Verzia	Autor	Dátum	Poznámka

Tento dokument skontroloval

	Meno	Dátum kontroly
1		
2		
3		
4		
5		

Tento dokument schválil

	Meno	Dátum schválenia
1		
2		
3		
4		
5		

Tento projekt je podporený z Európskeho sociálneho fondu.

Zoznam skratiek

Skratka	Význam
AEPD	Španielska agentúra na ochranu údajov
CIP[1]	Centrálna integračná platforma
CoE	Centrum excelentnosti (z angl. Center of Excellence)
CRZ	Centrálny register zmlúv
EDPS	Európsky dozorný úrad pre ochranu údajov
ELISE	Európska sieť centier excelencie umelej inteligencie
EMD	Vzdialenosť hýbateľa zeme - Metóda na vyhodnotenie odlišnosti medzi dvoma viacrozmernými distribúciami v priestore prvkov (z angl. Earth Mover's Distance)
GDPR	Všeobecné nariadenie o ochrane údajov
HIPAA	Štandard pre ochranu súkromia a bezpečnosť pacientov a ich zdravotných záznamov (z angl. The Health Insurance Portability and Accountability Act)
INSPIRE	Infraštruktúra pre priestorové informácie v Európe
ISVS	Informačný systém verejnej správy
KAV[2]	Konsolidovaná analytická vrstva, bližšie na uvedenom odkaze
MIRRI	Ministerstvo investícii, regionálneho rozvoja a informatizácie
MIT	Massachusettský technologický inštitút (z angl. Massachusetts Institute of Technology)
NCZI	Národné centrum zdravotníckych informácií
(NZ)	Nezverejniteľné údaje
OVM	Orgán verejnej moci
PaaS[3]	Platforma ako služba (z angl. platform as a service), tu používané ako súčasť projektu Rozvoj platformy integrácie údajov (centrálna integračná platforma) a Manažment osobných údajov, bližšie na uvedenom odkaze
PII	Osobné identifikovateľné informácie (z angl. Personal Identifying Information)
PSČ	Poštové smerovacie číslo
(Z)	Zverejniteľné údaje

[1] https://metais.vicepremier.gov.sk/refid/id/egov/project/346?tab=basicForm
[2] https://metais.vicepremier.gov.sk/detail/Projekt/15aae7d3-7149-4b5b-b606-9efd56cc8114/cimaster?tab=documentsForm
[3] https://metais.vicepremier.gov.sk/refid/id/egov/project/346?tab=documentsForm

Obsah

Tento projekt je podporený z Európskeho sociálneho fondu.

Tento projekt je podporený z Európskeho sociálneho fondu.

1 Úvod a zhrnutie

1.1 Kontext

Detailný výstup č. 5.2.1: Nastavenie pravidiel pre anonymizáciu osobných údajov v datasetoch vznikol ako rozšírenie výstupu 1.1.5: Štandardizácia anonymizácie údajov vo vzťahu k otvoreným údajom, so zameraním na osobné údaje v datasetoch.

Dokument bol pripravený v rámci projektu „Zlepšenie využívania údajov vo verejnej správe“. Tento projekt má ambíciu transformovať fungovanie inštitúcií verejnej správy tak, aby dokázali maximálne efektívne spravovať a zdieľať údaje, využívať údaje pre lepšie rozhodovanie na základe faktov a dôkazov, pre zlepšenie efektivity a adresnosti služieb na základe lepšieho využívania dát.

Projekt Zlepšenie využívania údajov vo verejnej správe realizuje Dátová kancelária verejnej správy ako špeciálna jednotka Ministerstva investícii, regionálneho rozvoja a informatizácie (ďalej aj MIRRI).

Obsahom dokumentu je súhrn metód, algoritmov a odporúčaní anonymizovania údajov s cieľom ich publikovania ako otvorených údajov. Výstup vznikol ako realizácia aktivity číslo 5 Podpora zvýšenia rozsahu a kvality otvorených údajov, navrhuje bezpečné metódy anonymizovania údajov s cieľom ich zverejnenia ako otvorených údajov a zároveň prezentuje využitie vybraných metód prostredníctvom algoritmov na syntetickom dátovom súbore.

1.2 Metodika realizácie výstupu

Dokument pokrýva tieto oblasti v rámci anonymizovania údajov s cieľom ich publikovania ako otvorených údajov:

- Analýza problematiky publikovania anonymizovaných otvorených údajov, v rámci ktorej sa venuje znovu-identifikácii citlivých údajov, kategorizácii citlivých údajov a riziku znehodnotenia údajov ich anonymizovaním
- Preskúmanie cieľov, ktoré je možné dosiahnuť zavedením anonymizovania, vrátane rozšírenia množstva otvorených údajov, ktoré sú publikované
- Metódy anonymizovania a ich využiteľnosť pri anonymizácii údajov s cieľom ich zverejňovania ako otvorených údajov
- Návrh správy, dohľadu a riadenia procesov anonymizácie
- Preskúmanie použitia anonymizácie pre vybrané prípady použitia a prezentovanie príkladov dobrej praxe zo zahraničia
- Príklad využitia algoritmu k-anonymity využitím Python na dosiahnutie k-5 anonymity pre prípad zdravotných údajov.
- Rámec na výber anonymizačného nástroja a možnosti fungovania v prostredí ISVS

Tento projekt je podporený z Európskeho sociálneho fondu.

2 Problematika anonymizovania údajov s cieľom ich publikovania ako otvorených údajov

Anonymizovanie údajov je proces, pri ktorom sa odstraňujú osobné údaje a iné údaje citlivej povahy, čím sa zamedzuje identifikácii konkrétnej osoby. V prípade anonymizovania osobných údajov v dátových súborov, ktoré majú byť publikované ako otvorené údaje je dôležité pristúpiť k anonymizovaniu obzvlášť zodpovedne, keďže v prípade zle vykonaného anonymizovania údajov a následného úniku údajov vnikajú vážne problémy.

V prípade anonymizovania údajov s cieľom ich zverejnenia ako otvorených údajov je nutné brať do úvahy množstvo premenných, ktoré musia byť nastavené správne, aby boli údaje po anonymizovaní zabezpečené.

Anonymizovanie údajov prináša mnoho výhod, ktoré umožňujú uchovávanie resp. dlhodobé skladovanie dát s menším rizikom porušenia zákonov na ochranu osobných údajov. Aj napriek anonymizácii však v prípade zverejňovania dát v ktorých boli anonymizované osobné a citlivé údaje vznikajú riziká, kedy môže dôjsť prostredníctvom spätnej identifikácie k úniku osobných údajov a následnej identifikácii osôb a citlivých informácii. Spätná identifikácia je teda jednou z kľúčových tém pri zverejňovaní otvorených dát, v ktorých boli anonymizované údaje.

Po publikovaní otvorených údajov, v ktorých boli anonymizované osobné údaje a iné údaje citlivej povahy, vlastník stráca kontrolu nad dátami a nie je schopný technickými prostriedkami obmedziť možné pokusy o spätné identifikovanie. Kvôli tomu je nutné vykonať anonymizovanie údajov s vedomím toho, aké hrozby spätnej -identifikácie existujú a aké techniky je možné využiť, aby riziko spätnej identifikácie bolo minimalizované.

2.1 Spätná identifikácia osobných údajov

Spätná identifikácia osobných údajov a akýchkoľvek iných citlivých údajoch, ktoré boli anonymizované predstavuje závažnú hrozbu. V princípe snahy o spätná identifikáciu využívajú 3 slabiny v rámci anonymizovaných dát:

- Nedostatočné anonymizovanie
- Zvrátenie pseudonymizácie
- Prepojenie viacerých záznamov

Zneužitiu týchto slabín je možné predchádzať a tým znižovať riziko spätnej identifikácie údajov. Pre anonymizovanie údajov s minimálnym rizikom ich spätnej identifikácie je nutné pochopiť, ako aplikovať rôzne metódy anonymizovania na rozličné typy údajov. Toto je možné dosiahnuť pochopením jednotlivých prípadov využitia anonymizácie a zároveň znalosťou hrozieb spätnej identifikácie a spôsobov na ich odstránenie. Bližšie jednotlivé hrozby spätnej identifikácie rozoberá výstup 1.1.5 Štandardizácia anonymizácie údajov, v ktorom sú navrhnuté aj postupy, ako je možné spätnej identifikácii predchádzať.

2.1.1 Nedostatočné anonymizovanie

Cieľom anonymizovania osobných údajov je vytvorenie záznamov, ktoré môžu byť zverejnené ako otvorené dáta. Pri vykonávaní anonymizácie je nutné zvážiť, či nie je nutné anonymizovať aj údaje, ktoré síce nespadajú do kategórie osobných ale svojou povahou, netradičnosťou alebo obsahom môžu v prípade skombinovania s inými údajmi viesť k spätnej identifikácii osôb alebo ohrozeniu utajovaných skutočností.

Za nedostatočne anonymizované údaje môžu byť považované aj tie, na ktoré bola použitá nesprávna technika anonymizovania s cieľom zachovať čo najväčšiu využiteľnosť otvorených dát. Zachovať otvorené dáta vo forme využiteľnej pre analytické účely alebo iné prípady využitia je dôležité, avšak môže sa tak stať na úkor ľahšej spätnej identifikácie. Je preto nutné vždy zvážiť využitie všetkých rozumne dostupných možností anonymizácie pre minimalizovanie tohto rizika.

2.1.2 Zvrátenie pseudonymizácie

V prípade, že v rámci záznamov, v ktorých sú dáta nie len anonymizované ale aj pseudonymizované môže dôjsť k zvráteniu pseudonymizácie na základe odhalenia pseudonymizačného kľúča alebo aj pomocou iných techník. Vo všeobecnosti cieľom útočníka je:

- získanie pseudonymizačného tajomstva („secret“),
- úplná opätovná identifikácia dotknutej osoby,
- čiastočné rozoznanie dotknutej osoby či skupiny („discrimination“).

Bližšie je táto problematika vysvetlená vo 8. kapitole výstupu 1.1.5 Štandardizácia anonymizácie údajov.

2.1.3 Prepojenie viacerých záznamov

Keďže predmetom tohto dokumentu je anonymizácia údajov za cieľom ich zverejňovania, je dôležité brať do úvahy aj hrozbu prepojenia viacerých záznamov. Pri zverejňovaní veľkého množstva otvorených údajov vzniká riziko, že v rámci dát sa budú objavovať údaje, ktoré bude možné kombinovať naprieč jednotlivými záznamami, a zároveň navyše existuje možnosť, že útočníci môžu využiť údaje z iných zdrojov. Pokiaľ sa tieto dáta podarí skombinovať existuje riziko, že útočníci nadobudnú čiastočné alebo úplné informácie o konkrétnej osobe alebo skupine osôb.

V súčasnosti sú napríklad voľne dostupné záznamy o konateľoch a majiteľoch firiem v rámci obchodného registra. Pokiaľ by v rámci otvorených údajov útočník dokázal objaviť prepojenia medzi takýmito zdrojmi údajov, vznikol by závažný problém úniku citlivých údajov. Príkladov záznamov typu obchodný register, ktoré môžu byť využité pri spájaní s otvorenými údajmi existuje viacero. Útok na citlivé údaje pomocou prepájania údajov teda hrozí vždy v prípade zle vykonanej anonymizácie.

2.2 Klasifikácia citlivosti údajov

Pri klasifikácii citlivosti údajov sa stretávame s 3 základnými triedami, ktoré priamo určujú ich citlivosť:

Tento projekt je podporený z Európskeho sociálneho fondu.

2.2.1 Citlivé údaje podľa zákona o ochrane osobných údajov

Prvá trieda citlivosti je určená Zákonom č. 18/2018 Z. z. o ochrane osobných údajov a GDPR, ktorý vymedzuje údaje z pohľadu zamedzenia identifikácie fyzickej osoby, ktoré nesmú byť zverejnené:

— Osobné údaje – teda každý údaj, ktorý identifikuje alebo umožňuje identifikovať fyzickú osobu

— Osobné údaje osobitných kategórií

2.2.2 Citlivé údaje podľa osobitných zákonov

Ako sa uvádza vo výstupe 1.1.5 štandardizácii anonymizácie údajov, druhou triedou sú skupiny údajov, na ktoré sa vzťahujú špecifické osobitné politiky správy aktív, pretože majú zásadný vplyv na ochranu osobných údajov alebo iných údajov špecifického charakteru považovaných za citlivé. Na základe toho aj v prípade prípravy údajov na zverejnenie ako otvorených dát existujú dve úrovne citlivých údajov:

— zverejniteľné údaje (Z) – údaje, ktorých zverejnenie neohrozuje fungovanie štátu a jeho systémov, a preto ich je možné kedykoľvek zverejniť.

— nezverejniteľné údaje (NZ) – údaje, ktoré nie je vhodné zverejniť v žiadnom prípade, pretože zverejnenie nesie riziko okamžitého alebo neskoršieho pokusu o narušenie informačnej bezpečnosti. Pre nezverejniteľné údaje môže existovať podmienka, za splnenia ktorej sa stanú zverejniteľnými.

Podľa znenia legislatívnej požiadavky zákona č. 95/2019 Z. z. sa v dokumente štandardizácii anonymizovania údajov navrhuje vytvorenie podkategórií nezverejniteľných údajov (NZ), ktoré budú nastavovať štandard pre stupnicu citlivosti nezverejniteľný informácii. Táto stupnica zároveň poskytne aj rámec, na základe ktorého bude možné zaradenie údajov podľa tejto stupnice a následný prístup k údajom v jednotlivých stupňoch citlivosti z hľadiska anonymizovania a pseudonymizácie.

Pre nezverejniteľné údaje boli vytvorené tieto stupne citlivosti:

- **Vyhradené údaje** – Osobné alebo iné údaje, ktoré podliehajú najprísnejším požiadavkám na spracovanie vzhľadom na ich citlivosť a riziko pre organizáciu a zákazníkov v prípade nesprávnej manipulácie.

- **Dôverné údaje** – Osobné alebo iné údaje, ktoré podliehajú prísnym požiadavkám na spracovanie vzhľadom na ich citlivosť a riziko v prípade nesprávneho zaobchádzania.

- **Interné údaje** – Údaje, ktoré sú zamestnancom a prípadným tretím stranám k dispozícii na základe zmluvy/dohody o (zachovaní) mlčanlivosti výlučne v dôsledku ich zamestnania v organizácii alebo prebiehajúceho projektu či poskytovania služby a ktoré nie sú kategorizované ako dôverné alebo vyhradené.

V rámci výstupu 1.1.5 boli do týchto stupňov citlivosti zatriedené jednotlivé druhy údajov z pohľadu klasifikácie osobných údajov na základe GDPR. Avšak je taktiež nutné klasifikovať do týchto kategórií aj údaje, ktoré sú určené ako citlivé podľa osobitných zákonov, touto problematikou sa hlbšie zaoberá výstup 1.1.3 štandardizácia pre

Tento projekt je podporený z Európskeho sociálneho fondu.

bezpečnosť a ochranu údajov. Z pohľadu osobitných zákonov budú v rámci stupnice citlivosti zatriedené aj tieto typy údajov:

— Utajované skutočnosti

— Informácie o kritickej infraštruktúre

— Daňové tajomstvo

— Iné údaje podľa osobitých zákonov

2.2.3 Iné citlivé údaje

Iné citlivé údaje nespadajú do kategórie osobných údajov ani osobných údajov osobitných kategórií, tieto údaje taktiež nie sú zahrnuté ani ako citlivé podľa osobitných zákonov. Ide teda o údaje, ktoré sú citlivé na základe inej skutočnosti, akou môže byť fakt, že obsahujú prvky, na základe ktorých je možné identifikovať vzájomné vzťahy medzi rôznymi dátovými súbormi, čo môže viesť k následnému odhaleniu citlivých údajov podľa osobitných zákonov alebo k identifikácii fyzickej osoby. Taktiež môže ísť o údaje, ktoré sa môžu stávať citlivými údajmi iba v špecifických prípadoch.

2.3 Znehodnotenie otvorených údajov

Vykonaním anonymizovania dochádza prirodzene k strate niektorých údajov z pôvodných záznamov. Toto je spôsobené aplikáciou jednotlivých metód anonymizovania na vybraný záznam, z ktorého sú odstránené určité údaje, iné sú zas umelo pretvorené na nepresné, čím sa vytvára viacero rovnakých hodnôt a znižuje sa teda pravdepodobnosť identifikácie.

V nasledujúcich podkapitolách sú priblížené jednotlivé modely merania ochrany osobných údajov, každý z týchto modelov má rôzne výhody a nevýhody v oblastiach ochrany citlivých údajov a uchovania presnosti údajov. Tieto modely pre dosiahnutie určitej úrovne anonymizácie využívajú rôzne metódy anonymizovania, a teda podľa ich výberu dochádza aj k rôznemu znehodnoteniu presnosti a kvality záznamov na ich následné druhotné využitie. Okrem nižšie spomenutých modelov existujú aj ďalšie, avšak ich využiteľnosť pri anonymizovaní údajov s cieľom publikovania, ako otvorených dát nemusí byť dostatočná vzhľadom na to, že nedokážu poskytnúť dostatočnú ochranu citlivých údajov.

2.3.1 K-anonymita

K-anonymita je jedným z modelov merania stupňa ochrany údajov. Na jeho základe je možné určiť, do akej miery sú údaje v dátovom súbore anonymizované. Ide teda o model, na základe ktorého je možné povedať do akej miery bolo realizovanie anonymizácie úspešné. Keď teda v tomto prípade hovoríme o k-anonymite, stupeň „K“ stanovuje, aké percento podobných údajov je možné nájsť pri zanechaní určitej presnosti záznamu.

Bližšie je k-anonymita rozpísaná vo výstupe 1.1.5 Štandardizácia anonymizácie údajov, avšak vo všeobecnosti platí, čím vyššia úroveň k-anonymity (vyjadrená vyšším K) je aplikovaná na údaje pri zachovaní súkromia cez podobnosť údaju blížiacej sa k 100%, tým dochádza k väčšej nepresnosti údajov, a teda ich znehodnoteniu. Na využiteľnosť

otvorených dát môže mať nepresnosť údajov zanesená anonymizovaním značný vplyv, avšak do veľkej miery nie je možné sa týmto nepresnostiam vyhnúť, keďže zachovanie ochrany citlivých údajov je prioritou.

2.3.2 L-diverzita

Ďalším modelom merania úspešnosti (stupňa) ochrany údajov je L-diverzita, ktorá rozširuje K-anonymitu a je jej nadstavbou, ale zároveň v niektorých prípadoch môže byť použitá aj samostatne. V prípade, keď sa pracuje s menej citlivými informáciami je L-diverzitu možné využiť samostatne na meranie stupňa anonymity a teda jej aplikáciou je častokrát možné zachovať väčšiu presnosť v údajoch, ale zároveň aj zabezpečiť dostatočnú ochranu citlivých údajov. V kapitole 7.2 a vo výstupe 1.1.5 Štandardizácia anonymizácie údajov je bližšie vysvetlené praktické využitie tohto modelu.

3 Ciele zavedenia anonymizácie

Anonymizáciou sledujeme najmä zväčšenie množstva otvorených údajov, ktoré sú publikované za súčasného zachovania dostatočného stupňa ochrany citlivých údajov. Tým, že sa rozšíri dostupná báza otvorených údajov sa zároveň podporí dosiahnutie dôležitých cieľov zverejňovania otvorených údajov, najmä:

— Zlepšenie transparentnosti a zodpovednosti: Vďaka zverejňovaniu ešte väčšieho množstva dát môžu občania a iné zainteresované strany lepšie pochopiť, ako vláda a verejné inštitúcie pracujú a ako sa využívajú verejné zdroje. Toto môže zvýšiť dôveru verejnosti v štátnu správu a pomôcť jej lepšie plniť svoje úlohy.

— Podpora inovácií a rozvoja: Výrazné rozšírenie množstva zverejňovaných dát môže pomôcť podnikateľom, výskumníkom a ďalším zainteresovaným stranám zlepšiť svoje služby a výrobky. Vďaka lepšiemu prístupu k dátam môžu vznikať nové inovatívne riešenia, ktoré by inak neboli možné. Ako veľmi populárny cieľ je dnes prezentovaný rozvoj digitálnej ekonomiky (pozri aj výstup 5.1.1 Štandardy pre zverejňovanie údajov verejnej správy vo formáte otvorených údajov k téme merania dopadu otvorených údajov na ekonomiku).

— Zlepšenie efektívnosti a kvality verejných služieb: Viac dát môže pomôcť verejným inštitúciám lepšie plánovať a poskytovať svoje služby, čo môže viesť k zlepšeniu efektívnosti a kvality verejných služieb.

— Zlepšenie rozhodovania: Lepší prístup k dátam môže viesť k lepšiemu plánovaniu a rozhodovaniu o politikách a opatreniach, ktoré bude založené na faktoch a údajoch, a nie iba na názoroch a predpokladoch,

Anonymizácia citlivých dát môže pomôcť dosiahnuť tieto ciele tým, že zabezpečí ochranu súkromia a osobných údajov. To môže pomôcť zvýšiť dôveru verejnosti v zverejňovanie otvorených dát a zabezpečiť, že sa tieto dáta budú využívať iba na legitímne účely. Detailnejšie je o tejto téme pojednané v rámci nasledujúcich podkapitol 3.1, 3.2 a 3.3

Je dôležité poznamenať, že v súčasnosti v štátnej správe neexistuje jednotný prístup k anonymizácii, a takisto ani spoločná správa tohto procesu (governance). Viac k dôvodom zložitosti dosiahnutia cieľov pomocou komplexného riešenia na anonymizovanie sa venuje kapitola 8.

