VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ
DEPARTMENT OF TELECOMMUNICATIONS

OSOBNÍ ÚDAJE A ANONYMIZACE
PERSONAL DATA AND ANONYMISATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Petra Kajánková**
AUTHOR

VEDOUCÍ PRÁCE — **JUDr. MgA. Jakub Míšek, Ph.D.**
SUPERVISOR

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Informační bezpečnost**

Ústav telekomunikací

Studentka: Petra Kajánková ***ID:*** 211550

Ročník: 3 ***Akademický rok:*** 2020/21

NÁZEV TÉMATU:

Osobní údaje a anonymizace

POKYNY PRO VYPRACOVÁNÍ:

Široké vymezení pojmu osobní údaj vede k situaci, kdy do režimu ochrany osobních údajů dopadá velké množství informačních procesů. Způsobem, který zároveň vede ke zvýšení úrovně ochrany práv a zájmů subjektu údajů, a který zároveň umožní volné nakládání s předmětnými informacemi mimo režim právní úpravy, je efektivní anonymizace. V teoretické části se bakalářská práce zaměří na definici pojmu osobní údaj. Dále analyzuje různé anonymizační techniky a vyhodnotí vhodnost jejich použití. Zároveň popíše a odůvodní výběr typových datových sad pro praktickou část. Praktická část bude spočívat ve vytvoření anonymizéru, který bude použitelný pro vybraný typ datových sad.

DOPORUČENÁ LITERATURA:

[1] Ohm, P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review. 2009, č. 6, s. 1701–1777.

[2] Nonnemann, F. Objektivní, či subjektivní pojetí osobních údajů? Právní rozhledy. 2015, č. 12, s. 425–431.

Termín zadání: 1.2.2021 ***Termín odevzdání:*** 31.5.2021

Vedoucí práce: JUDr. MgA. Jakub Míšek, Ph.D.

doc. Ing. Jan Hajný, Ph.D.
předseda rady studijního programu

ABSTRAKT

Bakalářská práce se zaměřuje především na problematiku osobních údajů společně s jejich anonymizací. Teoretická část specifikuje platné české a evropské legislativní dokumenty, přibližuje podmínky anonymizace. V neposlední řadě představuje různé anonymizační techniky (pseudonymizace, anonymizace), kterými lze vytvořit anonymizovaná data. Praktický výstup práce tvoří naprogramovaný anonymizátor, který je schopen na konkrétní datové sadě naplnit požadavky vycházející z práva evropského, a naplnit tedy i podmínky, kdy data již nejsou brána jako osobní údaje.

KLÍČOVÁ SLOVA

Osobní údaje, GDPR, Anonymizace, Pseudonymizace, Správce, Zpracovatel, Subjekt údajů, K-anonymita

ABSTRACT

The Bachelor thesis focuses mainly on the issue of personal data and their anonymisation. The theoretical part specifies valid Czech and European legislative documents and also introduces the issue of anonymisation. It presents also various anonymisation methods (Pseudonymisation, Anonymisation) which can be used to create anonymized data. The practical part of this thesis is a program which is able to fulfill the requirements based on European law on a specific dataset. Outgoing „personal" data from the attached program are no longer taken as personal data.

KEYWORDS

Personal data, GDPR, Anonymisation, Pseudonymisation, Data Controller, Data Processor, Data Subject, k-Anonymity

KAJÁNKOVÁ, Petra. *Osobní údaje a anonymizace*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2021, 58 s. Bakalářská práce. Vedoucí práce: JUDr. MgA. Jakub Míšek, Ph.D.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Petra Kajánková
VUT ID autora:	211550
Typ práce:	Bakalářská práce
Akademický rok:	2020/21
Téma závěrečné práce:	Osobní údaje a anonymizace

Prohlašuji, že svou závěrečnou práci jsem vypracovala samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autorky*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Zde bych ráda poděkovala vedoucímu bakalářské práce panu JUDr. MgA. Jakubovi Míškovi, Ph.D. za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Evropská unie, ve snaze sjednotit podmínky pro zpracování osobních údajů, k jejichž ochraně se zavázala Listinou základních práv EU, nejprve vydala v roce 1995 směrnici 95/46/ES, kterou později musela nahradit nařízením č. 2016/679. Vydáním nařízení však způsobila rozruch v oblasti zpracování osobních údajů z důvodu nepřesného stanovení vytyčených hranic kolem pojmu osobní údaj. Vznikl zde problém, a tedy i nejasnost, která data jsou chráněna garantovaným právem. Bylo potřeba stanovit jasné hranice vymezující vlastnosti osobního údaje. Tohoto úkolu se zhostil Soudní dvůr EU, který jednoznačně vymezil potřebné hranice ve zlomovém soudním procesu týkajícím se otázky, zda dynamická IP adresa je osobním údajem. Následkem rozhodnutí došlo k sjednocení výkladu napříč jednotlivými státy. Zmíněné nařízení však i oddělilo pojmy anonymizace, pseudonymizace a nastavilo tak podmínky, kdy data, ačkoliv prošla krokem pseudonymizace, nesplňují potřebné vlastnosti anonymizovaných dat, a tak musí být stále pod ochranou evropského práva. Tyto konkrétní kroky učiněné Evropskou unií popisuje podrobněji bakalářská práce.

Cílem teoretické části je nejen přiblížit hranice stanovené kolem pojmu osobní údaj, jinými slovy, která data jsou z právního aspektu chráněna jako osobní údaje, ale i objasnit další pojmy často zmiňované v legislativě jako je zpracování, správce, zpracovatel, subjekt údajů, a především představit teoretické možnosti způsobu anonymizace. Cíl praktické části je následně realizovat zvolený proces umožnující anonymizaci údajů.

První teoretická část se zaměřuje na aktuální problematiku sběru, zpracování dat takovým způsobem, aby byla v souladu s podmínkami vycházejícími z evropského práva. Kapitola představuje nejen legislativní dokumenty (omezující sběr, zpracování a předávání dat) na již zmíněné úrovni evropského práva, jako je například zakotvení práva na ochranu osobních údajů do Listiny základních práv EU, jejíž právní síla je srovnatelná s právní silou ústavního práva, ale i na úrovni ústavního práva aplikovaného v České republice. Druhá část následně představuje různé techniky umožňující vytvořit nejen anonymizovaná, ale i pseudonymizovaná data s využitím prvků z kryptografie, zobecněním nebo odstraněním informací a v neposlední řadě zanesením rozumných chyb do údajů.

Praktická část práce spočívá v navržení systému, který bude schopen realizovat vybraný proces anonymizace na konkrétní datové sadě. Kapitola nejprve uvádí důvody výběru dané datové skupiny (snaha o co nejširší využití v praktickém životě, nebo na již existující databáze), společně s postupným popisem, jak program od začátku načítá, kontroluje, krok po kroku zpracovává a vyhodnocuje výsledky. Praktický výstup práce bude tvořit naprogramovaná spustitelná aplikace, která bude schopna realizovat veškeré kroky uvedené v návrhu a vytvoří očekávaná anonymní

data, která ani po propojení s dalšími informacemi nepovedou k jednoznačné identifikaci.

1 Osobní údaje

Kapitola se zabývá pojmy: přímý, nepřímý osobní údaj, citlivý osobní údaj, identifikace, správce, zpracovatel, subjekt údajů. Dále zde budou popsány důvody pro vymezení nepřesných hranic pojmu osobní údaj, působnost, spolu s povinnostmi plynoucími z nařízení EU č. 2016/679.

1.1 Legislativa týkající se sběru, zpracování a předávání osobních údajů

Data nesoucí informace o uživatelích jsou mnohdy hlavními finančními zdroji pro mnohé nadnárodní korporace. Facebook, Google, Microsoft denně sbírají nespočet dat (Big Data)[1], na základě kterých je možné vytvořit vzorce chování, upravit reklamu na míru či poskytovat cílené marketingové nabídky. Jak uvádí Ohm ve své knize Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization: *„Full retreat to a tort-based privacy regime, which would abandon the forty-year preventative turn in privacy law, would be a grave mistake, because without regulation, the easy reidentification result will spark a frightening and unprecedented wave of privacy harm by increasing access to what I call the "database of ruin." The database of ruin exists only in potential: It is the worldwide collection of all of the facts held by third parties that can be used to cause privacy-related harm to almost every member of society."*[2], neomezený sběr, zpracování dat představují potencionální riziko pro každého člena společnosti. Tyto důvody tak zapříčinily vznik několika zákonů, směrnic, nařízení, jejichž snahou je regulovat nakládání s daty obsahujícími jakékoliv osobní údaje.[1, 2]

Ochrana před neoprávněným nakládáním s osobními údaji je již zakotvena v ústavním právu České republiky. Cílem Listiny základních práv a svobod vydané předsednictvem České národní rady 16. prosince 1992, je stanovit přirozená, neporušitelná, rovnocenná práva všem občanům ČR. V čl. 10 odst. 3 této listiny je stanoveno: *„Každý má právo na ochranu před neoprávněným shromažďováním, zveřejňováním nebo jiným zneužíváním údajů o své osobě.“* Touto listinou se tak Česká republika zavazuje k ochraně osobních údajů v rovině nejvyšší možné právní síly.[3]

Následky možného porušení, tedy zásahy do zmíněného ústavního práva (například zveřejňování, sdělení, přisvojování údajů) jsou pak vymezeny v Trestním zákoníku č. 40/2009 Sb. platném od 9. února 2009. Konkrétně v § 180 zabývajícím

[1] Jedná se o pojem označující enormní množství informací.

[2] Citováno z: OHM, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 2009, s.1746.[2]

se neoprávněným nakládáním s osobními údaji. Paragraf je rozdělen do několika odstavců hovořících o sběru s odlišnými okolnostmi. Odst. 1 hovoří o sběru v souvislosti s výkonem veřejné moci, odst. 2 se pak zabývá sběrem: *„v souvislosti s výkonem svého povolání, zaměstnání nebo funkce, a způsobí tím vážnou újmu na právech nebo oprávněných zájmech osoby, jíž se osobní údaje týkají.“* Dopustí-li se správce, zpracovatel takovéhoto činu, hrozí jim v případě rozsáhlého zásahu trest odnětí svobody až na 8 let, zákaz činnosti.[4]

O další regulaci se postarala i Evropská unie, která vydala omezení zpracování a podmínky sběru údajů platné po celém svém území. Daná nařízení a směrnice se však nevztahují pouze na státy spadající pod EU, ale i na všechny správce, kteří cílí ať webovými stránkami či jinými prostředky na její občany[3].

Obdobně jako v českém právním řádu, je i v evropském zakotvena ochrana osobních údajů již v Listině základních práv Evropské unie, jež byla vyhlášena 7. prosince 2000, s účinností však až od 1. prosince 2009. Listina, podobně jako ústavní právo České republiky, představuje základní, neporušitelná práva všech občanů spadajících pod Evropskou unii. Konkrétní ustanovení týkající se ochrany osobních údajů lze nalézt pod čl. 8, který zároveň i vymezuje podmínky sběru takovýchto údajů (odst. 2). Dále občanům garantuje dohled nad plněním všech povinností dané EU nezávislým orgánem (odst. 3). Tímto orgánem je v České republice Úřad na ochranu osobních údajů, jehož úkolem je nejen výše zmíněný dohled, ale i příjem podnětů a stížností týkajících se samotného sběru, zpracování.[5, 6]

Důležitým milníkem v regulaci zpracování osobních údajů se stal 24. říjen 1995, kdy Evropský parlament spolu s Radou Evropské unie vydal směrnici s označením 95/46/ES. Cílem bylo sjednotit a ustanovit rovná práva všem občanům při zpracování osobních údajů. S vývojem technologií pro zpracování dat však nastal problém nedostatečného omezení samotného zpracování, které se stalo komplexnějším a zároveň rizikovějším pro subjekty údajů. Správci osobních údajů ovšem i naráželi na problém nesjednocení všech pravidel zpracování napříč jednotlivými státy. Na rozdíl od nařízení, směrnice musí být implementována každým státem do právního řádu samostatně, což zapříčinilo vznik drobných odlišností. Správcům často hrozily nerovnocenné postihy za nedodržení podmínek napříč celým územím. Z těchto důvodu bylo třeba vymezit nařízení, které by řešilo veškeré nedostatky této směrnice.[7, 8]

Proto dne 27. dubna 2016 bylo vydáno nařízení EU 2016/679, které je též známo pod zkratkou GDPR[4] s účinností od 25. května 2018. Zároveň byla zrušena platnost výše zmíněné směrnice 95/46/ES. Hlavním cílem bylo tentokrát stanovit identická pravidla napříč celou EU. Nařízení navíc stanovuje vysoké pokuty za porušení vy-

[3]Více v kapitole 1.3 Zpracování osobních údajů.

[4]Celým názvem General Data Protection Regulation.

mezených podmínek, a to až do výše 4 procent ročního globálního obratu[5] či částku 20 000 000 EUR. Výše sankce je posuzovaná dle rozsahu a vážnosti zásahu. V případě vážného, rozsáhlého porušení pravidel je volen vyšší peněžní trest[6].[9, 10]

1.2 Pojem osobní údaj

Na první pohled pojem osobní údaj není těžké specifikovat. GDPR popisuje osobní údaj jako jakoukoliv informaci vedoucí k identifikaci subjektu údajů. Konkrétněji lze ustanovení najít v čl. 4 odst. 1, který blíže specifikuje, co vše je pod tímto pojmem zahrnuto: *„například jméno, identifikační číslo, lokační údaje, síťový identifikátor nebo na jeden či více zvláštních prvků fyzické, fyziologické, genetické, psychické, ekonomické, kulturní nebo společenské identity této fyzické osoby.“*

Výše zmíněný pojem zahrnuje i velmi citlivé informace, které na základě zveřejnění mohou poškodit subjekt údajů. EU v GDPR čl. 9 odst. 1, zabývajícím se zpracováním této zvláštní kategorie osobních údajů, zakazuje zpracování citlivých údajů, nenaplní-li se jakákoliv výjimka vycházející z čl. 9 odst. 2. Mezi stanovené zákonné výjimky patří:

- Výslovný souhlas subjektu údajů, není-li to v rozporu s přímým zákazem EU, členského státu, aby subjekt údajů zrušil zákaz vyplývající z čl. 9 odst. 1.
- Za účelem plnění: povinnosti, ochrany životně důležitých zájmů, určení, výkonu či obhajoby právních nároků, veřejného zájmu, archivace[7], vědeckého, historického výzkumu, zpracování údajů, které subjekt sám zveřejnil.