3.1 Zamedzenie únikov citlivých informácii

Ako už bolo spomenuté vyššie, v súčasnosti nie je zavedený jednotný štandard anonymizovania údajov a zároveň mnohé súčasné informačné systémy verejnej správy nie sú pripravené na príchod takéhoto riešenia. Vzniká teda oprávnené obava z úniku citlivých údajov. Neprospieva ani fakt, že ochrana osobných a citlivých údajov je aj interne v rámci orgánov verejnej moci len veľmi slabo regulovaná. Môže sa teda stať, že pokiaľ si určitý orgán verejnej moci vyžiada prístup k datasetom, obsahujúcim aj citlivé informácie, môže ho dostať v plnom rozsahu aj napriek tomu, že pri danom konkrétnom prípade využitia niektoré informácie nie sú potrebné a mohli byť anonymizované alebo v niektorých prípadoch aj pseudonymizované. Tým by sa riziko úniku mitigovalo oveľa lepšie, čo by zároveň umožnilo poskytnúť inak neposkytované údaje a získať inak nedostupné prínosy spracovania otvorených dát.

Ak teda chceme, aby OVM zverejňovali čo najviac údajov, treba im v prvom rade pomôcť s pochopením, ako by takýto proces mal byť vykonaný, a teda ako anonymizovať citlivé údaje. Okrem iného je dôležité, aby bolo k dispozícii dostatočné technické vybavenie a boli správne nastavené príslušné procesy, roly a zodpovednosti. Už len takáto základná znalosť a správne nastavenie manažmentu dát dokáže výrazne dopomôcť k zamedzeniu úniku citlivých informácií prostredníctvom publikovania otvorených dát.

Otvorené dáta sú okrem iných čŕt definované aj ako strojovo spracovateľné. Avšak, keď ide o anonymizovanie je nutné sa pozrieť aj na neštruktúrované údaje a to hneď z viacerých dôvodov.

— Prvým z nich je, že vo verejnej správe dochádza k zverejňovaniu množstva neštruktúrovaných údajov napríklad aj vo forme zmlúv, ktoré obsahujú rôzne citlivé informácie.

— Ďalším dôvodom je aj fakt, že takéto zverejnené neštruktúrované údaje je relatívne jednoduché dostať súčasnými technológiami do štruktúrovanej podoby, kde v prípade zle anonymizovaných neštruktúrovaných údajov získa útočník citlivé informácie, ktoré môže použiť k ďalším útokom na spätné identifikovanie osobných informácií alebo utajovaných skutočností.

Je teda zrejmé, že pri anonymizovaní citlivých údajov za cieľom ich publikovania ako otvorených údajov by sme sa mali zaoberať aj údajmi, ktoré sú publikované v neštruktúrovanej podobe.

Ako sa uvádza vo výstupe 1.1.5 Štandardizácia anonymizácie údajov, väčšina údajov, ktoré sa v súčasnosti nachádzajú v organizáciách nie sú štruktúrované, preto by sa nemalo zanedbávať nastavenie pravidiel pre ich anonymizovanie. Ako príklad vezmime Centrálny register zmlúv (www.crz.gov.sk). V tomto prípade sú neštruktúrovanými údajmi zmluvy, ktoré majú jednotlivé verejné inštitúcie povinnosť zverejňovať prostredníctvom tohto registra. Takto zverejnené zmluvy často obsahujú množstvo citlivých údajov, akými sú napríklad osobné údaje, v niektorých prípadoch (i keď zriedkavo) aj citlivé údaje podľa osobitných zákonov.

Jednotlivé verejné inštitúcie sú povinné zamedziť úniku citlivých informácii pri zverejnení takýchto zmlúv, avšak nie vždy sú pri tom úspešné. Úniky pramenia často z dôvodu, že neexistujú všeobecne prijaté pravidlá anonymizácie, a tak môže dochádzať k tomu, že neodborne vykonaná anonymizácia je ľahko zvrátiteľná a je možné odhaliť citlivé údaje.

Jedným z cieľov zavedenia anonymizácie by malo byť teda aj vytvorenie resp. zjednotenie pravidiel, ktoré by mali byť využité pri anonymizovaní neštruktúrovaných údajov. Vytvorením týchto pravidiel by verejné inštitúcie získali podklad k tomu, ako správne vykonávať anonymizovanie, čím by sa výrazne zamedzili úniky citlivých dát na základe neodborného anonymizovania. Príklady takýchto únikov a pravidlá anonymizovanie neštruktúrovaných dát sa nachádzajú v kapitole 6.2.

3.2 Rozšírenie množstva otvorených údajov

Zväčšovanie množstva otvorených údajov je všeobecnou snahou v rámci aktivít dátovej kancelárie, ktorá koná na základe európskej iniciatívy, ktorej cieľom je zverejnenie čo najväčšieho množstva dát, ktoré sú považované za všeobecne prospešné pre verejnosť z rôznych dôvodov. Avšak existuje aj časť dát, ktoré nie sú zverejniteľné ako otvorené dáta v ich zdrojovej podobe, keďže obsahujú citlivé dáta.

Tento projekt je podporený z Európskeho sociálneho fondu.

Dáta obsahujúce citlivé údaje pred ich publikovaním ako otvorených dát musia prejsť anonymizovaním. Na anonymizovanie údajov, ktoré budú publikované ako otvorené údaje pri zohľadnení súčasného stavu manažmentu údajov v štátnej správe sa výrazne znižuje počet metód anonymizovania, ktoré je možné využiť na tento účel. Vhodné metódy pre anonymizovanie údajov s cieľom publikovania otvorených dát sú rozpísané v kapitole 4.

3.3 Analýzy dátových súborov s citlivými údajmi

Analýzam údajov v dátových súbor obsahujúcich citlivé údaje musí predchádzať anonymizácia, ktorá zaručí ochranu týchto citlivých údajov. K tejto problematike sa čiastočne vyjadruje výstup 1.1.5 Štandardizácia anonymizácie údajov, ktorý opisuje prípad využitia analyzovanie údajov v rámci Konsolidovanej analytickej vrstvy[4] (KAV). Usudzujúc podľa dostupných informácií sa však analyzovanie otvorených údajov a analyzovanie údajov v rámci KAV od seba líši najmä tým, že v rámci Konsolidovanej analytickej vrstvy je po celú dobu nad dátami držaná určitá kontrola a nie sú verejne dostupné.

Cieľom anonymizovania dát v prostredí KAV je zamedziť spätnej identifikácii citlivých údajov, ktoré boli pseudonymizované v prípadoch, keď bude analyzovaných viacero prepojených dátových súborov. V pseudonymizovaných údajoch je zámerne zanechaný identifikačný kľúč, ktorý podporuje analyzovanie údajov naprieč viacerými súbormi, z čoho hrozí riziko, že pri analyzovaní viacerých súborov naraz bude možné pospájať pseudonymizované údaje tak, že budú odhalené osobné a citlivé údaje. Preto je dôležité v rámci KAV zaviesť aj anonymizovanie časti atribútov v dátových súboroch pre minimalizáciu tejto hrozby. Zároveň anonymizovanie dát v rámci KAV nemusí byť až tak striktné, keďže dáta sú udržované neustále pod kontrolou technickými opatreniami a teda tým, že sa využije menej striktné anonymizovanie dáta si zachovajú lepšiu využiteľnosť pri analýzach.

V prípade otvorených dát je situácia výrazne odlišná, je nutné myslieť na to, že dáta sú voľne dostupné a údaje, ktoré v nich budú ponechané je možné rôzne spájať, analyzovať a zároveň na tieto dáta bude možné aplikovať pokročilé technológie na odhalenie citlivých informácii.

Ako vyplýva z predtým uvedených skutočností, na anonymizáciu dát s cieľom ich publikovania ako otvorených dát je nutné použiť čo najstriktnejšie metódy anonymizovania, aby sme zamedzili úniku citlivých informácií. To znamená, že je nutné anonymizovať v týchto dátových súboroch okrem iného aj všetky referenčné identifikátory, ktoré by umožnili prepájanie rôznych dátových súborov. Takéto striktné anonymizovanie je nutné aj kvôli absencii štruktúrovaného manažmentu dát, ktorý by podporoval komplexné riešenie anonymizácie. Z hľadiska anonymizácie sa štruktúrovaný manažment dát chápe ako systém manažmentu dát, na základe ktorého je možné identifikovať vzťahy medzi dátami, mať pod kontrolou ich publikovanie a disponovať platformou / nástrojom, ktorý zabezpečí kontrolu prepojiteľnosti dátových súborov plánovaných na zverejnenie ako otvorených údajov s inými už publikovanými.

[4] https://metais.vicepremier.gov.sk/detail/Projekt/15aae7d3-7149-4b5b-b606-9efd56cc8114/cimaster?tab=documentsForm

4 Definícia metód anonymizácie otvorených údajov

Správne pristupovanie k akémukoľvek procesu anonymizácie je veľmi dôležité, je potrebné správne pochopiť jeho dôsledky a obmedzenia. Existuje sedem základných princípov Ochrany súkromia už od návrhu (Privacy by Design[5]), ktoré predstavila Dr. Ann Cavoukian, bývalá komisárka pre informácie a súkromie z Ontária v Kanade. Ide o tieto zásady:

1. Proaktívne, nie reaktívne; preventívne, nie nápravné: Tento princíp zdôrazňuje dôležitosť predchádzania problémom týkajúcim sa ochrany súkromia skôr, ako by sa riešili následky. Miesto riešenia problémov až po vzniku je lepšie identifikovať a zmierniť riziká vopred.

2. Súkromie ako predvolené nastavenie: Princíp sa zameriava na zabezpečenie, že súkromie používateľov je automaticky chránené, bez toho, aby bolo potrebné vykonávať akékoľvek špeciálne akcie alebo nastavenia zo strany používateľa.

3. Súkromie zabudované do dizajnu: Tento princíp presadzuje, že súkromie by malo byť zohľadnené už počas návrhu a vývoja produktov alebo služieb, a nie pridané dodatočne.

4. Plná funkčnosť - kladný súčet, nie nulový súčet: Ochrana súkromia by nemala byť na úkor funkčnosti alebo výkonnosti. Ide o zabezpečenie, že všetky funkcie a súkromie môžu spoločne existovať bez kompromisov.

5. End-to-end bezpečnosť - ochrana počas celého životného cyklu: Zabezpečenie dát od okamihu, keď sú získané, cez spracovanie až po ich zničenie. Dáta by mali byť chránené počas celého ich životného cyklu.

6. Viditeľnosť a transparentnosť – zachovanie otvorenosti: Organizácie by mali byť otvorené a transparentné v súvislosti so svojimi postupmi ochrany súkromia a bezpečnosti údajov, čím umožňujú používateľom a regulačným orgánom lepšie pochopiť, ako sú ich údaje chránené a spracovávané. Transparentnosť zahŕňa jasné a zrozumiteľné informácie o tom, aké údaje sa zbierajú, akým spôsobom sa spracovávajú, komu sú poskytované a aké opatrenia sa prijímajú na ich ochranu.

7. Rešpektovanie súkromia používateľa – zameranie na používateľa: Tento princíp zdôrazňuje dôležitosť zohľadnenia potrieb a očakávaní používateľov pri navrhovaní a implementácii opatrení na ochranu súkromia. Umožňuje používateľom mať kontrolu nad svojimi údajmi a poskytuje im možnosť spravovať svoje súkromie podľa vlastných preferencií.

Dodržiavaním týchto zásad pri výbere a implementácii metód anonymizácie môžu organizácie výrazne zlepšiť ochranu citlivých údajov a podporiť kultúru súkromia a bezpečnosti.

[5] Ochrana súkromia už od návrhu https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf

Tento projekt je podporený z Európskeho sociálneho fondu.

4.1 Všeobecné metódy anonymizácie

Anonymizácia sa týka procesu maskovania alebo odstraňovania osobných údajov (PII) s cieľom chrániť súkromie jednotlivcov alebo chrániť iné citlivé údaje. Dá sa to označiť aj ako proces mazania osobných údajov[6]. Hlavným cieľom anonymizácie je sťažiť alebo znemožniť opätovnú identifikáciu osôb alebo citlivých informácií. V prípade otvorených údajov by však malo byť cieľom anonymizovania znemožnenie, nielen sťaženie opätovnej identifikácie.

V priebehu rokov sa vyskytlo niekoľko príkladov nesprávne vykonanej anonymizácie údajov, ktoré viedli k opätovnej identifikácii jednotlivcov. Štúdia výskumníkov z MIT a Université Catholique de Louvain[7] zistila, že je možné identifikovať 95 % používateľov mobilných telefónov zo súboru anonymizovaných údajov 1,5 milióna ľudí pomocou iba štyroch referenčných bodov s nízkym priestorovým a časovým rozlíšením. Štúdia využívala informácie o hovoroch od mobilného operátora počas 15 mesiacov, kde každá interakcia so sieťou zaznamenáva umiestnenie pripájacej antény. Štúdia zdôrazňuje potenciálne riziká pre bezpečnosť anonymizovaných súborov údajov a potrebu lepších opatrení na ich ochranu.

Portály s otvorenými údajmi zohrávajú kľúčovú úlohu pri umožňovaní zdieľania údajov a podpore výskumu založeného na údajoch. Ak však údaje nie sú správne anonymizované, môžu predstavovať značnú hrozbu pre súkromie a bezpečnosť jednotlivcov. Portály s otvorenými údajmi v mnohých prípadoch poskytujú prístup k veľkým súborom údajov, ktoré obsahujú osobné informácie, ako sú lokalizačné údaje, finančné záznamy a zdravotné informácie. Hoci takéto údaje možno použiť na hodnotný výskum a analýzu, môžu ich zneužiť aj tretie strany na identifikáciu jednotlivcov a sledovanie ich pohybu.

Je teda nevyhnutné zabezpečiť, aby boli otvorené údaje pred zverejnením na portáloch s otvorenými údajmi riadne anonymizované. To zahŕňa odstránenie akýchkoľvek osobných informácií, ktoré by sa dali použiť na identifikáciu jednotlivcov, ako sú mená, telefónne čísla a adresy, ako aj zhromažďovanie alebo rozmazávanie údajov o polohe, pre sťaženie určenia konkrétnych jednotlivcov.

Spoločný dokument od AEPD a EDPS[8] poskytuje 10 skvelých príkladov nepresného pochopenia procesu anonymizácie:

1. “Pseudonymizácia je to isté ako anonymizácia ” – Pseudonymizácia je technika, ktorá nahrádza identifikačné informácie pseudonymom alebo kódom. Na rozdiel od toho anonymizácia úplne odstraňuje akúkoľvek možnosť identifikácie jednotlivcov z údajov. Pseudonymizované údaje môžu byť aj naďalej osobnými údajmi, keďže ich možno pomocou dodatočných informácií potenciálne vysledovať ku konkrétnej osobe.

2. „Šifrovanie je anomizácia “ – Šifrovanie je technika, ktorá využíva tajné kľúče, aby boli informácie nečitateľné bez dešifrovania. Šifrovanie síce môže znížiť riziko zneužitia údajov, ale úplne neodstráni možnosť identifikácie jednotlivcov. Ak sú k

[6] 10 nedorozumení súvisiacich s anonymizáciou: 21-04-27_aepd-edps_anonymisation_en_5.pdf (europa.eu)
[7] de Montjoye , YA., Hidalgo, C., Verleysen , M. *a kol.* Jedinečné v dave: Hranice súkromia ľudskej mobility. *Sci Rep* **3** , 1376 (2013). https://doi.org/10.1038/srep01376
[8] Ibid 6.

dispozícii tajné kľúče používané na dešifrovanie, zašifrované údaje možno stále spájať s konkrétnymi osobami.

3. „Anonymizácia údajov je vždy možná“ – Anonymizácia nie je vždy možná, najmä ak je súbor údajov malý, obsahuje vysoké úrovne demografických údajov alebo údajov o polohe alebo obsahuje jedinečné alebo zriedkavé údajové body, ktoré môžu identifikovať jednotlivcov.

4. „Anonymizácia je navždy“ – Anonymizácia nie je jednorazový proces a vždy existuje riziko, že sa v budúcnosti bude dať vrátiť späť. Nové technológie, výpočtové zdroje alebo dodatočné informácie môžu narušiť anonymitu predtým anonymizovaných údajov.

5. „Anonymizované údaje sa nikdy nedajú spätne identifikovať“ – Anonymizované údaje možno niekedy spätne identifikovať, najmä ak proces anonymizácie nebol dostatočne spoľahlivý. Techniky, ako je prepojenie údajov alebo interferenčné útoky môžu odhaliť identitu jednotlivcov v súboroch údajov, ktoré boli považované za anonymizované.

6. „Ochrana údajov nie je potrebná pre anonymizované údaje“ – Anonymizované údaje stále podliehajú nariadeniam o ochrane údajov, ak ide o osobné údaje, pretože údaje môžu byť potenciálne prepojené s konkrétnymi jednotlivcami. Prevádzkovatelia a spracovatelia údajov sú zodpovední za to, že anonymizované údaje zostanú anonymné a nebude ich možné opätovne identifikovať.

7. „Anonymizované údaje sú zbytočné“ – Anonymizované údaje môžu byť stále užitočné na určité účely, ako je výskum alebo štatistická analýza, ale užitočnosť údajov musí byť vyvážená rizikom opätovnej identifikácie.

8. „Anonymizáciu je možné vykonať rýchlo a jednoducho“ – Anonymizácia je zložitý a časovo náročný proces, ktorý si vyžaduje odborné znalosti v oblasti analýzy údajov a ochrany súkromia. Neexistuje žiadne univerzálne riešenie pre anonymizáciu , pretože každý súbor údajov a účel si vyžadujú prispôsobené techniky anonymizácie .

9. „Anonymizácia eliminuje všetky riziká ochrany súkromia“ – Anonymizácia môže zmierniť riziká ochrany osobných údajov, ale nemôže ich úplne odstrániť. Stále sa môžu vyskytnúť riziká, ako je porušenie ochrany údajov alebo neoprávnený prístup k údajom a správcovia údajov a spracovatelia musia prijať vhodné opatrenia na ochranu údajov.

10. „Anonymizácia je vždy najlepším riešením na ochranu súkromia“ – Anonymizácia je jednou z niekoľkých techník ochrany súkromia, nie je však vždy tým najlepším riešením. V závislosti od povahy údajov a zamýšľaného účelu môžu byť vhodnejšie iné techniky, ako je minimalizácia údajov, kontrola prístupu alebo šifrovanie.

V tabuľke nižšie je možné vidieť niekoľko najbežnejšie používaných anonymizačných techník.

Tento projekt je podporený z Európskeho sociálneho fondu.

Tabuľka 1 - Anonymizačné techniky

Technika	Popis	Príklad
Potlačenie	Zahŕňa odstránenie alebo skrytie konkrétnych častí údajov zo súboru údajov, aby sa zabránilo identifikácii jednotlivcov.	Odstránenie mien alebo iných identifikačných informácií z verejnej databázy zdravotných záznamov.
Generalizácia	Zahŕňa úpravu údajov nahradením konkrétnych hodnôt všeobecnejšou hodnotou.	Nahradenie presného veku vekovým rozsahom (napr. 25 – 30) v súbore údajov prieskumu.
Agregácia údajov	Zahŕňa kombinovanie údajových bodov do skupín na ochranu súkromia jednotlivca.	Kombinovanie PSČ do väčších geografických oblastí na ochranu súkromia jednotlivcov vo verejnom súbore údajov.
Randomizácia údajov	Táto metóda zahŕňa pridanie náhodného šumu alebo údajových bodov do súboru údajov, aby sa sťažila identifikácia jednotlivcov.	Pridávanie náhodných čísel do súboru údajov lekárskych záznamov, ktoré môžu zakryť diagnózu alebo liečbu jednotlivca.
Výmena dát	Táto technika zahŕňa výmenu určitých dátových polí medzi jednotlivcami v súbore údajov, aby sa sťažila identifikácia jednotlivcov.	Výmena veku jednotlivcov v súbore údajov tak, aby vek už nezodpovedal správnym jednotlivcom.

Tieto tri techniky – potlačenie, generalizácia a agregácia údajov[9] – patria napriek niektorým nedostatkom medzi najúčinnejšie metódy na anonymizáciu údajov pri príprave údajov na zverejnenie na portáloch s otvorenými údajmi, pretože:

— Sú relatívne jednoduché a ľahko implementovateľné: Tieto techniky možno použiť bez toho, aby si vyžadovali značné odborné znalosti alebo zdroje, čo je nevyhnutné pri publikovaní veľkých súborov údajov na portáloch s otvorenými údajmi, kde vlastníci údajov nemusia mať rozsiahle technické zručnosti alebo zdroje.