Samotné vymezení hranic kolem pojmu citlivý osobní údaj je pak zakomponováno v zmíněném čl. 9. konkrétně v odst. 1. Jedná se o údaje, které vypovídají o: *„rasovém nebo etnickém původu, politických názorech, náboženském vyznání nebo filosofickém přesvědčení nebo členství v odborové organizaci, genetický údaj*[8]*, biometrický údaj*[9] *zpracovávaný za účelem jedinečné identifikace fyzické osoby, údaj o zdravotním stavu*[10]*, o sexuálním chování, o sexuální orientaci a údaj týkající se rozsudků v trestních věcech a trestných činů nebo souvisejících bezpečnostních opatření.“*[11] Správce a zpracovatel v případě zpracování této zvláštní kategorie údajů musí zvážit citlivost

[5]Jedná se o celkový obrat nadnárodní společnosti působící na trhu, vydělaný za jeden rok ve všech zemích, kde působí.

[6]Porovnává se, zda 4 procentní globální výdělek je vyšší než částka 20 000 000 EUR.

[7]Pouze v případech veřejného zájmu.

[8]GDPR v čl. 4 odst. 13 tyto údaje charakterizuje jako jedinečné informace vypovídající o fyziologii, zdraví vyplývající z biologického vzorku.

[9]GDPR v čl. 4 odst. 14 tyto údaje charakterizuje jako údaje nesoucí fyzické, fyziologické znaky či znaky chování vedoucí k jedinečné identifikaci.

[10]GDPR v čl. 4 odst. 15 tyto údaje charakterizuje jako údaje týkající se duševního, fyzického zdraví, údaje o poskytnutí zdravotních služeb, které vypovídají o zdravotním stavu.

[11]Citováno z: GDPR čl. 9 odst. 1.[9]

samotné informace[12] z důvodů, že takovéto osobní údaje mohou poškodit subjekt údajů.[13][9, 11]

1.2.1 Proces identifikace

Zveřejněním každého osobního údaje vzniká jisté riziko identifikace. Tento proces lze chápat jako postupné odhalování identity subjektu údajů, kdy s každým dalším údajem se okruh potencionálních subjektů zužuje až do úplného stavu identifikace – spojení dat s konkrétním subjektem.[2]

Příkladem takového procesu může být případ pochybení americké společnosti AOL poskytující internetové služby, která 4. srpna 2006 zveřejnila výsledky vyhledávání 650 000 uživatelů sbíraných po dobu tří měsíců. AOL rozlišoval jednotlivé uživatele přidělenými identifikátory, což zapříčinilo oddělení jednotlivých výsledků vyhledávání. Ačkoliv na první pohled se výsledky vyhledávání samy o sobě nejeví jako osobní údaj, mohou právě obsahovat takováto data vedoucí k výše popsané identifikaci. Američtí novináři deníku The New York Times Michael Barbaro, Tom Zeller Jr. ve svém článku A Face Is Exposed for AOL Seacher No. 4417749 odhalili na základě zveřejněných údajů identitu uživatele s označením 4417749. Daný uživatel vyhledával nejen své příjmení, ale i informace o aktivitách v blízkém okolí. Ačkoliv data byla krátce po zveřejnění společností stažena, na internetu se i přesto objevovaly kopie, což vedlo k dalším dvěma identifikacím. I přestože identita ostatních zůstala skryta, byly v datech nalezeny citlivé osobní údaje vypovídající o psychickém rozpoložení, fyzickém stavu uživatelů. Vše vyústilo hromadnou žalobou, kdy americký soud obvinil AOL z porušení tehdejších zákonů o ochraně osobních údajů a požadoval odškodnění nejméně 5 000 $ za každou identifikovanou osobu. Z výše uvedeného příkladu tak vychází myšlenka, že i na první pohled nevýznamné informace mohou dopomoci dokončit proces identifikace, tak jak se tomu stalo u uživatele 4417749.[12, 13, 14]

1.2.2 Vytyčení hranic pojmu osobní údaj

Pojem osobní údaj popisuje dřívější směrnice 95/46/ES čl. 2 odst. a jako informaci vedoucí přímo nebo nepřímo k identifikaci jejího subjektu. Všeobecně lze pak na tyto údaje nahlížet dvěma možnými přístupy pojetí. Prvním přístupem je subjektivní pojetí, kdy je možnost identifikace hodnocena z pozice osoby pracující s danou informací. V takovém případě se jedná o osobní údaj, má-li zmíněna osoba reálnou možnost získat dodatečné informace (s ohledem například na časové, finanční náklady) a zároveň informaci lze přímo propojit se subjektem údajů. Druhým možným přístupem je objektivní pojetí, kdy je hodnoceno, zda jakákoliv osoba má reálnou

[12] Tato podmínka nastává i v případě zveřejnění osobního údaje.

[13] Podmínkou zpracování je i naplnění jakékoliv uvedené výjimky plynoucích z GDPR čl. 9 odst. 2.

možnost získat dodatečné informace v případě, že informace není přímo propojitelná s jejím subjektem. Na základě rozhodovací praxe týkající se osobních údajů lze tvrdit, že aktuální nařízení nahlíží na pojem osobní údaj objektivním přístupem - hodnotí tedy možnost identifikace z pohledu jakékoliv osoby.[15]

Z definice osobního údaje popsané výše vychází zřejmé rozdělení dle síly identifikace na přímý a nepřímý osobní údaj. Toto rozdělení následně převzalo i nařízení GDPR.

Jako přímý osobní údaj jsou charakterizovaná data, která nejen významně při procesu identifikace zužují potencionální skupinu subjektů, ale jsou i přímo propojitelná s konkrétním subjektem. Příkladem mohou být zmíněné citlivé údaje, lokační údaje, jméno, příjmení, datum narození, místo narození, rodné číslo, rodinný stav, videozáznamy, fotografie, titul, povolání, informace finančního, trestního charakteru.