— Zachovávajú užitočnosť údajov: Často sa požaduje, aby anonymizované údaje boli použiteľné pre výskum a ďalšiu analýzu. Tieto techniky môžu pomôcť zabezpečiť, aby údaje zostali užitočné a zároveň chránili súkromie jednotlivcov.

— Poskytujú vysokú úroveň ochrany osobných údajov: Tieto techniky môžu výrazne znížiť riziko opätovnej identifikácie, aj keď sa používajú v kombinácii.

— Sú široko akceptované a bežne používané: Tieto techniky sú dobre zavedené v oblasti anonymizácie údajov a sú široko akceptované ako najlepšie postupy na ochranu súkromia jednotlivcov.

Všeobecne sú tieto techniky účinné pri vyvažovaní meniacich sa cieľov ochrany súkromia a zachovania užitočnosti údajov, vďaka čomu sú preferované na použitie na portáloch s otvorenými údajmi.

[9] Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization (13. august 2009). UCLA Law Review, roč. 57, str. 1701, 2010, U of Colorado Law Legal Studies Research Paper č. 9-12, dostupné na SSRN: https://ssrn.com/abstract=1450006

Okrem toho budú techniky randomizácie a výmeny údajov diskutované ako príklady nie najlepších techník pri príprave otvorených údajov na zverejnenie vzhľadom na ich neefektívnosť pri ochrane osobných údajov a znižovaní kvality údajov.

4.2 Potlačenie (odstránenie atribútov)

Popis:

Jednou z bežne používaných techník pri anonymizácii údajov je potlačenie, ktoré zahŕňa odstránenie alebo maskovanie určitých častí údajov z množiny údajov. Niektoré z bežných techník potlačenia pri anonymizácii údajov spolu s ich výhodami a nevýhodami zahŕňajú:

— Úplné potlačenie: Zahŕňa odstránenie celého záznamu zo súboru údajov, ak obsahuje osobné informácie, ktoré by mohli identifikovať jednotlivca. Ak napríklad množina údajov obsahuje záznam s menom, adresou a dátumom narodenia konkrétnej osoby, celý tento záznam sa odstráni. Na jednej strane je sila tejto techniky v tom, že úplne odstraňuje riziko opätovnej identifikácie, no na druhej strane môže výrazne znížiť užitočnosť súboru údajov.

— Čiastočné potlačenie: Zahŕňa odstránenie alebo maskovanie niektorých, ale nie všetkých citlivých informácií v súbore údajov. Ak napríklad množina údajov obsahuje meno a adresu konkrétnej osoby, meno môže byť zamaskované, pričom adresa zostane nedotknutá. Výhodou tejto techniky je, že zachováva časť užitočnosti súboru údajov. Na druhej strane môže ponechať určité riziko opätovnej identifikácie, čo je zas jej nevýhodou.

— Náhodné potlačenie: Zahŕňa náhodné odstránenie alebo maskovanie údajových bodov zo súboru údajov bez ohľadu na to, či obsahujú osobné informácie alebo nie. Ak napríklad množina údajov obsahuje informácie o nákupoch zákazníkov, niektoré z nákupov môžu byť náhodne odstránené alebo maskované. Aj keď táto technika môže pomôcť zabrániť opätovnej identifikácii a zároveň zachovať celkovú štruktúru súboru údajov, môže znížiť presnosť akýchkoľvek analýz alebo modelov založených na údajoch.

— Potlačenie top-k: Zahŕňa odstránenie alebo maskovanie top k dátových bodov v množine údajov na základe nejakého kritéria, ako je ich frekvencia alebo hodnota. Ak napríklad množina údajov obsahuje informácie o návštevách webových stránok, 10 najčastejšie navštevovaných webových stránok môže byť odstránených alebo maskovaných. Výhodou tejto techniky je, že môže pomôcť zabrániť opätovnej identifikácii a zároveň zachovať niektoré z najdôležitejších informácií v súbore údajov, ale nevýhodou je, že môže byť ťažké vybrať vhodnú hodnotu k.

Kedy metódu použiť: keď sa atribút nevyžaduje v anonymizovanom súbore údajov alebo keď atribút nemôže byť inak vhodne anonymizovaný inou technikou. Táto technika by sa mala použiť na začiatku procesu anonymizácie, pretože v tomto bode je to jednoduchý spôsob, ako znížiť identifikovateľnosť[10].

[10]Sprievodca základnými technikami anonymizácie údajov (Komisia na ochranu osobných údajov Singapur, 2018) https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf

Tento projekt je podporený z Európskeho sociálneho fondu.

Spôsob použitia: odstrániť záznam alebo odstrániť atribút(y) alebo ak je potrebné zachovať štruktúru súboru údajov, vymazať údaje (a prípadne aj hlavičku). Všimnite si, že potlačenie by malo byť skutočné odstránenie (t. j. trvalé), a nie iba „skrytie stĺpca“. Podobne „redigovanie“ nemusí byť dostatočné, ak základné údaje zostanú do istej miery prístupné[11]. Hlavnou výhodou tejto techniky je, že pri trvalom vymazaní atribútu alebo záznamu je nemožné získať informácie[12].

Príklad techniky úplného potlačenia:

Ako príklad použime súbor údajov „Adult Census Income“ z UCI Machine Learning Repository[13]. Tento súbor údajov obsahuje informácie o jednotlivcoch, ako je ich vek, vzdelanie, povolanie, pracovná trieda, rodinný stav atď., spolu s ich úrovňou príjmu (viac alebo menej ako 50 000 USD ročne).

Ak chceme použiť techniku potlačenia, zvolíme potlačenie atribútu „povolania“, aby sme anonymizovali súbor údajov. To znamená, že by sme zo súboru údajov odstránili atribút „povolanie“, čím by sme zabránili útočníkovi identifikovať jednotlivcov na základe ich povolania.

Tu je príklad, ako by sme mohli použiť potlačenie atribútov na množinu údajov „Príjmy zo sčítania dospelých“:

Pôvodný súbor údajov:

Tabuľka 2 - Príklad údajov pred potlačením

vek	kategória práce	vzdelanie	rodinný stav	povolanie	príjem
39	štátna správa	Bakalári	Nikdy nebol ženatý	Administratíva	<=50 tis
50	Živnostník	Bakalári	Ženatý-civ-manžel	Manažér	>50 tis
38	Súkromné	HS-grad	Rozvedený	Psovodi-čistiace prostriedky	<=50 tis
53	Súkromné	11	Ženatý-civ-manžel	Psovodi-čistiace prostriedky	<=50 tis

Potlačená množiny údajov (odstránený atribút povolania):

Tabuľka 3 - Príklad údajov po potlačení

vek	kategória práce	vzdelanie	rodinný stav	príjem
39	štátna správa	Bakalári	Nikdy nebol ženatý	<=50 tis
50	Živnostník	Bakalári	Ženatý-civ-manžel	>50 tis
38	Súkromné	HS-grad	Rozvedený	<=50 tis
53	Súkromné	11	Ženatý-civ-manžel	<=50 tis

Ako vidíme, atribút „povolanie“ bol zo súboru údajov odstránený, čo môže pomôcť chrániť súkromie jednotlivcov v súbore údajov. Je však potrebné poznamenať, že samotné potlačenie atribútov nemusí byť dostatočné na to aby poskytovalo silné záruky anonymity, najmä ak sú v súbore údajov ďalšie atribúty, ktoré možno použiť na identifikáciu jednotlivcov. Preto je dôležité zvážiť ďalšie techniky na ochranu súkromia,

[11] Ibid 10
[12] Marques, Joana Ferreira a Jorge Bernardino. "Analýza techník anonymizácie údajov." KEOD. 2020. https://www.scitepress.org/Papers/2020/101423/101423.pdf
[13] https://archive.ics.uci.edu/ml/datasets/Adult

Tento projekt je podporený z Európskeho sociálneho fondu.

ako je k-anonymita alebo l-diverzita, v spojení s potlačením atribútov, aby sa zabezpečila ochrana súkromia jednotlivcov.

4.3 Generalizácia

Generalizovanie sa používa pri anonymizácii údajov na úpravu pôvodných údajov nahradením konkrétnych hodnôt všeobecnejšími, menej identifikovateľnými hodnotami. Cieľom je chrániť súkromie jednotlivcov a zároveň zachovať užitočné informácie v údajoch.

Techniky generalizácie sa bežne používajú v rôznych kontextoch, kde ide o ochranu osobných údajov. Napríklad v zdravotníctve sa generalizovanie používa na nahradenie konkrétnych lekárskych diagnóz širšími kategóriami s cieľom chrániť súkromie pacienta a zároveň umožniť výskum a analýzu. Vo financiách možno použiť generalizovanie na nahradenie konkrétnych súm v dolároch rozsahmi, aby sa predišlo podvodným aktivitám a zároveň umožnila presná analýza finančných trendov. V rámci orgánov verejnej moci sa generalizovanie môže použiť na nahradenie konkrétnych demografických údajov širšími kategóriami s cieľom chrániť súkromie jednotlivcov a zároveň umožniť presnú analýzu demografických trendov.

Celkovo sú techniky generalizovania dôležitým nástrojom na ochranu súkromia jednotlivcov v údajoch, pričom stále umožňujú užitočnú analýzu a výskum. Je však dôležité dôkladne zvážiť potenciálny vplyv na užitočnosť údajov, pretože generalizovanie príliš veľkého množstva informácií môže obmedziť užitočnosť údajov na analýzu a výskum. Tu je prehľad niektorých techník generalizovania pri anonymizácii údajov spolu s ich výhodami a nevýhodami:

4.3.1 Maskovanie údajov (znakov)

Popis:

Maskovanie údajov (znakov) je zmena znakov hodnoty údajov, napr. pomocou konštantného symbolu (napr. „*“ alebo „x“). Maskovanie je zvyčajne čiastočné, t. j. aplikuje sa len na niektoré znaky v atribúte.

Kedy ju použiť: Keď je hodnotou údajov reťazec znakov a skrytie jej časti je dostatočné na zabezpečenie požadovaného rozsahu anonymity[14].

Ako to použiť: V závislosti od povahy atribútu nahradiť príslušné znaky vybraným symbolom. V závislosti od typu atribútu sa možno rozhodnúť nahradiť pevný počet znakov (napr. čísla kreditných kariet) alebo variabilný počet znakov (napr. e-mailovú adresu) [15].

Príklad techniky maskovania údajov: Tu je príklad toho, ako môže maskovanie e-mailov vyzerať v množine údajov:

[14]Sprievodca základnými technikami anonymizácie údajov (Komisia na ochranu osobných údajov Singapur, 2018) https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf
[15]Ibid 14.

Tento projekt je podporený z Európskeho sociálneho fondu.

Pôvodný súbor údajov:

Tabuľka 4 - Príklad údajov pred maskovaním

Názov	Email	Telefón
John Doe	john.doe@example.com	555-123-4567
Jane Smith	jana.kovac@example.com	555-987-6543
Bob Johnson	bob.johnson@example.com	555-555-1212

Maskovaný súbor údajov:

Tabuľka 5 - Príklad údajov po maskovaní

Názov	Email	Telefón
John Doe	xxxxx.xxxx@example.com	555-123-4567
Jane Smith	xxxxx.xxxxxxxx@example.com	555-987-6543
Bob Johnson	xxx.xxxxxxxon@example.com	555-555-1212

Pôvodné e-mailové adresy boli nahradené maskovanými verziami, ktoré zachovávajú štruktúru pôvodných e-mailových adries, pričom skutočné e-mailové adresy zakrývajú. Pomáha to chrániť súkromie a dôvernosť jednotlivcov, ktorých údaje sú zahrnuté v súbore údajov, pričom sa stále zachováva užitočnosť údajov na účely analýzy a výskumu.

4.3.2 Zoskupovanie údajov

Popis:

Táto technika zoskupuje podobné hodnoty do rozsahov alebo „skupín“. Vekové skupiny môžu byť napríklad rozdelené do rozsahov, ako sú 0 – 10, 11 – 20, 21 – 30 atď. aj keď táto technika zachováva niektoré informácie v údajoch, ako napríklad rozdelenie hodnôt, môže to sťažiť analýzu údajov. Vytváranie takýchto skupín možno použiť v rôznych kontextoch, ako napríklad:

- Časové zoskupovanie: zoskupovanie údajových bodov podľa časových intervalov, ako sú hodinové, denné, týždenné alebo mesačné.
- Geo-priestorové zoskupovanie: zoskupovanie údajových bodov podľa geografických oblastí, ako sú PSČ, mestá, štáty alebo krajiny.
- Numerické zoskupovanie: zoskupovanie údajových bodov podľa číselných rozsahov, ako sú cenové rozsahy, rozsahy platov alebo rozsahy množstva.

Kedy ju použiť: Keď je hodnotou údajov reťazec znakov a skrytie jeho časti je dostatočné na zabezpečenie požadovaného rozsahu anonymity[16].

Ako ju použiť: Najprv určite vhodnú veľkosť a rozptyl skupiny na základe požadovanej úrovne súkromia a granularity údajov. Potom zoskupte podobné údajové body do týchto segmentov a uistite sa, že každý segment obsahuje dostatočný počet údajových bodov

[16]Sprievodca základnými technikami anonymizácie údajov (Komisia na ochranu osobných údajov Singapur, 2018) https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf

Tento projekt je podporený z Európskeho sociálneho fondu.

na zachovanie anonymity. Pri zoskupovaní geo-priestorových údajov zvážte použitie prístupov založených na geo-priestorovom maskovaní na agregovanie informácií o polohe, čím sa síce zníži presnosť údajov ale zvýši sa ochrana citlivých údajov.

Príklad techniky zberu údajov:

Tu je príklad toho, ako môže segmentovanie údajov vyzerať v množine geo-priestorových údajov:

Pôvodný súbor údajov:

Tabuľka 6 - Príklad údajov pred vytvorením skupín

Poloha	Zemepisná šírka	Zemepisná dĺžka
Mesto New York	40,7128	-74,0060
San Francisco	37,7749	-122,4194
Los Angeles	34,0522	-118,2437

Segmentovaná množina údajov:

Tabuľka 7 - Príklad údajov po vytvorení skupín

Poloha	Rozsah zemepisnej šírky	Rozsah zemepisnej dĺžky
Mesto New York	<40,5; 41,0>	<-74,5; -73,5>
San Francisco	<37,5; 38,0>	<-123,0; -122,0>
Los Angeles	<34,0; 34,5>	<-119,0; -118,0>

Súradnice zemepisnej šírky a dĺžky boli nahradené skupinami alebo rozsahmi hodnôt. To sťažuje identifikáciu presnej polohy jednotlivého záznamu a zároveň zachováva celkovú distribúciu údajov. Zrnitosť segmentov možno upraviť na základe požadovanej úrovne anonymizácie a presnosti pôvodných údajov.

4.4 Agregácia údajov

Popis:

Agregácia údajov je v podstate prevod súboru údajov zo zoznamu záznamov na súhrnné hodnoty. Agregácia zahŕňa kombináciu viacerých údajových bodov do jednej súhrnnej hodnoty, zvyčajne použitím štatistických funkcií, ako je priemer, medián, súčet alebo počet. Agregácia údajov znižuje ich granularitu a sťažuje identifikáciu jednotlivých záznamov, môže však viesť aj k strate podrobných informácií. Táto technika sa často používa v scenároch, kde na analýzu postačuje poskytnutie súhrnu alebo všeobecného prehľadu údajov.

Kedy ju použiť: keď sa nevyžadujú jednotlivé záznamy a na daný účel postačujú súhrnné údaje.

Ako ju používať: medzi typické spôsoby patrí použitie súčtov alebo priemerov atď. V prípade potreby dávať pozor na skupiny, ktoré majú po vykonaní agregácie príliš málo záznamov. Môže byť potrebné použiť agregáciu v kombinácii s potlačením. Možno bude potrebné odstrániť niektoré atribúty, pretože obsahujú podrobnosti, ktoré sa nedajú

Tento projekt je podporený z Európskeho sociálneho fondu.

agregovať, a možno bude potrebné pridať nové atribúty, napr. aby obsahovali novo vypočítané súhrnné hodnoty[17].

Príklad techniky agregácie údajov:

Pôvodný súbor údajov:

Tabuľka 8 - Príklad dátového súboru pred agregáciou

ID pacienta	Vek	Pohlavie	Hmotnosť (kg)	Výška (cm)	Krvný tlak (mmHg)	Pulz (bpm)	Diagnóza
1	45	M	80	175	120/80	75	Zdravý
2	38	F	65	160	130/85	72	Zdravý
3	55	M	90	180	150/95	80	Hypertenzia
4	29	F	50	157	115/75	70	Zdravý
5	65	M	85	168	160/100	85	Hypertenzia

Súhrnný súbor údajov (priemer podľa diagnózy):

Tabuľka 9 - Príklad dátového súboru po agregácii

Diagnóza	Priemerný vek	Priemerná hmotnosť (kg)	Priemerná výška (cm)	Priemerný tlak (mmHg)	Pulz (bpm)
Zdravý	37,33	65	164	121,67/80	72,33
Hypertenzia	60	87,5	174	155/97,5	82,5

V tomto príklade sme agregovali množinu klinických údajov podľa stĺpca diagnózy. Vypočítali sme priemerný vek, priemernú hmotnosť, priemernú výšku, priemerný krvný tlak a priemernú srdcovú frekvenciu pre každú kategóriu diagnózy (Zdravie a hypertenzia).

4.5 Nevhodné techniky anonymizácie

Ďalej uvádzame techniky anonymizácie, ktoré nie sú vhodné pri príprave údajov na zverejnenie online na portáli otvorených údajov.

4.5.1 Výmena dát

Výmena dát zahŕňa výmenu hodnôt medzi dvoma alebo viacerými jednotlivcami v súbore údajov, aby sa vytvoril nový súbor údajov, kde hodnoty už nie sú spojené s pôvodnými jednotlivcami. Táto technika zachováva štatistické vlastnosti údajov a zároveň chráni súkromie jednotlivca. Aj keď táto technika môže byť účinná pri ochrane súkromia jednotlivcov, môže tiež spôsobiť značné nepresnosti a nezrovnalosti v údajoch. V dôsledku toho nemusí byť výmena údajov vhodnou technikou na prípravu údajov na zverejnenie na portáli s otvorenými údajmi.

Primárnym účelom portálov s otvorenými údajmi je sprístupniť údaje verejnosti užitočným a informatívnym spôsobom, čo si vyžaduje, aby údaje boli čo najpresnejšie a

[17]Sprievodca základnými technikami anonymizácie údajov (Komisia na ochranu osobných údajov Singapur, 2018) https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(_250118).pdf

Tento projekt je podporený z Európskeho sociálneho fondu.

najkonzistentnejšie. Ak sa údaje výrazne zmenili prostredníctvom výmeny údajov, nemusia byť užitočné alebo informatívne pre používateľov, ktorí sa potrebujú spoliehať na údaje na účely analýzy, výskumu alebo rozhodovania.

Tu je príklad toho, ako môže výmena údajov spôsobiť nepresnosti v množine údajov:

Predpokladajme, že máme súbor údajov o výške a hmotnosti 100 jednotlivcov. Technika výmeny údajov môže zahŕňať náhodnú výmenu výšky a hmotnosti jednotlivcov v rámci súboru údajov. Napríklad hodnota výšky pre jednotlivca #1 môže byť zamenená s hodnotou hmotnosti pre jednotlivca #5.

Príklad techniky výmeny údajov:

Pôvodný súbor údajov:

Tabuľka 10 - Príklad dátového súboru pred výmenou údajov

ID	meno	vek	výška (cm)	Hmotnosť (kg)
1	John	30	180	100
2	Jane	25	165	60
3	Jim	40	175	80
4	Jack	35	185	85
5	Jill	28	162	58

Množina údajov po neúspešnej výmene údajov:

Tabuľka 11 - Príklad dátového súboru po neúspešnej výmene údajov

ID	meno	vek	výška (cm)	hmotnosť (kg)
1	John	30	58	100
2	Jane	25	165	60
3	Jim	40	175	80
4	Jack	35	185	85
5	Jill	28	162	180

Aj keď táto technika môže zachovať súkromie výšky a hmotnosti jednotlivcov, prináša aj značné nepresnosti v údajoch. Napríklad výška a hmotnosť osoby sú zvyčajne korelované, takže vyšší jednotlivci majú tendenciu vážiť viac ako nižší jednotlivci. Náhodnou zámenou hodnôt výšky a hmotnosti môžeme skončiť s nezmyselnými kombináciami údajov, ktoré túto koreláciu neodrážajú. Mohli by sme napríklad skončiť so súborom údajov, kde jednotlivec, ktorý je zaznamenaný ako 58 cm vysoký, váži 100kg, čo pravdepodobne nebude dávať zmysel pre účely analýzy.

4.5.2 Metóda náhodnosti (randomizácia)

Technika randomizácie údajov je proces úpravy pôvodných údajov zmenou hodnôt prvkov údajov náhodným spôsobom a zahŕňa zmenu hodnôt v súbore údajov pridaním náhodného šumu. Dá sa to urobiť pridaním náhodného čísla ku každej hodnote alebo náhodným premiešaním poradia hodnôt. Táto technika sa zvyčajne používa na ochranu súkromia jednotlivcov a zabezpečenie toho, aby dôverné informácie neboli vystavené neoprávneným stranám.

Tento projekt je podporený z Európskeho sociálneho fondu.

Technika randomizácie údajov však nie je dobrou technikou pri transformácii údajov na publikovanie ako otvorené údaje pre verejnosť. Randomizované údaje totiž stále môžu odhaliť citlivé informácie o jednotlivcoch alebo skupinách. Okrem toho môže randomizácia zničiť dôležité vzorce a vzťahy v údajoch, čím sa stanú nepoužiteľnými na účely analýzy.

Zoberme si napríklad súbor údajov obsahujúci informácie o anamnéze pacientov. Predpokladajme, že pôvodný súbor údajov obsahuje premenné, ako sú vek, pohlavie a zdravotný stav. Randomizácia údajov zmenou hodnôt týchto premenných môže stále odhaliť citlivé informácie o jednotlivcoch, ako je ich vekové rozpätie, rozdelenie podľa pohlavia a prevalencia určitých zdravotných stavov v populácii napríklad v rámci užších špecifických skupín.

Navyše, randomizované údaje môžu sťažiť vykonanie zmysluplnej analýzy údajov. Napríklad, ak má výskumník záujem o skúmanie vzťahu medzi vekom a zdravotným stavom, randomizácia vekovej premennej by zničila vzťah a znemožnila by vyvodenie zmysluplných záverov.

Tu je príklad pôvodného súboru údajov a randomizovanej verzie tých istých údajov:

V tomto príklade bol do pôvodného súboru údajov pridaný náhodný šum bezvýznamnou úpravou veku a príjmu pacientov.

Pôvodný súbor údajov:

Tabuľka 12 - Príklad dátového súboru pred zanesením šumu

vek	príjem	vzdelanie
25	35 000	Bakalársky titul
40	60 000	Magisterský stupeň
32	45 000	Diplom zo strednej školy

Randomizovaný súbor údajov:

Tabuľka 13 - Príklad dátového súboru po zanesení šumu

vek	príjem	vzdelanie
26	35 247	Bakalársky titul
39	59 875	Magisterský stupeň
33	45 219	Diplom zo strednej školy

Tento projekt je podporený z Európskeho sociálneho fondu.

5 Správa, dohľad a riadenie procesov anonymizácie

V tejto kapitole sa snažíme definovať proces anonymizácie, ku ktorému nevyhnutne patrí aj riadenie tohto procesu. Na účely tohto dokumentu je potrebné chápať anonymizáciu ako súčasť väčšieho celku (správy otvorených údajov), pričom cieľom je z pohľadu riadenia popísať iba riadenie anonymizácie a nie definovať celú oblasť správy otvorených údajov.

Pod pojmom riadenie anonymizácie rozumieme také aktivity riadenia, správy a dohľadu, ktoré majú zabezpečiť štandardizované vykonávanie procesu anonymizácie jednotlivými OVM. V rámci riadenia pritom rozlišujeme minimálne dve úrovne riadenia, a to úroveň samotného OVM a vyššiu úroveň v podobe centrálnej úrovne riadenia, ktorá prekračuje hranice OVM. Štandardizácia tohto procesu je obzvlášť dôležitá aj v prípade výberu a implementácie centrálneho komplexného anonymizačného nástroja v rámci PaaS, ktorému sa venuje 8 kapitola.

Nižšie je rámcovo vymedzený súbor minimálnych požiadaviek na riadenie anonymizácie, ktorého implementácia smeruje k jednotnému postupu OVM pri výkone anonymizácie:

Požiadavky na úroveň OVM:

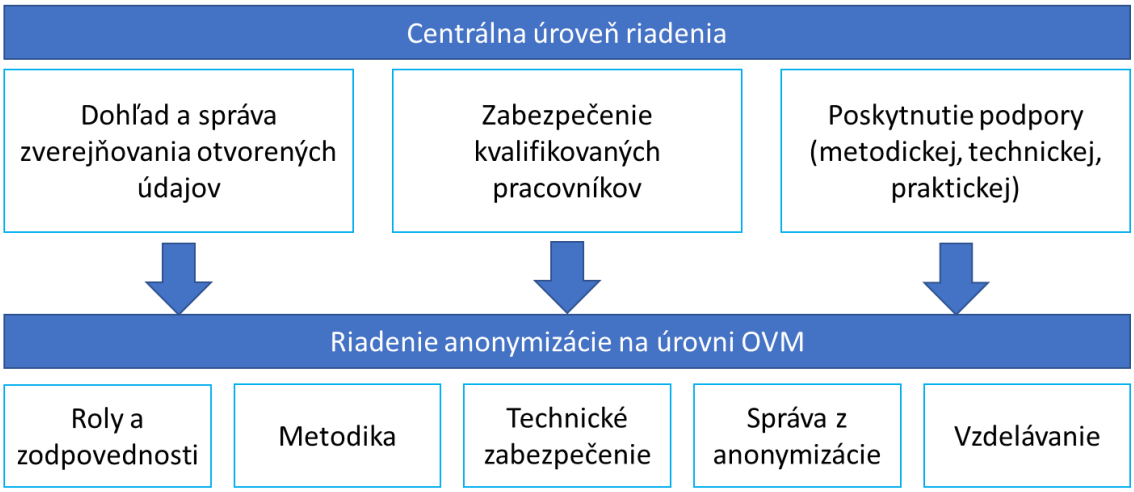
- Stanoviť roly a zodpovednosti minimálne v rozsahu
    - osoby zodpovednej (za výsledky anonymizácie)
    - osoby výkonnej (vykonávajúcej anonymizáciu)
- Stanoviť lokálnu metodiku / procesy
    - Určiť jasný postup (manuál) realizácie anonymizácie, kde základom sú centrálne stanovené štandardy, teda proces znázornený na obrázku 1, ktorý bude zahŕňať lokálne špecifiká pre dané OVM. Lokálne špecifiká môžu predstavovať rôzne právne úpravy vzťahujúce sa na údaje alebo špecifické typy údajov, s ktorými OVM narába. Je nutné si uvedomiť, že proces musí zodpovedať špecifikám každého OVM a byť súčasťou celkového procesu práce s otvorenými údajmi
- Rozhodnúť o technologickej podpore
    - začlenenie do technologického prostredia OVM, teda do business a aplikačnej architektúry a IT infraštruktúry, v rámci toho vykonať aj rozhodnutie o nástroji na anonymizáciu (lokálny, centrálny, zjednotiť požiadavky na funkcie)
- Definovať atribúty správy z vykonanej anonymizácie
    - Správa z procesu anonymizovania bude obsahovať minimálne tieto informácie: účel zverejnenia dát, analýzy citlivosti údajov a výsledky analýz jednotlivých atribútov dátového súboru, ktoré metódy boli využité na anonymizovanie jednotlivých atribútov a ich dopad na zníženie znovu identifikovania, schválenie zverejnenia všetkými potrebnými osobami zodpovednými za kontrolu anonymizovaného dátového súboru.

- Zabezpečiť kontinuálne vzdelávanie v oblasti anonymizácie
    - Dotknuté osoby v rámci OVM si musia pravidelne rozširovať svoje znalosti v oblasti anonymizácie a to najmä v oblasti nových hrozieb spätnej identifikácie a spôsobov, ako im predchádzať.

Požiadavky na centrálnu úroveň:

- Určiť spôsob dohľadu a správy zverejňovania otvorených údajov (vrátane ich anonymizácie), v rámci toho
    - Určiť orgán zodpovedný za dohľad (MIRRI alebo iná inštitúcia konajúca v úlohe Centrum excelentnosti - CoE)
    - Vydať príslušné štandardy, politiky a metodické usmernenia (napríklad na základe tohto výstupu), v rámci čoho sa špecifikuje proces riadenia a dohľadu, vizualizovaný na obrázku 1.
- Zabezpečiť pravidelnú osvetu a proaktívnu komunikáciu s OVM a vyškolenie zodpovedných a výkonných pracovníkov, vrátane prípravy školiacich materiálov.
- Zabezpečiť rozvoj anonymizačnej platformy prostredníctvom platformy PaaS v rámci projektu CIP.
- Zabezpečiť dostatočnú úroveň centrálnej podpory vrátane podpory technologickej (nástroj na anonymizáciu), metodickej a praktickej (FAQ, telefonická odborná podpora).

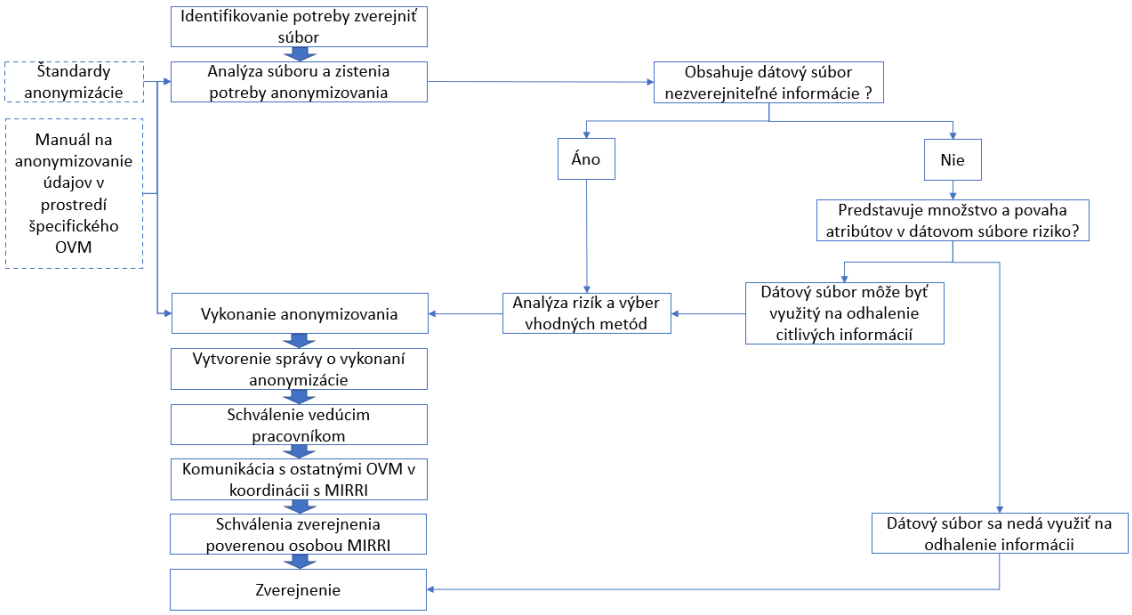
Obrázok 1 – Proces riadenia anonymizácie



Anonymizáciu je potrebné chápať ako časť procesu zverejňovania otvorených údajov, ktorá je realizovaná v prípade potreby zverejnenia súboru údajov s citlivými informáciami. Cieľom teda nie je popis celého procesu zverejňovania otvorených údajov ale iba popis procesu anonymizácie, ako súčasti procesu zverejňovania otvorených údajov.

Vstupom do procesu anonymizácie je potreba zverejnenia konkrétneho súboru údajov. Proces anonymizácie predstavuje tok aktivít (znázornených na obrázku 2 nižšie) smerujúcich k zverejneniu vstupného súboru údajov vo formáte otvorených údajov.

Obrázok 2 - Proces anonymizácie



Po identifikovaní potreby zverejniť dátový súbor nasleduje najdôležitejší krok celého procesu anonymizácie, ktorým je *analýza súboru a zistenia potreby anonymizovania.* V rámci tohto kroku dochádza k analýze jednotlivých atribútov dátového súboru a teda zodpovedanie otázky, či špecifický dátový súbor, ktorý chceme zverejniť obsahuje citlivé informácie podľa klasifikácie nezverejniteľných údajov (NZ). Za nezverejniteľné údaje podľa klasifikácie z výstupu 1.1.5 Štandardizácia anonymizácie údajov a výstupu 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov sú považované osobné údaje a ostatné citlivé údaje podľa osobitných zákonov.

Pre správne vykonanie analýzy dátového súboru je nutné identifikovať, či naň vplývajú ustanovenia osobitných zákonov, ktoré sa vzťahujú na niektoré atribúty dátového súboru. Tieto ustanovenia teda môžu upravovať, akým spôsobom je možné zverejňovať jednotlivé atribúty dátového súboru. Na obrázku 3, sa nachádza grafické znázornenie rámca prístupu k vykonávaniu analýzy dátového súboru z hľadiska potreby jeho anonymizácie.

Pokiaľ sa po prvkom kroku identifikovalo, že dátový súbor neobsahuje žiadne nezverejniteľné údaje je nutné ešte zvážiť tieto aspekty, predtým ako môže byť súbor zverejnený ako otvorené dáta, bez toho aby bol anonymizovaný:

— Nie je možné tieto dáta využiť na identifikáciu citlivých skutočností a osôb s využitím iných dát?

— Aké iné dáta sú verejne dostupné, ktoré je možné skombinovať s predmetným dátovým súborom?

— Ako a za akým cieľom môže byť predmetný dátový súbor spojený s inými verejne dostupnými informáciami?

Tento projekt je podporený z Európskeho sociálneho fondu.

— Je potrebné v rámci cieľa, s ktorým je dátový súbor zverejňovaný sprístupniť všetky atribúty vo forme otvorených údajov?

Ak po zvážení aspektov nedôjde k odhaleniu skutočnosti, že predmetný dátový súbor môže byť využitý na identifikáciu citlivých skutočností a osôb, je možné pristúpiť k jeho zverejneniu ako otvorených dát. Vyhodnoteniu, že dátový súbor je bezpečný na zverejnenie predchádza dôsledná analýza, v rámci ktorej sa musia dôkladne zvážiť vyššie spomenuté aspekty.

V prípade, že v rámci ďalšej analýzy sa preukáže, že je nutné anonymizovať niektoré údaje, je potrebné postupovať v anonymizovaní tak, ako sa postupuje pri dátovom súbore, v rámci ktorého boli hneď na začiatku odhalené nezverejniteľné údaje. Pokračuje sa teda analýzou rizík a identifikáciou vhodných anonymizačných metód na ich minimalizáciu opísaných v 6 kapitole.

Pred uvedením procesu anonymizácie je dôležité zmieniť, že každý prípad anonymizovania sa môže čiastočne líšiť od predchádzajúceho v dôsledku množstva faktorov. Preto nie je možné navrhnúť univerzálny podrobný proces, na základe ktorého je možné pristúpiť k anonymizovaniu v každej situácii. Preferovaným riešením je teda rámec, ktorý zodpovednej osobe bude slúžiť ako podklad pre analyzovanie dátového súboru a vykonanie činností s cieľom zabezpečenia citlivých informácií, pred zverejnením súboru ako otvorených údajov. Zároveň pri práci s rámcom anonymizovania je dôležité pamätať, že k zverejňovaným dátovým súborom je vždy nutné pristupovať ako by sme pracovali s maximálnym rizikom, ktoré zverejňovanie anonymizovaných otvorených dát predstavuje.

Proces anonymizácie v prostredí ISVS

Proces anonymizácie v prostredí informačných systémov verejnej správy na Slovensku je výrazne limitovaný jeho súčasným stavom, bližšie opísaným v 8 kapitole. Z dôvodu nevyspelých informačných systémov na účely anonymizovania je rozsah metód, ktoré môžu byť použité na anonymizovanie údajov zúžený, keďže pri niektorých metódach pri súčasnom stave ISVS by nebolo možné dosiahnuť dostatočnú ochranu citlivých údajov.

Tento projekt je podporený z Európskeho sociálneho fondu.

6 Výber vhodných metód pre jednotlivé prípady použitia

Každý prípad použitia môže byť čiastočne odlišný, a preto je dôležité, aby pre každý jednotlivý prípad boli vyberané metódy anonymizovania samostatne. Tento prístup je nutné zvoliť nielen z dôvodu, že aj rovnaký dátový súbor sa môže v budúcnosti jemne líšiť obsahom atribútov, ale aj z dôvodu, že anonymizovanie ako také sa neustále vyvíja a rôzne metódy a algoritmy, ktoré sa využívajú v súčasnosti už nemusia byť považované za efektívne v budúcnosti.

V súčasnosti sa odporúča vyberať metódy na anonymizovanie dátových súborov s cieľom ich publikovania ako otvorených údajov tak, ako sú zatriedené nižšie na základe kategórie citlivosti údajov (viď. kapitola 2.2). Zároveň pri výbere jednotlivých metód je nutné zohľadniť aj všetky ostatné riziká spojené so spätnou identifikáciou citlivých údajov, bližšie sa problematike spätnej identifikácie venuje výstup 1.1.5 Štandardizácia anonymizácie údajov.

Metódy anonymizácie zatriedené podľa kategórií citlivosti:

Vyhradené údaje

- Metóda potlačenia údajov (odstránenie údajov)

Dôverné

- Metóda potlačenia údajov (odstránenie údajov)
- Generalizácia
- Maskovania
- Globálne prekódovania

Interné

- Pri interných údajoch je nutné zvážiť, aká metóda anonymizovania bude využitá podľa povahy a typu konkrétneho údaju a je na vlastníkovi dát, ktorú z metód využije. Keďže ide o anonymizovanie za cieľom publikovania ako otvorených dát odporúča sa vyberať z vhodných metód popísaných pre ostatné dve kategórie spomenuté vyššie.

Okrem spomínaných metód pre rôzne kategórie rizík existuje ešte možnosť využitia agregovania celého dátového súboru, kedy nebudú publikované dáta zverejnené v granularite zdrojového súboru. Táto metóda pri jej správnej aplikácii prináša z hľadiska ochrany citlivých informácií najväčšiu bezpečnosť, avšak za cenu možnej straty využiteľnosti dát. Príkladom môžu byť zdravotné dáta zverejnené na portáli data.europe.eu (viď. Obrázok 3). Na tomto príklade môžeme vidieť, že táto metóda zabezpečuje vysokú bezpečnosť citlivých údajov, pričom použiteľnosť dát na ďalšie využitie samozrejme utrpela.

Tento projekt je podporený z Európskeho sociálneho fondu.

Obrázok 3 - Agregované zdravotné dáta[18]

lfd. Nr.	Pos.-Nr. der ICD-10/Hauptdiagnose	Ins-gesamt	davon im Alter von ...									
			0 - 1	1 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30	30 - 35	35 - 40	40 - 45	45 - 50
		Anzahl											
	Insgesamt	368 437	22 009	6 999	4 085	4 699	7 892	12 279	16 850	19 807	16 980	18 177	20 456
1	A00-T98 Krankheiten, Verletzungen und Vergiftungen	348 707	8 376	6 937	4 066	4 682	7 813	12 058	16 213	18 637	15 952	17 618	20 079
2	A00-B99 Bestimmte infektiöse u. parasitäre Krankheiten	11 740	721	1 290	434	399	365	371	407	421	353	443	466
3	A15-A19,B90 Tuberkulose einschl. ihrer Folgezustände	213	–	5	7	3	8	6	11	14	24	14	18
4	A39 Meningokokkeninfektion	6	–	–	1	–	1	–	–	2	–	–	–

Metódy anonymizovania využiteľné pri súčasnom stave ISVS sú obmedzené, avšak aj pred výberom konkrétnej, ktorá bude využitá v rámci špecifického typu údaju je nutné si spraviť analýzu, v ktorej sa zvážia tieto skutočnosti:

- Pravdepodobnosť, že sa niekto pokúsi spätne identifikovať citlivé údaje
- Pravdepodobnosť úspechu pri pokuse o spätnú identifikáciu
- Vhodnosť metódy anonymizácie pre daný účel publikovania údajov vo forme otvorených dát
- Ďalšia použiteľnosť údajov po anonymizovaní dátového súboru.

Po určení, na aké typy údajov sa využijú jednotlivé metódy anonymizovania sa pristúpi k samotnému vykonaniu anonymizovania citlivých údajov. V nasledujúcich kapitolách sú jednotlivé metódy aplikované na vzorky dát pre vybrané prípady použitia. Avšak tieto príklady by nemali byť považované za doslovný návod, ako vykonať anonymizovanie na danom type dát, keďže aj pri malých odlišnostiach v štruktúre, obsahu a možností kombinovateľnosti dátového súboru s inými údajmi už nemusí byť takto vykonaná anonymizácia dostatočne bezpečná.

6.1 Anonymizovanie štruktúrovaných údajov

V tejto podkapitole na príklade údajov v zdravotníctve a geo-priestorových dát demonštrujeme použitie rozhodovacieho stromu anonymizácie a približujeme dobrú prax zo zahraničia v tejto oblasti.

6.1.1 Anonymizovanie citlivých údajov v zdravotníctve

Zverejňovanie zdravotných dát na Slovensku pribieha v rámci striktných pravidiel vyplývajúcich zo Zákona č. 153/2013 Z. z. o národnom zdravotníckom informačnom systéme a o zmene a doplnení niektorých zákonov. Za zverejňovanie zdravotných údajov je zodpovedné Národné centrum zdravotníckych informácií (NCZI). V súčasnosti všetky informácie, ktoré sú nimi zverejňované ako otvorené dáta buď neobsahujú žiadne citlivé informácie a teda sú zverejňované mimo režimu anonymizovania, alebo sú anonymizované agregovaním. V rámci agregácii NCZI dosahuje vysokú úroveň bezpečnosti, keďže v rámci dát je venovaná pozornosť aj tomu, aby v každej skupine bolo vždy dostatočné množstvo agregovaných záznamov.