Z praktického hlediska se v případě nepřímého osobního údaje jedná především o informace, které samy o sobě nevedou k identifikaci, avšak v některých případech k ní mohou dopomoci (na základě propojení s dalšími údaji). V předběžné otázce položené soudnímu dvoru EU[14] v případu, kdy bylo nutno posoudit, zda dynamická IP adresa je osobním údajem, byl tento pojem charakterizován: *„aby mohla být informace kvalifikována jako osobní údaj, není nutné, aby tato informace sama o sobě umožňovala identifikovat subjekt údajů.“*[15][2, 16]

Zmíněný případ se týká věci Patrik Breyer, který se domáhal, aby Spolková republika Německo zakázala uchovávat IP adresu třetími osobami po odpojení z veřejně přístupných stránek německých spolkových orgánů. Z rozhodnutí německého soudu po výkladu nepřímého údaje SDEU vyplývá, že dynamická IP adresa je nepřímým osobním údajem z důvodu možného propojení s dalšími údaji, kterými disponuje poskytovatel služeb či subjekt sám poskytl internetové stránce. Na základě tohoto propojení by následně mohlo být možné realizovat proces re-identifikace a narušit tak soukromí subjektu údajů.[2, 17]

1.3 Zpracování osobních údajů

Cílem podkapitoly je především vymezit hranice mezi zákonným, nezákonným zpracováním a přiblížit pojmy subjekt údajů, správce, zpracovatel.

[14] Dále pouze SDEU.

[15] Citováno z: Rozsudek Soudního dvora Evropské unie ze dne 19. 10. 2016 ve věci č. C-582/14, Breyer, 2016, odst. 41.[16]

1.3.1 Působnost GDPR

Snaha specifikovat tentokrát co nejvšeobecnější nařízení vedla k výsledku, kdy je GDPR vztaženo na všechny osobní údaje bez ohledu na využité technologie. V čl. 2 je odst. 1 věnován právě této problematice. Z uvedeného odstavce vyplývá nutná aplikace všech podmínek zpracování plynoucí z GDPR, a to bez ohledu, zda jsou osobní údaje zpracovávány automatizovanými[16], neautomatizovanými[17] technologiemi, pouze ale na takové údaje, které budou nebo již jsou zařazeny v evidenci.[9, 18]

Nařízení navíc stanovuje výjimky, kdy není nutno aplikovat požadované podmínky na zpracovávané údaje, které budou i přesto zařazovány do evidence, a bude se tedy jednat o naplnění charakteristik vyplývajících z přechozího odstavce. Jedná se především o činnost:

- nespadající pod působnost evropského práva - typicky úpravy týkající se spolupráce;
- spadající pod působnost zahraniční a bezpečnostní politiky vyplývající z uzavřené Smlouvy o Evropské unii - smlouva zakládající EU. Cílem bezpečnostní politiky je především ochrana společných hodnot, bezpečnost, garance základních práv a svobod[19];
- osobní či domácí - kdy fyzická osoba zpracování uskuteční bez jakýchkoliv finančních výdělků za účelem osobní potřeby. Výjimka na takovéto zpracování nastává pouze tehdy, není-li databáze veřejně dostupná;
- v souladu s cíli směrnice 2016/680, kterými jsou: *„ochrana fyzických osob v souvislosti se zpracováním osobních údajů příslušnými orgány za účelem prevence, vyšetřování, odhalování či stíhání trestných činů nebo výkonu trestů, o volném pohybu těchto údajů.“*[18] Avšak stát, orgán poskytující na základě této směrnice osobní údaje, by měl ověřit bezpečnost zpracování osobních údajů státu, kterému budou údaje předány. V případě nedostatečné bezpečnosti může odmítnout předání. Na území EU by podmínky zpracování měly být rovnocenné, a tak poskytnutí údajů závisí na účelu žádosti.

Nařízení v neposlední řadě ani neopomíjí zpracování mimo území EU v těch případech, je-li činnost správce cílena na subjekty nacházející se na jejím území[19]. Ustanovení týkající zpracování mimo EU lze nalézt v čl. 3 odst. 1: *„Toto nařízení se vztahuje na zpracování osobních údajů v souvislosti s činnostmi provozovny správce nebo zpracovatele v Unii bez ohledu na to, zda zpracování probíhá v Unii či mimo ni.“*,

[16] Jedná se o zpracování využívající typicky libovolný automatizovaný software.

[17] Jedná se o zpracování bez využití automatizovaného softwaru.

[18] Citováno ze: Směrnice o ochraně fyzických osob 2016/680.[20]

[19] Může se jednat o doručení do ČR, platba v CZK, překlad stránek do českého jazyka, který nezajistí prohlížeč.

pouze týká-li se zpracování při činnosti spojené s nabídkou zboží, služeb[20], monitorování chování. Probíhá-li zpracování údajů mimo EU, avšak na území, na kterém se na základě mezinárodního práva veřejného uplatňuje právo členského státu, musí správce dodržet všechny podmínky stanovené nařízením GDPR. Ve všech jiných případech se na zahraniční správce podmínky nevztahují.[9]

1.3.2 Podmínky zpracování

Pojem zpracování chápe GDPR jako operaci či soubor operací s osobními údaji. Pro bližší specifikaci jsou v nařízení uvedeny příklady možného zpracování. Jedná se především o činnosti: *„shromáždění, zaznamenání, uspořádání, strukturování, uložení, přizpůsobení nebo pozměnění, vyhledání, nahlédnutí, použití, zpřístupnění přenosem, šíření nebo jakékoliv jiné zpřístupnění, seřazení či zkombinování, omezení, výmaz nebo zničení.“*[21]

Rozhoduje-li správce, zpracovatel u prováděného úkonu s daty, zda patří do vyhrazených hranic pro pojem zpracování a zároveň je prováděno pro stanovený účel, je úkon dle GDPR chápán jako zpracování. Kupříkladu zveřejnění pořízených fotografií svých zaměstnanců na webových stránkách firmy odpovídá výše popsanému zpracování. Pořídí-li zaměstnanec fotografie za účelem osobní potřeby zaměstnanců z výjezdové konference, nejedná se o zpracování osobních údajů.[9]

Tyto údaje následně musí být zpracovávány pouze v rozsahu stanoveného účelu. Podmínka vychází z GDPR čl. 5 odst. 1 písm. b: osobní údaje *„shromažďovány pro určité, výslovně vyjádřené a legitimní účely a nesmějí být dále zpracovávány způsobem, který je s těmito účely neslučitelný.“* Výjimka zmíněného ustanovení, popsaná čl. 6 odst. 4, dovoluje změnu pouze v případech, je-li nový účel slučitelný s předchozím. Podmínkou změny je ovšem fakt, že data nesmí být zpracovávána na základě výslovného souhlasu nebo zákonným zmocněním. V těchto případech nelze změnit účel zpracování a správce musí získat nový souhlas k novému účelu.[9, 21]

V neposlední řadě mezi další zásady nutné při sběru, zpracování, které musí být dodrženy patří:

- minimalizace dat – zpracování pouze nutných dat ke stanovému účelu;
- zákonnost zpracování[22];
- omezení časového uložení - stanovení doby, po jakou budou údaje uloženy;
- zajištění integrity - neoprávněná osoba nemůže data přepsat;
- důvěrnosti - neoprávněná osoba nemá přístup k osobním údajům.[22, 23]

[20] K naplnění této podmínky není vyžadována po subjektu údajů jakákoliv platba.
[21] Citováno z: GDPR čl. 4 odst. 2.[9]
[22] Podmínky zpracování lze nalézt v GDPR čl.6.[9]