Počas pandémie Covid-19 (v rokoch 2020 až 2022) NCZI zverejňovalo pod špeciálnym režimom aj dátové súbory, v rámci ktorých boli anonymizované identifikátory

[18] https://data.europa.eu/data/datasets/stanord_cms-55059?locale=en

Tento projekt je podporený z Európskeho sociálneho fondu.

konkrétnych osôb, vytvorením takzvaného pseudonymizačného tajomstva. Avšak tomuto typu zverejňovania údajov sa NCZI typicky kvôli bezpečnosti vyhýba. Z hľadiska ochrany citlivých informácii je to správny postup, pokiaľ sa zverejňovanie takýchto informácii nevyžaduje a zároveň nie je zavedená dostatočne rozvinutá politika na ochranu citlivých údajov, ktorá by detailne adresovala túto problematiku.

NCZI má metodiku, ktorá definuje ako je nutné pristupovať k zverejňovaniu anonymizovaných údajov v špeciálnom režime a zároveň aj to, ako treba správne pristupovať k vytváraniu agregovaných štatistických záznamov. Ako východisko pre metodické usmernenie na agregovanie údajov sa využíva aj zákon č. 540/2001 Z. z. o štátnej štatistike v znení neskorších predpisov.

Avšak pokiaľ nevznikne potreba zverejňovať anonymizované informácie v inej forme ako v agregovanej mimo špeciálneho režimu, takéto riešenie anonymizovania v rámci NCZI nie je nutné zavádzať. Zavedenie procesov anonymizovania pomocou iných metód by si zrejme vyžadovalo výraznejšie úpravy interného informačného systému, ktorý bol vybudovaný na mieru a teda by bolo do neho potrebné zaviesť napríklad funkciu mapovania dátových súborov medzi sebou a mnohé iné, ktoré by dopomohli ochrane citlivých údajov.

Rámec pre anonymizáciu zdravotných dát Európskej medicínskej agentúry

Jedným z príkladov dobrej praxe v rámci Európskej únie je Regulácia 0070: Usmernenie k Anonymizácii[19], tento dokument sa venuje problematike správnej anonymizácie dát v rámci klinických štúdií pred ich zverejnením a poskytuje podrobný prístup k tomu, aké rôzne možnosti a obmedzenia v oblasti anonymizácie zdravotných dát existujú.

Vo všeobecnosti regulácia považuje tieto metódy za vhodné na použitie pri anonymizovaní zdravotných dát z klinických štúdií:

- Potlačenie
- Generalizácia
- Randomizácia

Pri výbere akejkoľvek z týchto metód je vždy dôležité prihliadať na riziko spojené s využitím danej metódy na špecifický atribút a následnej sekundárnej využiteľnosti dát po aplikovaní vybranej metódy.

Európska medicínska agentúra v rámci regulácie ako príklad dobrej praxe uvádza, že k anonymizácii, by malo byť pristúpené v nasledovných krokoch:

1. **Určenie priamych identifikátorov –** Priame identifikátory sú často neužitočné na analýzu údajov a sú nimi napríklad: mená, e-mailové adresy, telefónne čísla a číslo poistenca.

2. **Určenie nepriamych identifikátorov –** Nepriame identifikátory sú napríklad: dátum narodenia, úmrtia alebo návštevy kliniky, PSČ, pohlavia a etnická

[19] https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-3.pdf

Tento projekt je podporený z Európskeho sociálneho fondu.

príslušnosť. Tieto identifikátory môžu nepriamo identifikovať jednotlivcov a preto je dôležité ich tiež zahrnúť do procesu anonymizácie

Určenie, či ide o priamy alebo nepriamy identifikátor je možné zistiť zodpovedaním týchto otázok:

— **Replikovateľnosť –** Hodnoty premenných musia byť dostatočne stabilné v čase, aby sa vyskytovali konzistentne vo vzťahu k dotknutej osobe

— **Rozlíšiteľnosť –** Hodnota musí mať dostatočnú variabilitu na rozlíšenie medzi jednotlivcami v dátovom súbore

— **Poznateľnosť –** Protivník musí poznať tento identifikátor dotknutej osoby, aby ich mohol využiť na opätovné identifikovanie

3. **Identifikácia možných útočníkov a pravdepodobných útokov na údaje –** Pri zverejnení otvorených údajov majú útočníci s najväčšou pravdepodobnosťou záujem o dosiahnutie týchto cieľov a teda pri údajoch o klinických testoch je potrebné sa zameriavať na zabránenie ich dosiahnutia:

— Útočník ma záujem zistiť, kto sú účastníci klinického testovania za cieľom finančného obohatenia sa

— Určitý účastník klinických testov je predmetom osobitného verejného záujmu a tlač alebo iná záujmová skupina chce získať tieto informácie.

— Skupina útočníkov alebo útočník z rôznych dôvodov má za cieľ podkopať verejnú podporu zverejňovania otvorených údajov odhalením akejkoľvek spojitosti na konkrétnu osobu.

— Jednotlivec, ktorý náhodou skúma správu obsahujúcu údaje o klinických testoch a má dobrú znalosť o pozadí určitého účastníka, a teda vie presne odhadnúť určité informácie a identifikovať daného účastníka.

4. **Analyzovanie použiteľnosti údajov –** Výsledky anonymizácie údajov klinických testov, ktoré sú anonymizované do takej miery, že už nie sú použiteľné na zamýšľané sekundárne účely, sú nesprávne anonymizované. Preto by malo byť skreslenie dát minimalizované. Dôležité je dosiahnuť rovnováhu, v rámci ktorej bude existovať prijateľne nízka pravdepodobnosť opätovnej identifikácie a vysoká použiteľnosť údajov na sekundárne účely.

5. **Určenie prahu rizika opätovnej identifikácie pre daný dátový súbor a vyhodnotenie skutočného rizika opätovnej identifikácie –** Meranie rizika opätovnej identifikácie zahŕňa výber vhodného spôsobu na meranie rizika, vhodnej prahovej hodnoty a následne vykonanie meranie rizika v špecifickom dátovom súbore o klinických testoch, ktorý sa má zverejniť. Pre publikovanie otvorených dát sa odporúča využiť maximálnu hodnotu rizika v rámci prahovej hodnoty.

6. **Metodológia anonymizácie –** Poverená osoba je zodpovedná za vytvorenie primeraného procesu anonymizácie a teda použitia vhodných anonymizačných metód. V rámci procesu je nevyhnutné opísať, ako sa zníži riziko opätovnej

Tento projekt je podporený z Európskeho sociálneho fondu.

identifikácie. Okrem toho je nutné poskytnúť aj odôvodnenie, prečo boli anonymizované vybrané údaje a výber použitých anonymizačných metód.

7. **Zdokumentovanie metodológie a procesu anonymizácie –** Dokumentácia použitého postupu anonymizovania je dôležitým krokom, pretože poskytuje informácie nielen o metódach, ktoré sa použili na anonymizáciu údajov, ale aj dôvody ich použitia.

Uvedený postup považujeme za vhodným rámec, aplikovateľný aj pre nastavenie procesov anonymizácie v iných oblastiach, nakoľko je prehľadne popísaný, zjavne overený praxou a aplikovaný na vysoko citlivé zdravotnícke dáta. Treba si však uvedomiť, že je určený pre anonymizovanie dát z klinických testov, a teda je dôležité zdôrazniť, že v prípade iného typu dát tento rámec bude musieť byť primerane prispôsobený. Už len v oblasti zdravotníckych dát je nutné zvažovať platnú legislatívnu úpravu, jedným z príkladov je aj Zákon o národnom zdravotníckom informačnom systéme (Zákon 153/2013 Z. z.). Podobné platí pre ostatné oblasti primerane.

Zverejňovanie anonymizovaných zdravotných údajov v EÚ na data.europa.eu

V rámci európskeho portálu otvorených dát data.europa.eu je možné nájsť množstvo dátových súborov s údajmi zo zdravotníctva, v prevažnej väčšine pokiaľ zdrojový súbor obsahuje citlivé informácie a je ho teda nevyhnutné anonymizovať, je na tento účel využité agregovanie. Na tomto portáli je však možné nájsť aj dáta, kde sa v rámci zdrojového súboru nachádzali citlivé informácie a bola pred ich publikovaním zvolená iná metóda anonymizovania.

Ako príklad využitia iných metód anonymizovania môže byť použitý dátový súbor[20] s granulárnou štruktúrou záznamov, publikovaný na data.europe.eu obsahujúci údaje o príjme psychiatrických pacientov do nemocnice.

V rámci tohto špecifického súboru môžeme vidieť, že na jeho anonymizovanie bola prevažne využitá metóda potlačenia, generalizovania a globálneho prekódovania.

Tabuľka 14 - Potlačenie identifikátora pacienta

Kód označenia	Štatistické označenie	Jednotka
HRA09C1	Číslo dospelého alebo dieťaťa prijatého na hospitalizáciu	Číslo
HRA09C1	Číslo dospelého alebo dieťaťa prijatého na hospitalizáciu	Číslo
HRA09C1	Číslo dospelého alebo dieťaťa prijatého na hospitalizáciu	Číslo

V tabuľke 14 môžeme vidieť, že došlo k potlačeniu špecifických identifikátorov, akými sú číslo pacienta a iných identifikačných čísel, ktoré by mohlo viesť k identifikácii jednotlivca.

[20] https://data.europa.eu/data/datasets/e2bd8b40-8b9d-49ba-8b2e-02068eb9ef4a?locale=en

Tabuľka 15 - Generalizácia dátumu hospitalizácie

Kód označenia	TLIST(A 1)	Rok
HRA09C1	2006	2006
HRA09C1	2006	2006
HRA09C1	2006	2006

V tabuľke 15 sa nachádza príklad využitia metódy generalizácie v predmetnom dátovom súbore, kde dátum hospitalizácie pacienta bol generalizovaný až na úroveň roku.

Pri úvahách o anonymizácii citlivých údajov v zdravotníctve je možné rátať s rôznymi spôsobmi a ich kombináciami na zabezpečenie ochrany citlivých údajov. Je však nutné mať na pamäti, že výber konkrétnej metódy je veľmi úzko viazaný na špecifické prípady použitia, ako potvrdzujú aj mnohé štúdie[21], ktoré sa venujú špecificky anonymizovaniu citlivých údajov v zdravotníckych dátach. To, že nie je možné odporučiť generickú metódu vychádza aj z faktu, že pred zverejňovaním akýchkoľvek otvorených dát je vždy nutné sa pozerať na tieto údaje z hľadiska ich možných prepojení s inými údajmi a podľa zistených špecifík prijímať potrebné opatrenia, ktoré zamedzia úniku citlivých údajov. Vo všeobecnosti najbezpečnejšími metódami je potlačenie všetkých typov údajov, ktoré môžu viesť akýmkoľvek spôsobom k spätnej identifikácii alebo agregáciou dátového súboru ako celku.

Využitie rozhodovacieho stromu anonymizácie

V rámci modelovej situácie využijeme pre tento príklad umelo vytvorenú štruktúru modelového dátového súboru zdravotníckych dát, na ktorej bude demonštrované využitie rozhodovacieho stromu.

Tabuľka 16 - Štruktúra modelového dátového súboru zdravotníckych dát

Atribút	Potreba anonymizovať
ID pacienta	Áno
Vek	Áno
Pohlavie	Áno
Výška	Áno
Telesná váha	Áno
BMI	V závislosti od cieľu využitia dátového súboru
Krvný tlak	V závislosti od cieľu využitia dátového súboru
Cholesterol	V závislosti od cieľu využitia dátového súboru
Množstvo cukru v krvi	V závislosti od cieľu využitia dátového súboru
Dátum návštevy doktora	áno
Poistenecké číslo pacienta	áno

Prezentovaný modelový dátový súbor v tabuľke 14, jednoznačne obsahuje atribúty, ktoré sú považované za citlivé a je nutné ich anonymizovať a pre ich určenie bol ako vodítko využitý rozhodovací strom. Tento dátový súbor obsahuje aj nepriame identifikátory, ktorých kombináciou môže dôjsť k odhaleniu citlivých informácii a teda po určení, na aký cieľ bude dátový súbor využitý zvyšné „nevýznamné atribúty“ pre daný

[21] https://www.researchgate.net/publication/359154769_A_Review_of_Anonymization_for_Healthcare_Data

Tento projekt je podporený z Európskeho sociálneho fondu.

cieľ by mali byť z dátového súboru odstránené úplne, čo výrazne zníži riziko spätnej identifikácie.

V rámci tohto príkladu by jednoznačne z tohto dátového súboru mali byť odstránené (potlačené) tieto atribúty: ID pacienta, Poistenecké číslo pacienta. Tieto atribúty nepridajú žiadnu hodnotu pri analýze a zároveň ani nesmú byť ponechané v dátovom súbore. Ostatné atribúty je možné anonymizovať aj inými technikami ako potlačením, avšak pri dodržaní dostatočnej úrovne anonymizácie, ktorá zaručí, že nedôjde k spätnej identifikácii. Okrem iného využitie iných techník ako agregácie údajov a potlačenia údajov v súčasnosti, kedy je komplikované zabezpečiť kontrolu možného prepojenia dátových súborov sa neodporúča, v dôsledku značného rizika úniky citlivých údajov.

6.1.2 Anonymizovanie geo-priestorových dát

Na Slovensku je situácia v rámci zverejňovania geo-priestorových dát ako otvorených dát pomerne živou témou v rámci iniciatív, súvisiacich s európskym nariadením INSPIRE. Zdrojové dáta zvyknú obsahovať aj citlivé údaje, ktoré je nutné anonymizovať pred ich zverejnením. V súčasnosti za anonymizovanie takýchto údajov zodpovedá poskytovateľ a proces anonymizovania jednotliví poskytovatelia nevykonávajú podľa jednotnej metodiky, ktorá by celý proces zjednotila.

Ochrana citlivých geo-priestorových údajov

Na úrovni Európskej únie je možné za najrozsiahlejšiu iniciatívu, ktorá sa venuje ochrane citlivých priestorových údajov považovať iniciatívu ELISE. Jej príručka *Report: Guidelines for public administrations on location privacy*[22] poskytuje usmernenia pre orgány verejnej správy o súkromí geo-priestorových údajov. Zaoberá sa spôsobmi, akými je možné pristúpiť k zachovaniu bezpečnosti takýchto údajov. V dokumente je možné nájsť rôzne modelové príklady, kedy poskytovateľ údajov narába s citlivými údajmi. Na týchto príkladoch je vysvetlené, ako v jednotlivých situáciách treba zaobchádzať s údajmi a ktoré legislatívne úpravy sa dotýkajú vybranej modelovej situácie.

Ďalším príkladom anonymizácie geo-priestorových a iných citlivých údajov, ktoré sú obsiahnuté v rámci takéhoto typu dátového súboru sa venuje štúdia[23] realizovaná vo Fínsku. Pri geo-priestorových dátach za všeobecne platné metódy anonymizovania je považovaná:

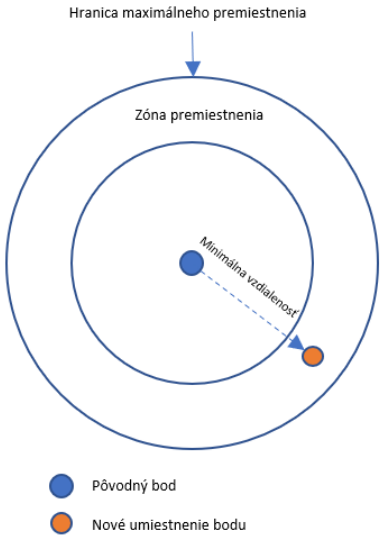
— Generalizácia – začlenenie individuálnych bodov do väčších celkov. Údaje o individuálnych bodoch sú začlenené do jedného bodu, ktorý zhŕňa body v rámci väčšej oblasti.

— Randomizácia – najčastejšie využívanou technikou v rámci tejto metódy je geo-maskovanie, ktoré zanáša do údajov šum a dislokuje pôvodnú polohu o určitú vzdialenosť v rámci určitého rádiusu.

[22] https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/document/report-guidelines-public-administrations-location-privacy-version-2
[23] https://www.sciencedirect.com/science/article/pii/S0198971520302465

Tento projekt je podporený z Európskeho sociálneho fondu.

Obrázok 4 - Geo-maskovanie (Randomizácia)



Pri iných typoch citlivých údajoch v rámci geo-priestorového dátového súboru predmetná štúdia odporúča využiť metódy potlačenia pri osobných údajoch, z hľadiska navrhnutého rozhodovacieho stromu v tomto dokumente by teda všetky údaje spadajúce do kategórie vyhradené mali byť potlačené. Následne údaje spadajúce do kategórie dôverné štúdia odporúča generalizovať, aby sa zachovala, čo najväčšia použiteľnosť dátového súboru.

Vo všeobecnosti aj pri geo-priestorových dátach sa teda odporúča zvoliť metódy anonymizácie na základe kategorizácie rizík podľa kapitoly 6. Avšak je dôležité poznamenať, že aj pri geo-priestorových dátach je nutné vždy pristupovať ku každému prípadu osobitne a analyzovať vhodnosť použitia rôznych metód anonymizovania v konkrétnom kontexte.

Využitie rozhodovacieho stromu anonymizácie

V rámci modelovej situácie využijeme túto umelo vytvorenú štruktúru dátového súboru geo-priestorových údajov, na ktorej bude demonštrované využitie rozhodovacieho stromu. V rámci modelového dátového súboru (tabuľka 15) bude uvažované, že v rámci granularity poskytuje dáta až na úroveň ulice.

Tabuľka 17 - Štruktúra modelového dátového súboru geo-priestorových dát

Atribút	Potreba anonymizovať
ID umiestnenia	Áno
Zemepisná šírka	Áno
Zemepisná dĺžka	Áno
Typ využitia územia	V závislosti od cieľu využitia dátového súboru
Počet obyvateľov	Nie
Priemerný príjem	Áno
Dátum aktualizácie údajov	Nie

Keďže modelový súbor má vysokú granularitu je nutné v prvom rade nutné zvážiť, či nie je nutné agregovať tieto dáta na vyššiu úroveň, keďže pri niektorých uliciach je možné, že nebude možné dosiahnuť dostatočne veľkú vzorku a teda nebude možné dosiahnuť

Tento projekt je podporený z Európskeho sociálneho fondu.

dostatočnú anonymitu. Pre tento príklad však predpokladajme, že každý záznam ulice obsahuje dostatočne veľkú vzorku a nedôjde k odhaleniu citlivých informácií.

Pri využití navrhnutého procesu anonymizácie ako rámca pre navigovanie rozhodnutí, či je nutné podstúpiť anonymizáciu je jednoznačne zistené, že tento dátový súbor obsahuje aspoň jeden atribút, ktorý by mal byť s určitosťou potlačený. ID umiestnenia musí byť jednoznačne odstránené z dátového súboru nakoľko môže slúžiť na prepájanie dátových súborov. Na ostatné údaje je možné použiť ďalšie techniky anonymizovania, akými je napríklad geo-maskovanie, ktoré zaručí dostatočnú anonymitu aj pre skupiny osôb, kde odhalením napríklad ich priemerných príjmov v určitej lokalite môže vzniknúť riziko ďalších útokov na odhalenia citlivých informácií. Zároveň zverejnenie takejto informácie samotnej sa dá považovať za citlivú informáciu, pokiaľ nebola anonymizovaná napríklad geo-maskovaním.

Špecifickým prípadom je typ využitia územia, ktorý by mal byť z dátového súboru odstránený úplne, keďže môže pri kombinácii s inými údajmi slúžiť k odhaleniu rôznych utajovaných skutočností. Zároveň sa v tomto prípade prízvukuje, že pri dátových súboroch tohto typu je dôležité myslieť aj na takýto typ útoku na spätnú identifikáciu a podľa toho aj pristupovať k realizácii anonymizácie.

6.2 Anonymizovanie neštruktúrovaných údajov

Dôležitosť anonymizácie neštruktúrovaných dát je zvýšená tým, že ich počet je výrazne väčší. O to dôležitejšie je zaviesť jednotný prístup k anonymizovaniu citlivých údajov, ktoré sa v neštruktúrovaných dátach nachádzajú.

Jedným z typov neštruktúrovaných dát sú zmluvy, v ktorých sa zvyčajne nachádza množstvo citlivých informácií, ktoré je potrebné anonymizovať. Aktuálne môžeme nájsť v rámci centrálneho registru zmlúv (www.crz.gov.sk, ďalej CRZ) viac ako 3 milióny zmlúv, v rámci ktorých sa nachádza množstvo citlivých informácií. Zverejňovať zmluvy v CRZ má povinnosť množstvo subjektov verejnej správy. Postupy, akými vykonávajú anonymizovanie prostredníctvom metódy potlačenia sa líšia subjekt od subjektu. Už rýchla sonda odhalí, že v mnohých prípadoch bola anonymizácia vykonaná neodborne, nedostatočne, a teda je možné z týchto zmlúv jednoducho odhaliť citlivé informácie, akými sú napríklad: rodné číslo, adresa trvalého pobytu, číslo občianskeho preukazu, číslo účtu a mnohé iné citlivé údaje. V zriedkavých prípadoch je anonymizovanie úplne opomenuté, v iných prípadoch zas nie sú anonymizované všetky údaje, ktoré podľa zákona o ochrane osobných údajov nesmú byť zverejnené.

Jedným z príkladov nesprávne vykonaného anonymizovania je „začiernenie“ všetkých alebo časti citlivých údajov v mylnej viere, že táto metóda potlačenia údajov je dostatočná. Prípad od prípadu bolo však toto začiernenie realizované digitálne v rámci inej údajovej vrstvy dokumentu než v tej, v ktorej sa nachádza samotný text zmluvy a teda je možné citlivé informácie z tejto zmluvy veľmi jednoducho skopírovať. Na druhej strane je treba uviesť, že v iných prípadoch anonymizovania zmlúv v CRZ sa používa fyzické prelepovanie citlivých informácii v dokumentoch pred naskenovaním, prípadne manuálne vyhľadanie citlivých informácii v dokumentoch a ich následné mazanie.

Je teda zrejmé, že jednotlivé subjekty verejnej správy nemajú dostatočné povedomie o tom, aké informácie by mali byť anonymizované a zároveň, aké možnosti existujú na vykonanie anonymizovania a ktoré z metód sú pre danú situáciu vhodné, čo potom vedie k zlyhaniu ľudského faktoru a úniku citlivých informácií.

Pre zníženia rizika úniku citlivých informácií z neštruktúrovaných údajov, akými sú spomínané zmluvy je nutné zlepšiť a zjednotiť proces anonymizácie, najmä:

— Zaviesť jasnú internú politiku resp. metodiku vykonávania anonymizácie neštruktúrovaných údajov, pre každú z dotknutých inštitúcií.

— Vyškoliť zamestnancov, ktorí sú zodpovední za anonymizovanie a ktorí ho vykonávajú.

— Zaviesť nástroje, ktoré dokážu overiť správnosť anonymizovania neštruktúrovaných citlivých údajov. Kapitola 8.1 pojednáva bližšie o výbere nástrojov na anonymizovanie neštruktúrovaných údajov.

Tento projekt je podporený z Európskeho sociálneho fondu.

7 Definícia algoritmov pre vybrané metódy

7.1 Algoritmy k-anonymity

K-Anonymita je koncept ochrany osobných údajov, ktorý sa používa na ochranu identity jednotlivcov v súbore údajov. Ide o metódu, ktorej cieľom je zabezpečiť, aby jednotlivca nebolo možné opätovne identifikovať zo súboru údajov jeho kombináciou s inými zdrojmi údajov. K-anonymizácia sa často označuje ako sila „skrytia sa v dave“. Údaje o jednotlivcoch sa zhromažďujú vo väčšej skupine, čo znamená, že informácie v skupine môžu zodpovedať akémukoľvek jednotlivému členovi, čím sa zakrýva identita jednotlivca alebo dotknutých jednotlivcov.

Súbor údajov je k-anonymný, keď atribúty v ňom sú generalizované alebo potlačené, až kým každý riadok nie je identický s aspoň k-1 ďalšími riadkami. Čím vyššia je teda hodnota k, tým nižšie je riziko opätovnej identifikácie. Podobne ako s veľkosťou davu klesá pravdepodobnosť, že nájdete presne osobu, ktorú hľadáte, k-anonymizácia funguje obzvlášť dobre pri veľkých súboroch údajov.

Ak však nie je dostatok údajov na anonymizáciu nepriamych identifikátorov, možno bude potrebné zredigovať riadky údajov, čím sa stanú nepoužiteľnými. Výskumníci preto kráčajú po tenkej hranici medzi užitočnosťou a súkromím, ktoré je však nevyhnutné z právneho a etického hľadiska. Bežne používané pravidlo je nastaviť **k aspoň na 5**, pretože sa ukázalo, že táto úroveň k-anonymity poskytuje primeranú úroveň ochrany súkromia pri zachovaní užitočného množstva dát. Pre vysoko citlivé údaje však môže byť potrebná vyššia úroveň k-anonymity, napríklad 10 alebo dokonca 100, aby sa zabezpečila ochrana súkromia.

Na dosiahnutie K-anonymity sa údaje zvyčajne upravujú potlačením určitých polí alebo generalizáciou určitých atribútov v súbore údajov. Napríklad vek možno generalizovať na vekové skupiny alebo PSČ možno nahradiť širšími geografickými oblasťami. Údaje sa tak môžu stále používať na analýzu a je aj zabezpečená ochrana súkromia jednotlivcov.

7.1.1 Praktický príklad aplikácie K-anonymity na súbor údajov

V tejto podkapitole je prezentovaný jednoduchý príklad toho, ako možno použiť K-anonymitu na dátovom súbore so zdravotnými údajmi. Aby sme ilustrovali, ako k-anonymizácia funguje v praxi, využijeme fiktívny súbor údajov (úplné pole obsahuje 1000 riadkov a 7 stĺpcov), ktorý je synteticky generovaný v Python prostredníctvom Pandas DataFrame.

Nasledujúce úložisko Github[24] obsahuje kód a výsledok anonymizovaných údajov tohto prípadu použitia.

Súbor so skutočnými údajmi nie je použitý, pretože pokiaľ sú takéto údaje zverejnené ako otvorené dáta, tak sú už anonymizované. Zároveň využitie reálnych zdrojových neanonymizovaných údajov by v rámci príkladu nemohlo byť použité, keďže by došlo k odhalenie citlivých informácií. Okrem toho sa tento súbor údajov generuje s riadenou

[24] https://github.com/olesyagrabova/anonymization-example-/blob/c8391057c71bfe9a3d8a950a79c9777d1f7fcc81/k-anonymity%20techniques.py

Tento projekt je podporený z Európskeho sociálneho fondu.

distribúciou, aby sa zabezpečilo, že syntetické údaje budú mať špecifické vlastnosti a budú sa riadiť vopred určeným vzorcom distribúcie. Definovaním distribúcie každej premennej môžeme simulovať scenáre reálneho sveta a vytvoriť súbor údajov, ktorý sa veľmi podobá skutočným údajom a navyše tento syntetický súbor údajov môže byť reprodukovaný inými, čím sa zabezpečí konzistentnosť a opakovateľnosť experimentov alebo analýz.

Štruktúra súboru údajov obsahuje 7 premenných:

1) vek
2) pohlavie
3) krvná skupina
4) fajčiar
5) výška
6) hmotnosť
7) telefónne číslo

Obrázok 5 - generovanie dátového súboru

```
import pandas as pd
import numpy as np

np.random.seed(42)

# Generate synthetic data with controlled distributions (1,000 rows)
age = np.random.choice([25, 30, 35, 40, 45, 50, 55, 60, 65, 70], size=1000, p=[0.1, 0.1, 0.1, 0.1, 0.1, 0.1,
sex = np.random.choice(['M', 'F'], size=1000, p=[0.5, 0.5])
blood_type = np.random.choice(['A', 'B', 'AB', 'O'], size=1000, p=[0.4, 0.3, 0.2, 0.1])
smoker = np.random.choice([0, 1], size=1000, p=[0.7, 0.3])
height = np.random.choice([150, 160, 170, 180, 190], size=1000, p=[0.2, 0.3, 0.3, 0.1, 0.1])
weight = np.random.choice([60, 70, 80, 90, 100], size=1000, p=[0.2, 0.3, 0.3, 0.1, 0.1])
phone_numbers = ['555-555-' + '{0:04d}'.format(x) for x in range(1000)]

# Create a Pandas dataframe
data = pd.DataFrame({'Age': age, 'Sex': sex, 'Blood_Type': blood_type, 'Smoker': smoker, 'Height': height, 'W
print(data)
```

Začíname vykonaním týchto krokov:

1. Importujte Pandas a knižnice Numpy: Pandas slúži na manipuláciu a analýzu údajov a Numpy na numerické operácie.

2. Nastavte náhodné číslo generovania na 42: tým sa zabezpečí, že generované náhodné hodnoty budú reprodukovateľné.

3. Vygenerujte syntetické údaje: vytvárame syntetické údaje pre 1 000 riadkov, pričom riadime distribúciu premenných podľa veku, pohlavia, krvnej skupiny , fajčiara, výšky, hmotnosti a telefónneho čísla.

Tento projekt je podporený z Európskeho sociálneho fondu.

4. Dátový rámec Pandas: skombinujeme syntetické údaje do dátového rámca Pandas nazývaného „data“.

Tu je výňatok zo súboru údajov ako príklad:

Obrázok 6 - Výňatok údajov z dátového súboru

```
>>> print(data)
     Age Sex Blood_Type  Smoker  Height  Weight  Phone_Number
0     40   M          A       0     170      70  555-555-0000
1     70   F          A       1     180      70  555-555-0001
2     60   F          O       0     170      90  555-555-0002
3     50   F          A       0     150      70  555-555-0003
4     30   F          A       0     150      90  555-555-0004
..   ...  ..        ...     ...     ...     ...           ...
995   25   F         AB       0     180      80  555-555-0995
996   70   F          A       0     150      80  555-555-0996
997   30   M          A       0     160      80  555-555-0997
998   70   M          A       0     180      70  555-555-0998
999   45   M         AB       0     170      60  555-555-0999

[1000 rows x 7 columns]
```

7.1.1.1 *Definovanie troch jednoduchých funkcií, ktoré ukazujú, či dátový súbor vyhovuje špecifickej hodnote K*

Aby sme overili jednotlivé kroky praktického prípadu, zadefinujeme tri jednoduché funkcie, ktoré nám umožnia zistiť, či množina vyhovuje konkrétnej hodnote K. Na začiatku množina spĺňa triviálnu anonymitu K=1, keďže každý jednotlivec tvorí so sebou skupinu s rovnakým typom záznamu o veľkosti jedna alebo väčšej ako jedna, avšak nespĺňa anonymitu K=2 alebo väčšiu.

Tieto tri funkcie sú inšpirované príkladom načrtnutým v „Úvode do anonymizácie údajov: Techniky a prípadové štúdie“[25], ktorá iniciatívou pre otvorené údaje španielskej vlády.

1. **Definovanie dopytu na vyhľadávanie záznamov**

Pre validáciu je potrebné definovať dopyt na vyhľadávanie záznamov pre skupinu premenných, ktoré boli definované pre proces k-anonymity. V tomto konkrétnom prípade použijeme všetky premenné okrem telefónneho čísla, keďže ide o jedinečný identifikátor a neskôr ho budeme riešiť samostatne:

[25] https://datos.gob.es/es/documentacion/introduccion-la-anonimizacion-de-datos-tecnicas-y-casos-practicos

Tento projekt je podporený z Európskeho sociálneho fondu.

Obrázok 7 - Definovanie dopytu vyhľadávania záznamov

```
def queryKAnonymized(row):
    return f'Age == {row.Age}' \
           f' & Sex == \'{row.Sex}\'' \
           f' & Blood_Type == \'{row.Blood_Type}\'' \
           f' & Smoker == {row.Smoker}' \
           f' & Height == {row.Height}' \
           f' & Weight == {row.Weight}'\
           f' & Phone_Number == \'{row.Phone_Number}\''
```

Na obrázku 7 je možné vidieť ako dotaz jednoducho definuje, ktoré polia sa budú kontrolovať z údajov konkrétneho riadku množiny.

2. Vyhľadanie skupiny v rámci množiny údajov

Po definovaní špecifického dopytu pre konkrétny prípad ho možno posunúť ako argument generickej funkcie, ktorú sme definovali na vyhľadávanie skupín v rámci množiny údajov (*queryKAnonymized* je definovaná ako predvolená hodnota, aby ju nebolo nutné uvádzať). Účelom tejto funkcie je overiť, či všetky príklady patria do skupiny najmenej K jednotlivcov, a teda či súbor spĺňal požiadavky na k-anonymity pre stanovenú hodnotu K (a zároveň všetky nižšie hodnoty K).

Vo všeobecnosti platí, že na určenie, či je súbor údajov k-anonymizovaný, je potrebné zabezpečiť, aby bol každý záznam v súbore údajov nerozoznateľný od najmenej k-iných záznamov s ohľadom na možné identifikátory (ako je vek, pohlavie, PSČ, atď.). To sa dá dosiahnuť zoskupením záznamov na základe identifikátorov a následnou kontrolou, či každá skupina má aspoň k záznamov. Ak majú všetky skupiny aspoň K záznamov, potom je súbor údajov k-anonymizovaný. Ak nie, potom súbor údajov nie je k-anonymizovaný.

Na obrázku 8 sa nachádza funkcia, ktorá umožňuje pre každý riadok skontrolovať, či patrí do skupiny aspoň K záznamov. Ak sa všetky riadky zhodujú s podmienkou, vráti hodnotu True (pravda), ak nájde aspoň 1 riadok, ktorý nespĺňa podmienku, vráti False (nepravda).

Obrázok 8 - Vyhľadanie skupiny v rámci množiny údajov

```
def isKAnonymized(df, k, queryFunction = queryKAnonymized):
    for index, row in df.iterrows():
        if df.query(queryFunction(row)).shape[0] < k:
            return False
    return True

print(isKAnonymized(data, 1))
print(isKAnonymized(data, 2))
```

Ak použijeme funkciu isKAnonymized na pôvodnú množinu, zistíme, že výsledok je platný pre anonymitu K=1, ale nie pre K=2 alebo K>2, to znamená, že pôvodný súbor vyhovuje K-anonymite len do K=1.

Obrázok 9 - Výsledok funkcie *"isKAnonymized"* na dátovom súbore

```
>>> print(isKAnonymized(data, 1))
True
>>> print(isKAnonymized(data, 2))
False
```

3. Vyhľadanie nezoskupených záznamov

Ďalším krokom je definovanie doplnkovej funkcie, aby bolo možné nájsť konkrétne záznamy nespĺňajúce definovanú podmienku K (ako predvolená hodnota je uvedený aj dotaz KAnonymized), to znamená, že cieľom je identifikovať záznamy, ktoré nespĺňajú uvedené podmienky, analyzovať ich a pokúsiť sa ich opraviť prostredníctvom novej iterácie (viď. obrázok 10)

Obrázok 10 - Vyhľadanie nezoskupených záznamov

```
def getNotKAnonymized(df, k, queryFunction = queryKAnonymized):
    rowsNotKAnonymized = pd.DataFrame()
    for index, row in df.iterrows():
        group = df.query(queryFunction(row))
        if group.shape[0] < k:
            rowsNotKAnonymized = pd.concat([rowsNotKAnonymized, group])
    return rowsNotKAnonymized.drop_duplicates()

print(getNotKAnonymized(data, 1))
print(getNotKAnonymized(data, 2))
```

Pri aplikácii tejto funkcie na pôvodnú množinu s K=1 vráti prázdnu množinu, pretože všetky riadky spĺňajú podmienku. Naopak, ak ho použijeme s K=2, vráti kompletnú množinu, pretože žiadny riadok nespĺňa Anonymitu-K pre K=2 alebo vyššie.

Toto je bežný stav pred použitím procesu anonymizácie, bežne záznamy vyhovujú anonymite K=1 (tvoria skupinu s jedným jednotlivcom) a normálne žiadny záznam (alebo väčšina) nevyhovuje vyšším hodnotám K.

V príkladoch nižšie sú prezentované rôzne techniky na zníženie počtu nevyhovujúcich záznamov, na dosiahnutie anonymizovania súborov údajov v súlade s K-anonymitou pre zadané K.

7.1.1.2 Maskovanie údajov zmenou znakov z číslic na hviezdičky

Jednoduchý spôsob, ako generalizovať premennú, je zaokrúhliť alebo zamaskovať najmenej významné hodnoty čísla alebo kód. V tomto prípade sa snažíme generalizovať telefónne číslo maskovaním posledných 8 číslic.

Obrázok 11 - Maskovanie telefónneho čísla

```
# Mask phone number
data['Phone_Number'] = data['Phone_Number'].apply(lambda x: x[:-8] + '*'*8)
```

Výsledok po použití algoritmu na maskovanie telefónneho čísla vyzerá takto:

Obrázok 12 - Maskovanie telefónneho čísla

```
>>> print(data_masked)
    Age_Range Sex Blood_Type  Smoker Height_Range Weight_Range  Phone_Number
0       36-65   M          A       0      140-170        40-70  555-********
1      66-100   F          A       1      171-210        40-70  555-********
2       36-65   F          O       0      140-170       71-100  555-********
3       36-65   F          A       0      140-170        40-70  555-********
4       18-35   F          A       0      140-170       71-100  555-********
..        ...  ..        ...     ...          ...          ...           ...
995     18-35   F         AB       0      171-210       71-100  555-********
996    66-100   F          A       0      140-170       71-100  555-********
997     18-35   M          A       0      140-170       71-100  555-********
998    66-100   M          A       0      171-210        40-70  555-********
999     36-65   M         AB       0      140-170        40-70  555-********
```

Po vykonaní maskovania telefónneho čísla skontrolujeme, či nová množina spĺňa anonymitu K=2. Na obrázku 12 vidíme, že ešte stále existuje 27 riadkov, ktoré nespĺňajú anonymitu K=2.

Obrázok 13 - Kontrola anonymity K=2

```
>>> len(getNotKAnonymized(data_masked,2,queryKAnonymized_masked))
27
```

7.1.1.3 Generalizovanie údajov zoskupovaním

Bežnou formou zovšeobecňovania je hierarchia a/alebo diskretizácia hodnôt. To znamená, že cieľom je zoskupiť rozsahy hodnôt do vopred definovaných kategórií pomocou pravidiel.

Časť kódu nižšie na obrázku 14 zovšeobecňuje atribúty veku, výšky a hmotnosti v množine údajov, aby boli údaje anonymnejšie.

1. **Generalizovanie veku:** Vek je rozdelený do troch širších rozsahov: '18-35', '36-65' a '66-100'. Na vytvorenie týchto zásobníkov s príslušnými štítkami sa používa funkcia ' pd.cut ()'.

2. **Generalizovanie výšky:** Podobne je výška rozdelená do dvoch rozsahov: '140-170' a '171-210' pomocou funkcie ' pd.cut ()'.

3. **Generalizovanie hmotnosti:** Hmotnosť je tiež rozdelená do dvoch rozsahov: '40-70' a '71-100' pomocou funkcie ' pd.cut ()'.

Obrázok 14 - Generalizovanie údajov zoskupovaním

```
# Generalize age into broader age ranges
data['Age_Range'] = pd.cut(data['Age'], bins=[18, 35, 65, 100], labels=['18-35', '36-65', '66-100'])

# Generalize height into broader height ranges
data['Height_Range'] = pd.cut(data['Height'], bins=[140, 170, 210], labels=['140-170', '171-210'])

# Generalize weight into broader weight ranges
data['Weight_Range'] = pd.cut(data['Weight'], bins=[40, 70, 100], labels=['40-70', '71-100'])
```


Tento projekt je podporený z Európskeho sociálneho fondu.

Po generalizovaní je definovaná funkcia s názvom 'queryKAnonymized_masked_generalised ()' na vytvorenie reťazca dopytu, ktorý možno použiť na filtrovanie riadkov s rovnakými generalizovanými hodnotami atribútov.

Anonymizovaná množina údajov 'data_masked_generalised' sa vytvorí výberom generalizovaných stĺpcov spolu s ďalšími stĺpcami, ako sú 'Sex', 'Blood_Type', 'Smoker' a už maskované 'Phone_Number'.

Obrázok 15 - Funkcia data_masked_generalised

```
def queryKAnonymized_masked_generalised(row):
    return f'Age_Range == \'{row.Age_Range}\'' \
           f' & Sex == \'{row.Sex}\'' \
           f' & Blood_Type == \'{row.Blood_Type}\'' \
           f' & Smoker == {row.Smoker}' \
           f' & Height_Range == \'{row.Height_Range}\'' \
           f' & Weight_Range == \'{row.Weight_Range}\''\
           f' & Phone_Number == \'{row.Phone_Number}\''

# Use the anonymized variables in the anonymized dataset
data_masked_generalised = data[['Age_Range', 'Sex', 'Blood_Type', 'Smoker', 'Height_Range', 'Weight_Range',
print(data_masked_generalised)
```

Výstup po použití funkcie 'data_masked_generalised':

Obrázok 16 - Výstup po 'data_masked_generalised'

```
>>> print(data_masked_generalised)
    Age_Range Sex Blood_Type  Smoker Height_Range Weight_Range  Phone_Number
0       36-65   M          A       0      140-170        40-70  555-********
1      66-100   F          A       1      171-210        40-70  555-********
2       36-65   F          O       0      140-170       71-100  555-********
3       36-65   F          A       0      140-170        40-70  555-********
4       18-35   F          A       0      140-170       71-100  555-********
..        ...  ..        ...     ...          ...          ...           ...
995     18-35   F         AB       0      171-210       71-100  555-********
996    66-100   F          A       0      140-170       71-100  555-********
997     18-35   M          A       0      140-170       71-100  555-********
998    66-100   M          A       0      171-210        40-70  555-********
999     36-65   M         AB       0      140-170        40-70  555-********
```

Následne sa počet riadkov, ktoré nie sú anonymizované, vypočíta pomocou funkcie „getNotKAnonymized()“ v generalizovanom súbore údajov pomocou funkcie „queryKAnonymized_masked_generalised“.