1.3.3 Vztah mezi správcem a zpracovatelem

Cílem GDPR je především sjednotit podmínky zpracování, předávání osobních údajů. Díky těmto podmínkám jsou regulovány činnosti správců a zpracovatelů. Přesné vymezení pojmu správce popisuje GDPR čl. 4 odst. 7: *„fyzická nebo právnická osoba, orgán veřejné moci, agentura nebo jiný subjekt, který sám nebo společně s jinými určuje účely a prostředky zpracování osobních údajů.“* Dle znění tedy správcem mohou být nejen fyzické osoby, ale i právnické osoby, úřady, obce, státy. Zpracovatele lze následně charakterizovat jako entitu, která pro správce vykonává samotný akt zpracování. Uvedené pojmy nařízení chápe jako zodpovědné subjekty, které jsou pověřeny dodržováním veškerých podmínek.[23] V opačném případě hrozí sankce právě takovýmto subjektům.[9]

Zpracování pak může být vykonáváno správcem či jím pověřenou osobou – zpracovatelem. V GDPR je tento pojem chápán čl. 4. odst. 8 jako: *„fyzická nebo právnická osoba, orgán veřejné moci, agentura nebo jiný subjekt, který zpracovává osobní údaje pro správce.“*, na něž stále dopadají vyplývající podmínky z nařízení.[18]

Mezi správcem a zpracovatelem musí být uzavřena smlouva, kde se zpracovatel zaváže k plnění zpracování, mlčenlivosti a dodržení všech dohodnutých podmínek. Součástí smlouvy by měl být účel, kategorie osobních údajů, doba trvání, povaha zpracování, povinnosti a práva správce. Při výběru vhodného zpracovatele má správce povinnost přihlédnout k podmínce[24] hovořící o posouzení, zda zpracovatel nabývá dostatečné kompetence k žádané činnosti. Posuzovanými prvky jsou organizační, technické, bezpečnostní záruky. Zpracovatel tyto kompetence může dokázat například certifikací ISO[25].[9]

1.3.4 Subjekt údajů

Subjektem údajů je dle GDPR živá[26] fyzická osoba, kterou na základě přímých, nepřímých osobních údajů lze identifikovat. Ochrana vyplývající z nařízení nedopadá na právnické osoby a jejich data nejsou chráněna tímto nařízením.[24]

Vzhledem k postavení subjektu údajů, správce a zpracovatele, je subjekt chápán jako slabší smluvní strana a jsou mu přiřčena práva, kterých se může domáhat:

Poskytnutí sbíraných informací na žádost subjektu. Správce musí poskytnout účel, kategorii údajů, dobu zpracování, způsob zpracování[27], příjemce spolu s informacemi o předávání dat třetím stranám, mezinárodním organizacím, zpracovává-li

[23] Ustanoveno v GDPR čl. 5 Zásady zpracování osobních údajů odst. 2.

[24] Podmínka vyplývá z GDPR čl. 28 odst. 1.

[25] Jedná o se organizaci stanovující bezpečnost jednotlivých certifikací.

[26] Podmínka vyplývá z GDPR recitálu 27, který hovoří o ochraně dat pouze žijících osob.

[27] Subjekt údajů má právo odmítnout být součástí rozhodování založeného na automatizovaném zpracování. Právo vychází z GDPR čl. 22.

data subjektu, který toto právo využil. První vydání přehledu informací je poskytnuto bez úplaty, další mohou být zpoplatněna přiměřenou částkou. Správce na základě tohoto práva nemusí vydávat kopie nosičů – listin, dokumentů.

Obdobné je pak právo na přenositelnost údajů, kdy má subjekt možnost požadovat po správci údaje, které vědomě, aktivně poskytl na základě výslovného souhlasu a jsou zpracovávány automatizovaným způsobem, ve strukturovaném, strojově čitelném, běžném formátu. Tyto údaje má pak právo subjekt předat dalšímu správci, nejsou-li tím dotčena práva dalších osob. Správce nemůže subjektu bránit ve výkonu svých práv nepředáním jeho osobních údajů.

Subjekt údajů má dále právo požádat správce o doplnění neúplných informací, opravu zpracovávaných údajů, jedná-li se o nepřesné, zastaralé osobní údaje. Toto právo se opírá o povinnost správce zpracovávat pouze přesné informace[28]. Upozorní-li subjekt na tuto skutečnost, správce musí pozastavit zpracování údajů a ověřit jejich korektnost.

Právo na výmaz, též známé jako „právo být zapomenut“, které nařizuje správcům neprodlené odstranění všech osobních údajů tykající se daného subjektu. Zmíněné právo se však aplikuje pouze v případech, kdy:

- zpracovávané údaje nejsou potřebné pro daný účel. Jedná se o porušení podmínky minimalizace dat vycházející z povinnosti správce;
- údaje byly získány na základě odvolaného souhlasu, na jehož základě správce nesmí nadále pokračovat se sběrem;
- subjekt vznesl námitku proti zpracování[29];
- správce se dopustil protiprávního zpracování - kupříkladu zpracování osobních údajů bez zákonné výjimky a souhlasu subjektu;
- ke splnění povinnosti vycházející z práva evropského, členského státu;
- nabídky služeb informační společností dětem.

Toto právo ovšem není absolutní a v některých případech může správce žádost odmítnout. Jedná se především o situace, kdy na něj dopadá právní povinnost uchovávat takovéto údaje. Příkladem může být žádost o výmaz veškerých osobních údajů po ukončení pracovního poměru.[25]

Na základě posledního práva může subjekt požádat správce o omezení zpracování, nabyl-li dojmu porušení jakéhokoliv výše uvedeného práva. Jedná se především o zpracování při uplatnění práva na opravu, nebo vznesení námitky proti zpracování. GDPR čl. 18 odst. 1 stanovuje i práva správcům, kdy mohou pozastavit zpracování - jsou-li údaje zpracovávány protiprávně, avšak subjekt odmítá výmaz, či správce data

[28]Tyto údaje však nemusí správce aktivně vyhledávat.

[29]Pokud se účely týkají přímého marketingu, nebo zájmy subjektu převažují nad zájmy správce. Podmínka vychází z GDPR čl. 22.

již za daným účelem nepotřebuje, subjekt je ale požaduje pro určení, výkon, obhajobu právních nároků.[9]

1.4 Cíle nařízení GDPR

Druhá část bakalářské práce je věnována vysvětlení technického postupu při procesu anonymizace, pseudonymizace, kdy v případě pseudonymizace nenastává stav anonymity.[30] Cílem této části je naplnit předpoklad stanovený v GDPR, který nepožaduje dodržení podmínek nařízení pouze v případech, nespadají-li údaje do vymezených hranic pro pojem osobní údaj. Předpoklad se opírá o recitál 26 hovořící o anonymních, anonymizovaných informacích. Ustanovení specifikuje, že anonymní informace (netýkající se subjektu) spolu s anonymizovanými daty (nesplňující charakteristické vlastnosti osobních údajů, tudíž nelze předpokládat spojení se subjektem údajů) již nadále nepodléhají tomuto nařízení. Pod anonymizovanými informacemi jsou myšleny veškeré osobní údaje, na které byl aplikován proces dostatečné anonymizace.[9]

Krok anonymizace představuje nevratný proces, který na základě různých technik postupně navyšuje obecnost informace až do stavu plné anonymizace, kdy zaniká možnost jakéhokoliv propojení s konkrétním subjektem údajů. Dojde tedy k přetrhání veškerých vazeb mezi informací a subjektem. V tomto případě lze o takovéto informaci tvrdit, že již nadále nesplňuje charakteristiky osobního údaje. Při posuzování anonymity informace ovšem správce musí vzít v potaz možnost re-identifikace na základě veřejných, dostupných informací, databází. *„If you imagine that your left hand is anonymized data, your right hand is outside information, and your interleaved fingers are places where information from the left matches the right, this image basically how reidentification is achieved.“*[31] Problematikou „dokonalé" anonymizace se zabývá následující kapitola, která představuje a zhodnocuje různé anonymizační techniky.