Obrázok 17 - Výsledok funkcie "getNotKAnonymized"

```
>>> len(getNotKAnonymized(data_masked_generalised,2,queryKAnonymized_masked_generalised))
27
```

Po použití funckie 'data_masked_generalised' na zistenie stavu anonymizácie je výsledkom stále 27 riadkov, ktoré nespĺňajú požiadavku anonymity K=2 (viď obrázok 17), čo znamená, že použité techniky generalizovania neboli úspešné pri zvyšovaní K-anonymity.

Tento projekt je podporený z Európskeho sociálneho fondu.

7.1.1.4 Potlačenie

Príklad kódu demonštruje techniku nazývanú potlačenie, čo je metóda anonymizácie používaná pri analýze údajov na ochranu súkromia. Potlačenie zahŕňa úplné odstránenie určitých atribútov zo súboru údajov na ochranu súkromia jednotlivcov.

Obrázok 18 - Kód na potlačenie typu údaju

```
# Drop Smoker column
data_suppressed = data.drop(columns=['Smoker'])
```

V predmetnom súbore údajov je stĺpec „Fajčiar“ potlačený, aby sa chránilo súkromie jednotlivcov. Rozhodnutie potlačiť tento atribút môže byť založené na niekoľkých dôvodoch:

- **Citlivosť:** Fajčenie možno považovať za citlivú informáciu, pretože súvisí s osobnými návykmi a zdravím jednotlivca. Odhalenie takýchto informácií môže v určitých kontextoch viesť k diskriminácii alebo stigmatizácii jednotlivcov.
- **Riziko opätovnej identifikácie:** Zahrnutie stĺpca „fajčiar“ do súboru údajov môže zvýšiť riziko opätovnej identifikácie, najmä v kombinácii s inými atribútmi. Potlačením tohto stĺpca sa znižuje pravdepodobnosť jednoznačnej identifikácie jednotlivca, čím sa zvyšuje ochrana súkromia.
- **Súlad s nariadeniami:** V niektorých prípadoch môžu nariadenia o ochrane údajov, ako je GDPR (alebo HIPAA), vyžadovať potlačenie určitých atribútov, ktoré môžu byť potenciálne použité na identifikáciu jednotlivcov, či už priamo alebo nepriamo. Odstránenie stĺpca „Fajčiar“ môže pomôcť zabezpečiť súlad s týmito nariadeniami.

Po potlačení stĺpca 'Fajčiar' sa počet riadkov, ktoré nespĺňajú anonymitu K=2, zníži a zostane iba 9 záznamov, ktoré napĺňajú anonymitu K=2. To naznačuje, že potlačenie tohto atribútu zlepšilo ochranu súkromia množiny údajov.

Obrázok 19 - Kontrola záznamov nespĺňajúcich K=2 po potlačení

```
>>> len(getNotKAnonymized(data_masked_generalised_supressed,2,queryKAnonymized_masked_generalised_supressed))
9
```

V ďalšom kroku sa potlačí stĺpec telefónne číslo a krvná skupiny, pretože sa tiež považujú za citlivé informácie, ktoré možno potenciálne použiť na identifikáciu jednotlivcov.

Tento projekt je podporený z Európskeho sociálneho fondu.

Obrázok 20 - Potlačenie telefónneho čísla a krvnej skupiny

```
# Drop Phone Number and Blood Type columns
data_suppressed = data.drop(columns=['Phone_Number', 'Blood_Type'])

# Generalize variables in the anonymized dataset
data_masked_generalised_supressed = data_suppressed[['Age_Range', 'Sex', 'Height_Range', 'Weight_Range']]
print(data_masked_generalised_supressed)

def queryKAnonymized_masked_generalised_supressed(row):
    return f'Age_Range == \'{row.Age_Range}\'' \
           f' & Sex == \'{row.Sex}\'' \
           f' & Height_Range == \'{row.Height_Range}\'' \
           f' & Weight_Range == \'{row.Weight_Range}\''

len(getNotKAnonymized(data_masked_generalised_supressed,2,queryKAnonymized_masked_generalised_supressed))
```

Znova skontrolujeme K=2 a v tomto prípade sa to úspešne dosiahne:

Obrázok 21 - Kontrola záznamov nespĺňajúcich K=2 po potlačení ďalších údajov

```
>>> len(getNotKAnonymized(data_masked_generalised_supressed,2,queryKAnonymized_masked_generalised_supressed))
0
```

Dátový súbor, ktorý spĺňa anonymitu K=2 vyzerá takto:

Obrázok 22 - Dátový súbor spĺňajúci K=2

	Age_Range	Sex	Height_Range	Weight_Range
0	36-65	M	140-170	40-70
1	66-100	F	171-210	40-70
2	36-65	F	140-170	71-100
3	36-65	F	140-170	40-70
4	18-35	F	140-170	71-100
..	...	..	...	...
995	18-35	F	171-210	71-100
996	66-100	F	140-170	71-100
997	18-35	M	140-170	71-100
998	66-100	M	171-210	40-70
999	36-65	M	140-170	40-70

Teraz skontrolujeme, či množina údajov spĺňa podmienku k=3, k=4 a k=5, čo je náš konečný cieľ. (viď. obrázok 23)

Obrázok 23 - Kontrola K=3,4,5

```
# Check k-anonymity and number of not anonymized rows
k_values = [1, 2, 3, 4, 5]

not_k_anonymized_rows = []

for k in k_values:
    anonymized = isKAnonymized(data_masked_generalised_supressed, k, queryKAnonymized_masked_generalised_supr
    not_k_anonymized_rows.append(len(getNotKAnonymized(data_masked_generalised_supressed, k, queryKAnonymized
    print(f"For k = {k}, is the dataset k-anonymized? {anonymized}. Number of not k-anonymized rows: {not_k_a

print(f"Number of not k-anonymized rows for k={k_values}: {not_k_anonymized_rows}")
```

Chyba! Nenašiel sa

Po kontrole vyšších stupňov anonymity vidíme, že k-anonymita až do k=3 je splnená, pričom stále existuje jeden riadok, ktorý nespĺňa k=4, a dva riadky, ktoré nevyhovujú k=5.

Obrázok 24 - Výsledok kontroly K=3,4,5

```
For k = 1, is the dataset k-anonymized? True. Number of not k-anonymized rows: 0
For k = 2, is the dataset k-anonymized? True. Number of not k-anonymized rows: 0
For k = 3, is the dataset k-anonymized? True. Number of not k-anonymized rows: 0
For k = 4, is the dataset k-anonymized? False. Number of not k-anonymized rows: 1
For k = 5, is the dataset k-anonymized? False. Number of not k-anonymized rows: 2
```

Aby sme vyriešili problém riadkov, ktoré nie sú k-anonymizované (pre k=4 a k=5), rozhodli sme sa ich odstrániť z množiny údajov namiesto použitia ďalších techník generalizovania alebo potlačenia. Naše odôvodnenie tohto rozhodnutia je založené na malom počte ovplyvnených riadkov, čo znamená, že ich odstránenie bude mať minimálny vplyv na celkový súbor údajov. Okrem toho by použitie ďalších techník anonymizácie mohlo ďalej znížiť užitočnosť súboru údajov. Odstránením týchto neanonymizovaných riadkov môžeme zachovať rovnováhu medzi ochranou súkromia jednotlivcov a zachovaním užitočnosti súboru údajov na analýzu.

Obrázok 25 - Odstránenie problémových záznamov

```
# Drop the not k-anonymized rows
data_k_anonymized = data_masked_generalised_supressed.drop(np.concatenate([not_k_anonymized_rows_4, not_k_anc

# Iterate until all non-k-anonymized rows are dropped
while True:
    not_k_anonymized_rows = []
    for k in k_values:
        not_k_anonymized_rows.append(getNotKAnonymized(data_k_anonymized, k, queryKAnonymized_masked_generali
    not_k_anonymized_rows = np.concatenate(not_k_anonymized_rows)
    if len(not_k_anonymized_rows) == 0:
        break
    data_k_anonymized = data_k_anonymized.drop(not_k_anonymized_rows)

# Check k-anonymity and number of not anonymized rows
k_values = [1, 2, 3, 4, 5]

not_k_anonymized_rows = []

for k in k_values:
    anonymized = isKAnonymized(data_k_anonymized, k, queryKAnonymized_masked_generalised_supressed)
    not_k_anonymized_rows.append(len(getNotKAnonymized(data_k_anonymized, k, queryKAnonymized_masked_generali
    print(f"For k = {k}, is the dataset k-anonymized? {anonymized}. Number of not k-anonymized rows: {not_k_a

print(f"Number of not k-anonymized rows for k={k_values}: {not_k_anonymized_rows}")
```

Aby sme tak urobili, najprv získame riadky, ktoré nespĺňajú anonymitu K=4 a K=5, využitím funkcie „ getNotKAnonymized ()“ z množiny údajov, s hodnotou k = 4 alebo 5 a funkciou dopytu. Indexy riadkov uložíme do 'not_k_anonymized_rows_4' a 'not_k_anonymized_rows_5'.

Následne vytvoríme novú množinu údajov s názvom ' data_k_anonymized ' vypustením riadkov z ' data_masked_generalised_supressed ', ktoré nie sú 4-anonymné alebo 5-anonymné. Spustíme cyklus, ktorý pokračuje, kým sa z množiny údajov neodstránia všetky neanonymné riadky. Vo vnútri cyklu inicializujeme prázdny zoznam s názvom ' not_k_anonymized_rows '. Iterujeme cez každú hodnotu k v 'k_values' a získame indexy ne-anonymných riadkov pomocou funkcie 'getNotKAnonymized()'. Tieto indexy pripájame k 'not_k_anonymized_rows'. Ďalej zreťazíme všetky indexy v 'not_k_anonymized_rows' a skontrolujeme, či je dĺžka nula. Pokiaľ už neexistujú žiadne ďalšie záznamy nevyhovujúce K-anonymite, prerušíme slučku. V opačnom prípade pokračujeme vypúšťaním riadkov z 'data_k_anonymized'.

Tento projekt je podporený z Európskeho sociálneho fondu.

Ďalším krokom je skontrolovanie k-anonymity výsledného súboru údajov 'data_k_anonymized' pre každú hodnotu k v 'k_values' využitím funkcie 'isKAnonymized ()' a funkciou 'getNotKAnonymized ()' si zobrazíme výsledok, či je súbor údajov k-anonymný a počet riadkov, ktoré nevyhovujú pre každú hodnotu k.

Výstup vyzerá takto:

Obrázok 26 - Finálny výsledok kontroly K=3,4,5

```
For k = 1, is the dataset k-anonymized? True. Number of not k-anonymized rows: 0
For k = 2, is the dataset k-anonymized? True. Number of not k-anonymized rows: 0
For k = 3, is the dataset k-anonymized? True. Number of not k-anonymized rows: 0
For k = 4, is the dataset k-anonymized? True. Number of not k-anonymized rows: 0
For k = 5, is the dataset k-anonymized? True. Number of not k-anonymized rows: 0
```

Anonymizovaná množina údajov, ktorá dosiahla k=5, je vyobrazená na obrázku 27, pričom celkovo bolo odstránených 7 riadkov v dôsledku vypustenia neanonymizovaných riadkov.

Obrázok 27 - Obsah dátového súboru po anonymizovaní minimálne na k=5

	Age_Range	Sex	Height_Range	Weight_Range
0	36-65	M	140-170	40-70
1	66-100	F	171-210	40-70
2	36-65	F	140-170	71-100
3	36-65	F	140-170	40-70
4	18-35	F	140-170	71-100
..	...	..	...	...
995	18-35	F	171-210	71-100
996	66-100	F	140-170	71-100
997	18-35	M	140-170	71-100
998	66-100	M	171-210	40-70
999	36-65	M	140-170	40-70

[993 rows x 4 columns]

Ako sme videli, existujú rôzne stratégie a algoritmy, ktoré umožňujú dosiahnuť anonymizáciu súboru údajov, ide o zložitý proces a úzko súvisí s účelom použitia a úrovňou zachovania užitočnosti dát a požadovaných záruk kvality a bezpečnosti dát.

Na tomto príklade sme videli využitie niekoľkých jednoduchých metód a algoritmov anonymizovania bez využitia vyčerpávajúcich analýz, cieľom príkladu bolo prezentovať aplikáciu rôznych techník. Keďže tiež neexistuje žiadny univerzálny predpis týkajúci sa použitia akejkoľvek špecifickej techniky, je najlepšie použiť vhodnú kombináciu, ktorá zahŕňa aspoň generalizovania, na zmenu mierok alebo rádov veľkosti a potlačenie.

7.2 L-diverzita & T-blízkosť

K-anonymita, tradičná technika zvyšujúca súkromie, ponúka základnú úroveň ochrany generalizovaním a potlačením atribútov v súbore údajov, aby sa vytvorili nerozoznateľné záznamy. K-anonymita však zaostáva, keď čelíme útokom na podobnosť údajov a znalosti pozadia (dodatočných informácií) pre špecifické záznamy. Na prekonanie týchto obmedzení výskumníci vyvinuli robustnejšie metódy anonymizácie údajov, ako je L-diverzita a T-blízkosť. L-diverzita vyžaduje, aby každá trieda ekvivalencie v rámci súboru údajov obsahovala aspoň 'L' odlišných citlivých hodnôt, čím sa pridáva ďalšia vrstva ochrany proti útokom na prepojenie atribútov. Na druhej strane v prípade T-blízkosti ide

o krok ďalej tým, že zabezpečuje, aby sa distribúcia citlivých atribútov v rámci každej triedy ekvivalencie veľmi podobala celkovej distribúcii v celom súbore údajov, čím sa znižuje riziko narušenia súkromia vyplývajúce z využívania distribúcie citlivých atribútov. Tieto pokročilé techniky spoločne dopĺňajú k-anonymitu a výrazne zlepšujú ochranu súkromia, čo organizáciám umožňuje využiť silu analýzy údajov a zároveň chrániť dôvernosť jednotlivcov a dodržiavať predpisy o ochrane osobných údajov.

7.2.1 L-diverzita

L-diverzita je pokročilá technika na zvýšenie ochrany súkromia, ktorá výrazne zlepšuje anonymizáciu údajov tým, že rieši prirodzené obmedzenia tradičných metód, ako je napríklad k-anonymita. Snaží sa zabezpečiť, aby každá trieda ekvivalencie (skupina záznamov zdieľajúcich podobné atribúty) v rámci súboru údajov obsahovala aspoň „L“ odlišných citlivých hodnôt. Tento inovatívny prístup účinne znemožňuje odhalenie identity a prepojenie atribútov, a to aj v situáciách, keď má útočník predchádzajúce znalosti o dotknutých osobách.

Vo prelomovom dokumente „ℓ-Diversity: Privacy Beyond k-Anonymity“[26] výskumníci zistili, že k-anonymita samotná nezaručuje súkromie kvôli svojej zraniteľnosti voči homogenite a útokom na znalosti na pozadí. V dôsledku toho navrhli model L-diverzity ako robustnejšie riešenie na presadenie rozmanitosti v citlivých atribútoch v každej triede ekvivalencie. Okrem toho dokument predstavil rôzne typy L-diverzity – vrátane L-diverzity založenej na entropii, rekurzívnej (c, l)-diverzity a odlišnej L-diverzity – z ktorých každá poskytuje rôzne úrovne ochrany súkromia prispôsobené špecifickým požiadavkám aplikácie.

Uvažujme o hypotetickom súbore zdravotných záznamov pacientov z nemocnice, ktorý obsahuje nasledujúce atribúty: vek, pohlavie, PSČ a diagnóza. Nemocnica chce tento súbor údajov zdieľať na výskumné účely a zároveň zabezpečiť súkromie pacientov. Pomocou k-anonymity sa súbor údajov generalizuje a záznamy sa zoskupia do tried ekvivalencie na základe ich spoločných atribútov (vek, pohlavie a PSČ).

Pracujme s nasledujúcim súborom údajov s k=2:

Tabuľka 18 - Príklad anonymizovaného dátového súboru s k=2

Vek	rod	PSČ	Diagnóza
25-34	M	123**	Chrípka
25-34	M	123**	Chrípka
25-34	F	123**	Astma
25-34	F	123**	Diabetes
45-54	M	456**	Diabetes
45-54	M	456**	Diabetes
45-54	F	456**	Chrípka
45-54	F	456**	Astma

[26] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 2006, pp. 24-24, doi: 10.1109/ICDE.2006.1.

Tento súbor údajov spĺňa anonymitu k=2, pretože každý záznam je na nerozoznanie od aspoň jedného iného záznamu. Stále je však náchylný na útoky spojené s atribútom, pretože prvé dva záznamy s rovnakým vekom, pohlavím a PSČ zdieľajú rovnakú diagnózu (chrípka).

Na vyriešenie tejto zraniteľnosti sa použila l-diverzita. Táto metóda zaisťuje, že každá trieda ekvivalencie má aspoň 'l' odlišné citlivé hodnoty (v tomto prípade diagnózy). Útočníkovi tak výrazne sťaží opätovnú identifikáciu pacienta prepojením diagnózy s inými dostupnými informáciami.

Aby sme to dosiahli, musíme upraviť súbor údajov tak, aby každá skupina mala aspoň dve rôzne diagnózy. Jedným zo spôsobov, ako to dosiahnuť, je upraviť diagnózy v rámci každej skupiny tak, aby sa navzájom líšili.

Aby sme mohli použiť L-diverzitu, zabezpečíme, aby každá trieda ekvivalencie mala aspoň 'L' odlišných citlivých hodnôt (v tomto prípade ide o diagnózu). Povedzme, že potrebujeme 2-diverzitu:

Tabuľka 19 - Aplikovanie l-diverzity (l=2) na príklade

Vek	rod	PSČ	Diagnóza
25-34	M	123**	Chrípka
25-34	M	123**	Astma
25-34	F	123**	Astma
25-34	F	123**	Diabetes
45-54	M	456**	Diabetes
45-54	M	456**	Chrípka
45-54	F	456**	Chrípka
45-54	F	456**	Astma

Vidíme, že v každej skupine sú minimálne dve rôzne diagnózy. Napríklad prvá skupina (vek 25-34 rokov, muž, PSČ 123**) obsahuje jeden záznam s diagnózou „Chrípka“ a druhý s diagnózou „Astma“. Tým sa dosiahne požiadavka 2-diverzity, ktorú sme sa rozhodli dosiahnuť.

Aplikovaním l-diverzity sme zlepšili ochranu súkromia súboru údajov pri zachovaní jeho analytickej hodnoty. Je to preto, že sme neupravili žiadne necitlivé atribúty (vek, pohlavie a PSČ), takže množinu údajov možno stále použiť na analýzu bez toho, aby došlo k ohrozeniu súkromia jednotlivcov v množine údajov.

Aplikovaním L-diverzity môžeme ochrániť citlivé informácie tým, že zabezpečíme, aby každá trieda ekvivalencie mala aspoň 'L' odlišných hodnôt (v tomto prípade diagnózy), čím sa pre útočníka zvyšuje problém presne určiť citlivé informácie jednotlivca.

Tento súbor údajov demonštruje, ako pomocu L-diverzity navrhnutej v článku dôjde k zlepšeniu anonymizácie údajov poskytovaním komplexnej ochrany proti útokom na prepojenie atribútov. Udržiavaním vyššieho stupňa variability citlivých informácií v rámci každej triedy ekvivalencie L-diverzita úspešne zmierňuje riziko opätovnej identifikácie a posilňuje ochranu súkromia jednotlivca. Výsledkom je, že táto metóda umožňuje organizáciám využiť silu analýzy údajov a súčasne dodržiavať nariadenia o ochrane osobných údajov a chrániť dôvernosť osobných informácií.

7.2.2 T-blízkosť

T-blízkosť je pokročilá metóda na ochranu súkromia, ktorá stavia na základoch položených k-anonymitou a L-diverzitou, čím sa ďalej zvyšuje ochrana citlivých údajov v anonymizovaných súboroch údajov. T-blízkosť je vyvinutá na riešenie obmedzení L-diverzity, zameriava sa na udržanie distribúcie citlivých atribútov v rámci každej triedy ekvivalencie, čím sa zabezpečí, že zostane dostatočne podobná celkovej distribúcii týchto atribútov v celom súbore údajov. Týmto spôsobom T-blízkosť účinne zmierňuje riziko narušenia súkromia vyplývajúce zo zverejnenia hodnôt citlivých atribútov alebo zneužitia distribúcie týchto hodnôt v rámci tried ekvivalencie.

Hlavným princípom T-blízkosti je, že vzdialenosť hýbateľa zeme (EMD) medzi distribúciou citlivého atribútu v triede ekvivalencie a distribúciou toho istého atribútu v celom súbore údajov nesmie byť väčšia ako vopred definovaný prah „T“. Tento prístup zaručuje, že útočník nemôže odvodiť citlivý atribút jednotlivca s pravdepodobnosťou výrazne vyššou, ako je pravdepodobnosť odvodená z celkového rozloženia tohto atribútu, čím sa zabezpečí silnejšia ochrana súkromia.