Cílem celého procesu je vytvořit informaci, která ani po propojení s dalšími veřejnými informacemi nelze spojit s konkrétním subjektem. Opačným procesem, při němž se „útočník“ snaží propojit takovéto informace, se označuje pojmem re-identifikace. Lze jej chápat jako „útok“ na soukromí subjektu údajů. Smyslem celého „útoku“ je pokus rekonstruovat proces identifikace s anonymizovanými informacemi. Podaří-li se uskutečnit takovéto kroky, „útočník“ nabude automaticky statusu správce.[2]

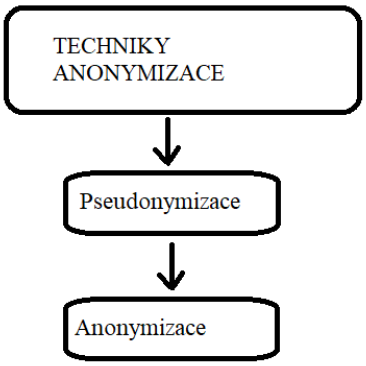
[30] Problematika anonymizace, pseudonymizace je řešena v následující kapitole.

[31] OHM, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 2009, s.1725.[2]

V kontextu GDPR je proces anonymizace zpracováním osobních údajů. Z těchto důvodů je nutno dodržet veškeré podmínky zmíněné v kapitole 1.3 Zpracování osobních údajů. Při porušení podmínek dojde k nezákonnému zpracování a k porušení výše zmíněného nařízení.[26]

2 Techniky anonymizace

Techniky anonymizace lze rozdělit z několika pohledů do různých skupin. Všechny ovšem vycházejí ze záměru odstranění všech osobních údajů (přímých, nepřímých), které by mohly vést k re-identifikaci. Jednotlivé databáze se liší strukturou, charakteristikou, záměrem, citlivostí dat, nelze tedy posoudit jednu vhodnou techniku, která by zajistila dokonalou anonymitu všem subjektům spolu s minimální ztrátou informací. Právě z těchto důvodů jsou techniky rozděleny dle škály anonymizačního výsledku.[27]



Obr. 2.1: Základní rozdělení technik anonymizace dle škály zásahu do informace.

2.1 Pseudonymizace

Pseudonymizace je první představenou metodou zajišťující částečnou ochranu soukromí pouze na základě spojení s přímými identifikátory. Evropská unie v GDPR oddělila pojmy anonymizace, pseudonymizace. Zatím co znění anonymizace je popsáno v předcházející kapitole, pseudonymizace je charakteristická jako vratný proces, zajišťující ochranu subjektům pouze odstraněním všech přímých identifikátorů z dat. Z definice je patrné, že nezajišťuje stav anonymity. Stále jsou zde nepřímé identifikátory, s kterými metoda nepracuje. Z těchto důvodů GDPR upozorňuje správce recitálem 26: *„Zásady ochrany údajů by se měly uplatňovat na všechny informace týkající se identifikované nebo identifikovatelné fyzické osoby. Osobní údaje, na něž byla uplatněna pseudonymizace a jež by mohly být přiřazeny fyzické osobě na základě dodatečných informací, by měly být považovány za informace o identifikovatelné fyzické osobě.“*, že i takto „anonymizované" údaje jsou stále pod ochranou nařízení, neboť obsahují data, která mohou být vztažena ke konkrétnímu subjektu.

Cílem pseudonymizace je tedy snížení rizika re-identifikace při procesu zpracování. Její využití lze nalézt například v diskuzních fórech, kryptoměnách a na sociálních sítích.[9, 27]

2.1.1 Způsoby implementace

Jak již bylo zmíněno, pseudonymizace je postavena pouze na zaměnění přímých identifikátorů pseudonymy – údaji, které pro veřejnost nemají informativní přínos o subjektu. Příkladem je nahrazení jmenných údajů pseudonymem jedinečného čísla, na jehož základě nemohou ostatní uživatelé zjistit jméno, příjmení, iniciály původních dat, tak jak je ukázáno v tabulce níže.[28]

Tab. 2.1: Příklad tabulky s fiktivními údaji v originální podobě.

Jméno	**Rok narození**	**Pohlaví**	**Závislost na cigaretách**
Karel K.	1996	Muž	Ano
Hana Z.	1986	Žena	Ano
Karel K.	1990	Muž	Ne
Hana Z.	2001	Žena	Ano

Veškeré údaje slouží pouze jako příklad, a tak nejsou spojeny s žádnou fyzickou osobou. Tabulka ukazuje případ obsahující kolizi jmen. Jestliže by správce generoval pseudonymy libovolným způsobem, avšak pouze na základě jména, mohlo by zde docházet k duplicitám takto vygenerovaných pseudonymů. GDPR ukládá povinnost předcházet jakýmkoliv duplicitám. Správce proto musí zvolit libovolné, avšak „bezpečné“ způsoby generování. Příkladem „bezpečného“ způsobu generování může být následující popsaný postup. Z důvodu minimalizace možnosti duplicity se veškeré charakteristiky zapojí do procesu generování pseudonymu, a to způsobem, který bude sčítat jejich hodnoty. Uvedený příklad popisuje výpočet pro první subjekt. Z charakteristiky označující jméno je spočítána absolutní hodnota – součet všech písmen ve jméně a příjmení. V konkrétním případě subjekt Karel K. nabývá absolutní hodnoty rovné šesti. K této hodnotě je přičten rok narození (aktuální vypočítaná hodnota = 1996 + 6 = 2002), konstanta pět v případě ženského pohlaví, šest v případě mužského pohlaví (aktuální vypočítaná hodnota = 2002 + 6 = 2008) a číslo 33, je-li subjekt kuřák (aktuální vypočítaná hodnota = 2008 + 33 = 2041). V tomto stavu by však stále mohlo docházet k duplicitám, z tohoto důvodu se k součtu přiřadí náhodně vygenerovaných pět znaků. Obsáhne-li databáze dva identické subjekty, je výrazně snížena možnost vzniku duplicity generovaného pseudonymu, ne

však zcela vyloučena. Při zařazení nového subjektu by databáze měla aktivně vyhledávat pseudonymy vedoucí k duplicitě a opakovat proces generování náhodných znaků. V uvedeném případě je nutné brát v úvahu, že algoritmus výpočtu je uzpůsoben konkrétní tabulce, konkrétnímu počtu a charakteristice subjektů.[29]

Tab. 2.2: Ukázka vygenerovaných pseudonymů vycházející z tabulky 2.1.

Pseudonym	**Rok narození**	**Pohlaví**	**Závislost na cigaretách**
76093	1996	Muž	Ano
44683	1986	Žena	Ano
93206	1990	Muž	Ne
18533	2001	Žena	Ano

Nařízení dále správci stanovuje povinnost uchovávat odděleně dodatečné informace – data, která prošla krokem nahrazení pseudonymy v originální podobě spolu s daty k nim přidruženými. Na ně je následně povinen aplikovat veškerá technická, organizační a bezpečnostní opatření[1], která jsou stanovena v podmínkám GDPR, jelikož se stále jedná o přímé, nepřímé osobní údaje.[9]

Tab. 2.3: Oddělené informace v procesu pseudonymizace tabulky 2.1.

Pseudonym	**Jméno**
76093	Karel K.
44683	Hana Z.
93206	Karel K.
18533	Hana Z.

Pseudonymy jsou generovány různými způsoby, nejčastěji metody využívají hash, šifrování, spojení obou zmíněných a maskování.[29]

Implementace šifrováním

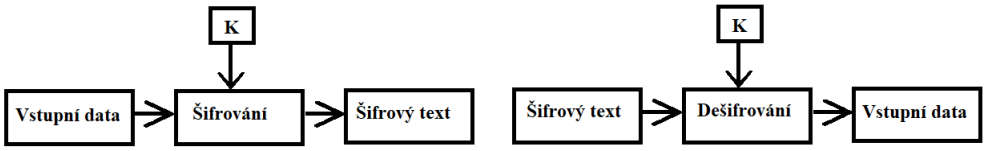
Metoda se zakládá na využití vlastností kryptografie[2], jejímž hlavním cílem je převést data do nečitelné podoby pro všechny uživatele s výjimkou koncových (například příjemce, odesílatel). Příjemce pak na základě ustanovených vlastností dokáže získat původní, originální informace. Nikdo jiný (útočník) by neměl být schopen realizovat popsaný převod. Kryptografické procesy se však ne vždy zakládají na bezpečných

[1] Příkladem může být zajištění přístupu pouze autorizovaným osobám.

[2] Jedná se o vědu zabývající se ochranou dat.

algoritmech, které by byly odolné vůči jakémukoliv útoku. Metoda hledání postupu, jež je schopna data převést bez znalostí, s kterými disponují pouze koncoví uživatelé, na původní informace, se nazývá kryptoanalýza. Příkladem postupu může být případ, kdy odesílatel na základě určitého systému mění jednotlivá písmena ve zprávě za jiná (např. posun o určitý počet v abecedě). Útočník je schopen na základě analýzy zachycené „zašifrované“ zprávy (hledání nejčastěji opakujících se znaků, jazykových schopností, intuice) vytvořit kopii aplikovaného algoritmu a zároveň tak rozluštit pro něj původně nečitelná data.

Z uvedeného příkladu vyplývá, že nejdůležitějším krokem je převedení informace do stavu, který bude schopen skrýt její původní význam (šifrování) natolik, aby nebylo možné realizovat proces dešifrování (získání původních dat) bez stanovených klíčů. Uvedený obrázek níže zobrazuje pouze obecný princip zabezpečení důvěrnosti v oblasti kryptografie.[30]



Obr. 2.2: Příklad obecného procesu šifrování, dešifrování v kryptografii.

Kryptografie je dělena dle počtu použitých vygenerovaných klíčů v komunikaci. Je-li použit jediný klíč pro oba procesy (šifrování, dešifrování), jedná se o symetrickou kryptografii. Hlavní výhodou je především rychlost celého procesu. Typickým zástupcem skupiny jsou šifry s názvem DES, AES. Využije-li se v komunikaci dvou klíčů, jeden pro šifrování, druhý pro dešifrování, jedná se o asymetrickou kryptografii. Proces ustanovení klíčů je pak časově složitější. Typickými zástupci jsou RSA, DSA. Pro asymetrické klíče platí nemožnost odvození jednoho klíče při znalosti klíče druhého.[30]

Využije-li správce pro pseudonymizaci metodu šifrování, stačí aby na identifikátory aplikoval vybraného zástupce libovolné šifry. Data se následně pro ostatní osoby stanou nečitelnými.[29]

Implementace hashem

Funkcionalita této implementace se opět opírá o prvky využité v kryptografii. V tomto případě konkrétněji o prvek zajišťující například autentičnost dokumentů. Je schopen zajistit, aby podepsaný dokument[3] nebylo možné pozměnit (datum, místo vy-

[3]Podepsaným dokumentem je myšlen dokument, ke kterému je připojen elektronický podpis. Elektronický podpis označuje souhrn metod zaručující totožnost, autentičnost odesílatele.

dání, samotná zpráva) bez povšimnutí. Vzhledem k citlivosti využití jsou kladeny na bezpečné hashe čtyři požadavky:

- fixní výstup - jakkoliv velký vstupní blok bude mít vždy fixní velikost výstupu. Je důležité zajistit, aby potencionální útočník nemohl odhadnout rozsah vstupních bloků a tím zjistit v případě pseudonymizace například počet písmen ve jméně;
- vysoká reakce na změny - představuje objem reakce i na malé změny vstupního bloku. Chrání pseudonymy před možností dedukce na základě pravděpodobnosti opakujících se posloupností;
- jednosměrnost - zabraňuje možnosti dešifrování hashové informace. Je zajištěna ochrana proti zpětnému získání vstupních, původních bloků při použití stejné hashové funkce, na již hashované informace;
- odolnost proti kolizím - výsledky jednotlivých vstupních bloků by se měly lišit. Jestliže tato skutečnost není zajištěna, je nutno z bezpečnostních důvodů zvolit jiný typ hashe, nebo zařadit ke vstupnímu bloku i kryptografickou sůl. Pro správce pak platí pravidlo výskytu každého pseudonymu v databázi pouze jednou.

Kryptografická bezpečnost jde uměle navyšovat přidáním solení. Solení označuje operaci připojení náhodně vygenerovaných bitů[4] ke vstupnímu bloku. Tím je zabráněno generování duplicitních hashů na základě vstupního bloku. Dále brání nejčastějším útokům proti dešifrování jako jsou slovníkové[5], Rainbow tables[6] útoky.[30]

Praktická implementace v oblasti pseudonymizace na osobní údaje pak znamená využití libovolného generátoru bezpečného hashe. Díky velkému rozvoji v oblasti kryptografie a technologie (výpočetní kapacita zařízení) již některé známé hashové funkce nejsou zcela bezpečné. Respektive je na ně znám útok, kterým je útočník schopen získat na základě hashe původní informace, aniž by vlastnil vstupní bloky. Tento negativní vývoj zapříčinil vznik organizací, zaměřujících se na kontrolu bezpečnosti hashů a klíčů. Nejznámější americká agentura vydávající standarty se nazývá NIST – National Institute of Standards and Technology.[31]

Spojení obou metod

Spojení obou výše uvedených metod přináší do jisté míry výhody. Správce snižuje riziko prolomení systému pro přidělování pseudonymů. Implementace je realizována dvěma způsoby: výstupní hashe jsou dále šifrovány, vstupní blok hashe je tvořen šifrovaným textem.[32]

[4]Bit představuje nejmenší datově-informační jednotku, nabývá pouze hodnot 0 nebo 1.