Využijeme nasledujúci hypotetický súbor údajov o zdravotných záznamoch pacientov, ktorý obsahuje atribúty ako vek, pohlavie, PSČ a diagnóza. Súbor údajov bol anonymizovaný pomocou K-anonymity a L-diverzity:

Tabuľka 20 - Príklad dátového súboru po aplikovaní k-anonymita a l-diverzity

Vek	rod	PSČ	Diagnóza
25-34	M	123**	Chrípka
25-34	M	123**	Astma
25-34	F	123**	Astma
25-34	F	123**	Diabetes
45-54	M	456**	Diabetes
45-54	M	456**	Chrípka
45-54	F	456**	Chrípka
45-54	F	456**	Astma

Teraz predpokladajme, že celkové rozloženie diagnóz v súbore údajov je nasledovné: chrípka (40%), astma (30%) a diabetes (30%). Na dosiahnutie T-blízkosti zabezpečujeme, aby sa distribúcia diagnóz v rámci každej triedy ekvivalencie veľmi podobala celkovej distribúcii. Príklad množiny údajov vyhovujúcej t-blízkosti:

Tabuľka 21 - Využitie t-blízkosti na príklade

Vek	rod	PSČ	Diagnóza
25-34	M	123**	Chrípka
25-34	M	123**	Astma
25-34	F	123**	Chrípka
25-34	F	123**	Diabetes
45-54	M	456**	Diabetes
45-54	M	456**	Chrípka
45-54	F	456**	Astma

Tento projekt je podporený z Európskeho sociálneho fondu.

Chyba! Nenašiel sa

Vek	rod	PSČ	Diagnóza
45-54	F	456**	Chrípka

V tomto príklade má každá trieda ekvivalencie rozdelenie diagnóz, ktoré je podobné celkovému rozdeleniu: chrípka (50 %), astma (25 %) a diabetes (25 %). Zachovaním tejto podobnosti T-blízkosť účinne chráni citlivé údaje pred narušením súkromia vyplývajúcim zo zneužitia distribúcie citlivých atribútov v rámci tried ekvivalencie, čím poskytuje robustnejšie riešenie anonymizácie údajov. Pri väčšine dátových súborov by malo byť cieľom dosiahnutie t-blízkosti rovnej 0,2, ktorá zaučuje dostatočnú anonymitu a zároveň výrazne neznehodnotí dáta, avšak ideálna hodnota sa môže pri jednotlivých dátových súborov líšiť, a teda je nutné pre každý dátový súbor vypočítať túto hodnotu samostatne.

Tento projekt je podporený z Európskeho sociálneho fondu.

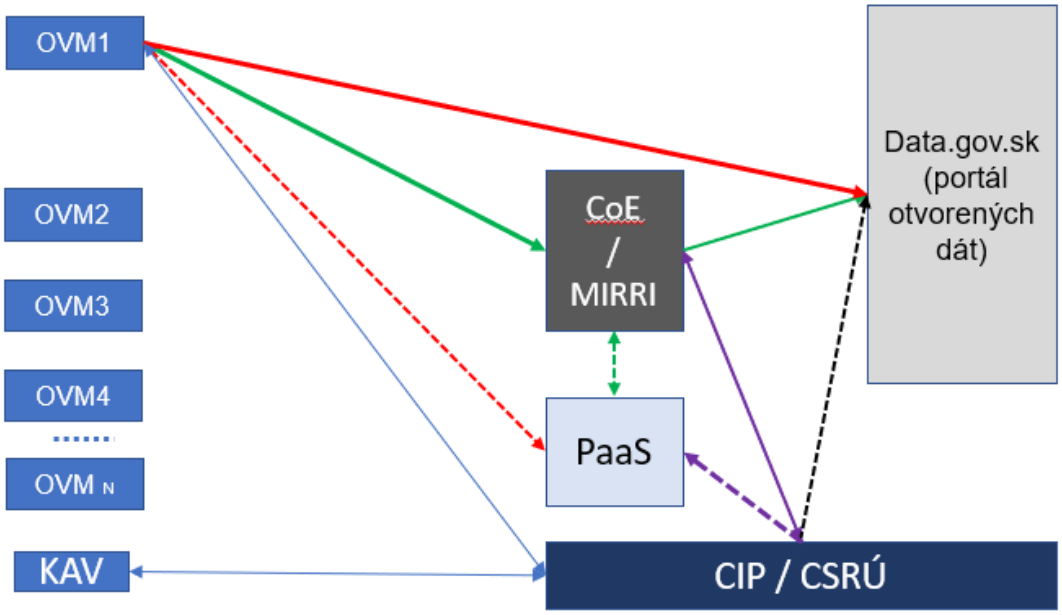
8 Výber nástrojov pre anonymizáciu otvorených údajov

Táto kapitola sa v prvej časti venuje kritériám na výber nástroja na anonymizáciu štruktúrovaných dát. V druhej časti v krátkosti pojednáva o možnostiach nástroja na anonymizovanie neštruktúrovaných údajov pre vyššie opísaný prípad použitia CRZ.

Výber nástroja na anonymizovanie štruktúrovaných údajov musí zohľadňovať súčasnú situáciu ISVS, do ktorej bude implementovaný, keďže stav manažmentu údajov verejnej správy výrazne ovplyvňuje možnosti výberu takéhoto nástroja. Anonymizácia je zameraná na ochranu citlivých údajov, preto od akéhokoľvek nástroja, ktorý je zvažovaný musíme na prvom mieste vyžadovať maximalizáciu ich bezpečnosti. V súčasnosti sa zdrojové dáta nenachádzajú v jednom dátovom úložisku, a teda nie je jednoduché kontrolovať vzťahy medzi údajmi naprieč rôznymi dátovými súbormi, ktoré budú anonymizované a zverejňované ako otvorené dáta, z čoho pramení riziko úniku citlivých dát spájaním rôznych dátových súborov útočníkom. To znižuje efektivitu ľubovoľného vybratého nástroja. Na jej zabezpečenie je nutné v prvom rade nastaviť manažment údajov tak, aby podporoval proces anonymizácie napríklad tak, ako je to navrhnuté v 5 kapitole.

Ďalej sa venujeme prístupu k výberu anonymizačného nástroja pre centrálne riešenie, ktoré by bolo poskytované prostredníctvom prístupu PaaS pre jednotlivé OVM. Pri výbere takéhoto centrálneho riešenia je dôležité zdôrazniť, že by mali byť vypracované, zjednotené, prioritizované a štruktúrované biznis požiadavky a pravidlá – pričom cieľová dátová architektúra a plán na dosiahnutie výsledného modelu fungovania anonymizovania už musí byť vytvorený. Požiadavky na výber anonymizačného nástroja umiestneného do prostredia PaaS by mali byť v ideálnom prípade brané do úvahy spoločne s požiadavkami na výber nástroja na riadenie dátovej kvality, o ktorom predpokladáme, že bude v krátkom čase implementovaný v prostredí PaaS.

Obrázok 28 Možnosti realizácie anonymizácie prostredníctvom PaaS



Uvažujeme s 3 možnosťami toku anonymizácie (pričom je možná aj ich prípadná kombinácia). Z prezentovaných možností s ohľadom na súčasnú organizáciu správy dát preferujeme možnosť na obrázku vyobrazenú zelenou farbou.

1. **Centrálna koordinácia (zelená) –** každé OVM samostatne realizuje anonymizáciu prostredníctvom služby PaaS v koordinácii s vyhradenou autoritou, ktorou môže byť napríklad MIRRI alebo na to určená iná organizácia (koncept Centra excelentnosti). Táto autorita okrem iného informuje ostatné OVM o tom, aké údaje budú zverejňované prostredníctvom anonymizovaných dátových súborov. Zároveň je táto autorita zodpovedná za schválenie predtým, ako je anonymizovaný dátový súbor publikovaný na portáli otvorených dát. Krížové informovanie je pri zverejňovaní takýto dát obzvlášť dôležité s ohľadom na krížovú kontrolu (identifikácia prípadných rizík opätovnej identifikácie citlivých dát použitím útoku spájaním rôzny ch datasetov). V tomto bode bude dôležitá zodpovedná príprava správ z anonymizácie, ktoré vypracuje OVM pri realizácii anonymizovania dátového súboru.

2. **Decentrálna koordinácia (červená) –** v porovnaní s prvou možnosťou (zelená) OVM nahrávajú anonymizované údaje prostredníctvom služby PaaS priamo na portál otvorených dát a teda schvaľovanie zverejnenia anonymizovaného dátového súboru a koordinácia s ostatnými OVM je v kompetencii OVM, ktoré anonymizovanie vykonáva. Tento variant odporúčame len v krajnom prípade, nakoľko predstavuje ešte väčší nápor na správnu koordináciu a vyššie riziko nedostatočnej informovanosti ostatných OVM a súvisiacu vyššiu hrozbu potencionálneho úniku citlivých dát.

3. **Centrálna realizácia (fialová) –** Táto možnosť počíta so scenárom, že všetky súbory údajov sú uložené centrálne (napr. nasadenie dátového skladu v rámci IS CSRÚ). Dátové súbory sú následne anonymizované prostredníctvom PaaS a publikované na portál otvorených dát. Za tento proces bude zodpovedať vyhradená autorita ako v prvom prípade (zelený). Táto možnosť je výhodná kvôli tomu, že anonymizovanie je realizované v rámci jedného úložiska a je teda najjednoduchšie sledovať vzťahy medzi anonymizovanými dátovými súbormi, ktoré budú zverejňované. Avšak realizácia tejto možnosti je značne komplikovanejšia v iných aspektoch ako pri predchádzajúcich variantoch (medzi iným aj pre nutnosť budovania robustného dátového skladu).

Okrem uvedených možností anonymizácie pomocou centrálneho nástroja je samozrejme možné aj riešenie lokálne, teda anonymizácia na úrovni OVM s použitím už v existujúcich databázových riešení, pomocou nástrojov v nich obsiahnutých alebo vytvorením vlastných anonymizačných skriptov. Podobne je možné vybrať hotové „krabicové“ riešenia na lokálne použitie. Tieto riešenia pri súčasnom množstve publikovania otvorených dát, a teda reálneho množstva dátových súborov, ktoré bude potrebné anonymizovať možno považovať za dostačujúci, pokiaľ bude dodržaný postup riadenia procesu anonymizácie uvedený v kapitole 5. Požiadavkám na takéto nástroje sa s ohľadom na množstvo lokálnych špecifík nevenujeme.

8.1 Faktory a kritéria výberu anonymizačných nástrojov

V závislosti od definovaných cieľov a účelov výsledku anonymizácie (na interné použitie, na externé/verejné použitie, na generovanie vzoriek AI atď.) sa prístup k spracovaniu údajov môže líšiť.

Dátová stratégia, biznis pravidlá a zásady súvisiace s údajmi (vrátane riadenia prístupu, kybernetickej bezpečnosti a rolí) by mali byť vypracované pred úvahami o výbere nástroja na anonymizáciu.

Tento projekt je podporený z Európskeho sociálneho fondu.

Vytvorenie dátovej stratégie, biznis pravidiel a zásad súvisiacich s údajmi musí zohľadniť aj priamo požiadavky, ktoré vychádzajú z potrieb na anonymizáciu. Toto sú niektoré z globálnych požiadaviek, ktoré vyplývajú z nám známeho architektonického rámca riadenia dát v SR:

- Nástroj by mal fungovať v private cloud režime pre OVM
- Nástroj by mal zabezpečiť kontrolu voči spájaniu anonymizovaného dátového súboru s už zverejnenými dátovými súbormi.
- Nástroj by mal spĺňať rôzne úrovne nastavenia anonymity dátových súborov

Nasledujúce faktory/kritériá by mali slúžiť ako podklad na posúdenie a hodnotenie nástroja pri jeho výbere. Prvým krokom pri použití tohto podkladu by malo byť určenie dôležitosti jednotlivých faktorov od 1 do 3 (3 najvyššia dôležitosť, 1 najnižšia dôležitosť) v rámci špecifického prípadu, pre ktorý je vyberaný nástroj na anonymizovanie. Následne pre každý uvažovaný nástroj treba stanoviť hodnotu od 1 do 5 pre každý faktor, pričom vyššia hodnota hovorí o tom, že daný nástroj spĺňa dané kritérium viac, ako ten s nižšou hodnotou. Násobením faktora dôležitosti s faktorom zhody dostaneme hodnotenie zvažovaných nástrojov pre výber.

Tabuľka 22 - Kritériá výberu anonymizačného nástroja

Faktor	Popis	Dôležitosť & Hodnotenie
Nariadenia	Zodpovedá všetkým príslušným nariadeniam a zákonným požiadavkám, vrátane povinných noriem zverejňovania (napr. európske iniciatívy o otvorených údajoch a API).	
Vlastnosti anonymizovanej množiny údajov	Je schopný pokryť požiadavky na vlastnosti, kvalitu a zložitosť požadovaného výsledku anonymizácie, výsledný formát súboru údajov, štruktúru a spracovávaný objem dát („enterprise grade“ riešenie)	
Frekvencia	Je schopný pokryť požiadavky na frekvenciu prípravy na zverejnenie a zverejňovanie požadovaných anonymizovaných súborov údajov	
Personál a roly	Sú dostupní odborníci na úlohy anonymizácie údajov s požadovanou odbornosťou a schopnosťou vykonávať požadované postupy. Úlohy, ktoré sa majú prideliť, najmä zodpovednosť za spracovanie údajov, zvolený prístup (podľa popisu v tabuľke vyššie – centrálny vs. decentrálny)	
Ekosystém	Súlad s existujúcou IT/dátovou infraštruktúrou, požiadavkami na integrácie, dátové toky a ďalšími požiadavkami súvisiacimi so životným cyklom dát.	
Cieľové umiestnenie	Súlad s požiadavkami na zverejňovanie (na aké účely bude zverejnený výsledný anonymizovaný súbor údajov, kde a akým spôsobom má byť zverejnený a pod.)	
Dostupné iné zdroje	Aké funkcie anonymizácie sú už dostupné v súčasnosti nainštalovaných/vyvíjaných riešeniach a aplikáciách, ktoré nie je potrebné duplikovať?	

Podrobný popis prístupu k štandardizácii anonymizácie údajov vrátane techník, metodológie, osvedčených postupov, politík, rizík / výziev, možných scenárov / prípadov použitia, knižníc a nástrojov je uvedený v dokumente 1.1.5.

Príklady často využívaných anonymizačných nástrojov

Pre komplexné riešenie anonymizácie, ktoré by bolo integrované do IS CSRÚ vo forme PaaS, je možné zvážiť viacero nástrojov, ktoré sú dostupné na trhu. Výber bude zrejme ovplyvnený rozhodnutím o prístupe k výberu nástroja na riadenie kvality dát, ktorý práve prebieha pre túto platformu v rámci projektu CIP/CSRU. Nižšie uvádzame tri anonymizačné nástroje (resp. nástroje obsahujúce možnosti anonymizácie), ktoré sa veľmi často opakujú ako bežná voľba pri takomto výbere. Ako sme uviedli, existujú aj ďalšie, ktoré je možné zvážiť, najmä ak by bol zvolený „single vendor“ prístup k výberu nástroja na riadenie kvalitu dát pre PaaS.

Talend – Nie je úplne plnohodnotným anonymizačným nástrojom, avšak v rámci svojich funkcionalít obsahuje aj funkciu maskovania, ktorú je možné využiť na maskovanie údajov podľa zadefinovaných kritérií na vybranom sémantickom type v rámci celého dátového súboru, pre ktorý je maskovanie vybrané.

ARX[27] **–** Je to open source anonymizačný nástroj, ktorý podporuje rôzne techniky anonymizácie, ako je generalizácia, škálovanie a permutácia. Je vyvíjaný na Viedenskej univerzite technológií a ekonomiky a je k dispozícii pod licenciou Apache 2.0. ARX podporuje anonymizáciu veľkého objemu údajov.

G9 (Esito.no)[28] **–** Je anonymizačný nástroj, ktorý poskytuje plne programovateľnú logiku anonymizácie. Je ľahké ju replikovať a možno ju vykonávať vo viacerých databázach. Nástroj tiež dokáže vymazať vybrané údaje. Umožňuje aj vytváranie databáz úplne od počiatku aj so syntetickými dátami alebo je možné pri špeciálnych požiadavkách kombinovať anonymizáciu so syntetickými dátami.

8.2 Nástroj na anonymizovanie neštruktúrovaných údajov

Na trhu sa nachádza množstvo riešení na anonymizovanie citlivých informácií v neštruktúrovaných údajoch. Pre účely tohto výstupu uvažuje prípad použitia pre zmluvy v Centrálnom registri zmlúv, ktorý považujeme za typ neštruktúrovaných dokumentov, ktoré obsahujú citlivé informácie a pred ich zverejnením je teda nutné ich anonymizovať. Požadujeme hlavne, aby umožňoval tak manuálne ako aj automatické vyhľadávanie citlivých informácií v dokumente a anonymizáciu tak, aby nebolo možné opätovné odhalenie citlivých informácií.

Podľa týchto požiadaviek sme ako príklad vybrali nástroj **Docbyte**[29], ktorý užívateľom poskytuje tieto funkcionality:

— **Pokročilé strojové učenie a umelá inteligencia** – umelú inteligenciu a strojové učenie tento nástroj využíva na rozpoznanie a anonymizáciu v jednotlivých dokumentoch. Strojové učenie umožňuje tomuto nástroju pracovať bez presne

[27] https://arx.deidentifier.org/
[28] https://www.esito.no/en/products/anonymizer/
[29] https://www.docbyte.com/anonymization/

definovaných šablón na vykonávanie anonymizácie, čo zaručuje, že citlivý obsah v dokumentoch bude vždy správne spracovaný.

— **Rozpoznávanie objektov** – pomocou rôznych algoritmov dokáže tento nástroj analyzovať obrázky, čo umožňuje nielen redigovať text, ale aj rozmazávať alebo zatemňovať obrázky.

— **Spracovanie v reálnom čase** – nástroj zaisťuje anonymizáciu informácií a obrázkov hneď, ako prechádzajú cez určitý bod v rámci systému.

— **Spĺňa všetky GDPR požiadavky** – Docbyte je uspôsobený na spĺňanie potrieb ochrany osobných údajov, ktoré vyplývajú z požiadaviek GDPR.

Tento nástroj je jedným z najlepších riešení pre anonymizovanie dokumentov v reálnom čase s využitím strojového učenia. Strojové učenie je to, čo tento nástroj robí jasnou voľbou pre anonymizovanie dokumentov v reálnom čase, keďže nevyžaduje vytváranie šablón na anonymizovanie, čo je veľkou výhodou, keďže v prípade CRZ by museli existovať stovky takýchto šablón pre zabezpečenie anonymizácie všetkých zmlúv vstupujúcich do CRZ. Tento nástroj pomocou strojového učenia automaticky určí všetky informácie, ktoré sú považované za citlivé a podľa nastavenia ich anonymizuje vo vybraných dokumentoch na vstupe do systému. Z tohto dôvodu je viac ako vhodným kandidátom na integráciu do systému CRZ pre zabezpečenie dodržiavania ochrany osobných údajov a iných citlivých údajov.

Tento projekt je podporený z Európskeho sociálneho fondu.

9 Záver – zhrnutie odporúčaní

Pre zjednodušenie orientácie v predkladanom texte uvádzame zhrnutie odporúčaní a kompetencií pre OVM a MIRRI (dátová kancelária / CoE) vyplývajúcich z tohto výstupu:

MIRRI

- Určiť spôsob dohľadu a správy zverejňovania otvorených údajov (vrátane ich anonymizácie) (viď. kapitola 5)
- Zabezpečiť pravidelnú osvetu a proaktívnu komunikáciu s OVM a vyškolenie zodpovedných a výkonných pracovníkov, vrátane prípravy školiacich materiálov. (viď. kapitola 5)
- Zabezpečiť rozvoj anonymizačnej platformy prostredníctvom platformy PaaS v rámci projektu CIP. (viď. kapitola 5)
- Zabezpečiť definovanie a správne nastavenie biznis požiadaviek na výber anonymizačného nástroja, výber varianty modelu fungovania nástroja a jeho implementácia do prostredia ISVS (viď. Kapitola 8)
- Zabezpečiť dostatočnú úroveň centrálnej podpory vrátane podpory technologickej (nástroj na anonymizáciu), metodickej a praktickej (FAQ, telefonická odborná podpora). (viď. kapitola 5)

OVM

- Stanoviť roly a zodpovednosti (viď. kapitola 5)
- Stanoviť lokálnu metodiku / procesy anonymizácie údajov na zverejnenie ako otvorených dát (viď. kapitola 5)
- Rozhodnúť o technológii využívanej na úrovni OMV v oblasti anonymizácie (viď. kapitola 5)
- Definovať atribúty správy z vykonanej anonymizácie (viď. kapitola 5)
- Zabezpečiť kontinuálne vzdelávanie v oblasti anonymizácie (viď. kapitola 5)

Contact us

Rudolf Sedmina
partner
Management Consulting
E rsedmina@kpmg.sk

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

www.kpmg.com

Tento projekt je podporený z Európskeho sociálneho fondu.