[5]Útok založený na prohledávání slovníků, kdy ke všem variantám je vypočítáván hash. Výsledkem stejného hashe mohou být dva rozdílné vstupní bloky.

[6]Rainbow tables představují přepočítané tabulky hashů, opět je hledána shoda.

Maskování

Maskování je poslední představenou možnou pomocnou technikou, jejíž cílem je převést data do pseudonymní podoby. Tentokrát však bez využití pseudonymu.[27]

Tab. 2.4: Ukázka metody maskování „subjektů“ z tabulky 2.1.

Pohlaví	**Závislost na cigaretách**
Muž	Ano
Žena	Ano
Muž	Ne
Žena	Ano

Její princip je postaven na odstranění všech klíčových hodnot vedoucích k identifikaci konkrétního subjektu. Nevýhodou metody je ovšem fakt, že takto upravené databáze neuchovávají žádné pseudonymy, a tak není možné data aktualizovat, měnit nebo přidávat (rozšiřovat zadané informace).[32]

Zhodnocení

Pseudonymizace nezajišťuje anonymitu, přesto je využívanou technikou i pro zveřejňovaná data. Stále zde zůstávají nepřímé identifikátory, které na základě útoku dedukcí mohou být klíčem k prolomení soukromí subjektů. Správce by měl vždy pamatovat na nahrazení všech klíčových vlastností pseudonymy.

Jednotlivé jmenované metody přinášejí jisté výhody. Šifrování nelze prolomit bez znalosti klíče, není tak potřeba ani na základě dešifrování ukládat data odděleně. Hash je pouze jednosměrný, dovoluje spojení několika údajů do jednoho bloku. Vývoj kryptografie však vytváří neustálé riziko prolomení (změny bezpečných velikostí klíčů, typy hashů). Správce by neměl spoléhat pouze na bezpečnost kryptografických prvků, ale zanést do dat určitou náhodnost (šum)[7] a tím navýšit ochranu svých databází.[9]

2.2 Anonymizace

Hranice pro pojem anonymizovaná data GDPR vytyčilo způsobem, aby byla zajištěna úplná ochrana subjektů. V tomto případě je tak nutno odstranit z dat nejen přímé, ale i nepřímé identifikátory a znemožnit zpětnou identifikaci. Ačkoliv správce zvolí dostatečně silné prvky anonymizace, stále zde ovšem zůstává možnost útoku na

[7]Metoda přidání šumu je blíže specifikována v kapitole 2.3 Přidání šumu.

základě dedukce - útočník je uvědomen, že vybraný subjekt se stal součástí zveřejněné databáze. Následně na základě svých znalostí jej dokáže re-identifikovat. Právě těmto útokům má zabránit k-anonymita.[9, 33]

2.2.1 K-anonymita

K-anonymita je postavena na myšlence skupinové ochrany anonymizovaných údajů. Jinak řečeno, pokud správce chce využít k-anonymity, musí zajistit v databázi identickou posloupnost údajů s četností k, platí zde podmínka k > jedna (dva a více), jak ukazuje příklad níže.[34]

Tab. 2.5: Ukázka k-anonymity na „subjekty“ z tabulky 2.1.

Rok narození	**Pohlaví**	**Závislost na cigaretách**
1990 - 1999	Muž	xx
1980 - 2009	Žena	Ano
1990 - 1999	Muž	xx
1980 - 2009	Žena	Ano

Jestliže by hodnota k (frekvence identických záznamů) byla rovna číslu jedna, tak na základě vzorce pro riziko re-identifikace[33]:

$$Riziko = \frac{1}{k} \tag{2.1}$$

lze vyvodit riziko rovno jedné, což odpovídá sto procentům. K re-identifikaci by útočníkovi stačila pouze základní znalost osobních údajů, protože žádný jiný záznam by těmto charakteristikám neodpovídal. V praxi se běžně doporučuje minimální hodnota k: 2, 3, 5, 10, maximálně 15. Rizikový práh pro vypsané hodnoty se pak přibližuje k: 50, 33, 20, 10, 7 procentům. Příklad uvádí výpočet rizika re-identifikace pro uvedenou tabulku výše (2.5):

$$Riziko = \frac{1}{k} = \frac{1}{2} = 0,5 \tag{2.2}$$

Z výše uvedeného je patrné, že se zvyšujícím se číslem k klesá riziko re-identifikace. Jestliže hodnota k nenabývá minimálně čísla dva (případ tabulky 2.1), je nutno využít technik, které jsou schopny tuto skutečnost zajistit - generalizace, suprese, avšak na základě určité ztráty dat.[33]

Generalizace

Generalizace se vyznačuje zobecňováním jednotlivých osobních údajů. Při využití metody se snižuje celkový informativní přínos – čím více jsou data obecnější, tím

méně charakterizují daný subjekt. Naopak klesá riziko re-identifikace. Platí zde dvě pravidla:

1. Každý prvek je zastoupen v generalizované doméně.
2. Všechny hodnoty v dané doméně jsou převedeny maximálně na jednu vyšší doménu (obecnější).

Uvedené pravidla jsou zobrazena na obrázku 2.3 B, který vychází z tabulky níže. Z takto generalizovaných domén můžeme postupně vykreslit doménovou generalizační hierarchii DGHD[8]. Při tom musí být zajištěna kompatibilita mezi DGHD a originální tabulkou prvků označovanou často PII (neanonymizovaná tabulka s identifikátory). Dále je nutno zajistit v hierarchii koordinované, logické uspořádání.[35]

Tab. 2.6: Tabulka fiktivních údajů pro příklad generalizace, suprese.

Jméno	**Datum narození**	**Pohlaví**	**PSČ**	**Počet dětí**
Marcela H.	3.9.1990	Žena	61500	1
Petr Z.	27.4.1995	Muž	60300	3
Barbora V.	6.12.1998	Žena	61400	2
Alena D.	16.5.2000	Žena	63500	1
Zdeněk W.	31.12.1988	Muž	61500	8

Výše uvedená tabulka má celkem pět identifikátorů, na základě kterých je možné subjekty re-identifikovat. Obsahuje také informace, které by neměly být v žádném případě zveřejněny pospolu v originální podobě. Jak dokázala L. Sweeney ve své práci: Simple Demographics Often Identify People Uniquely, kterou prováděla na území USA - pohlaví, datum narození, spolu s poštovním směrovacím číslem, dokáží identifikovat přes 87 procent americké populace. Tato práce se vztahuje k sčítání lidu v roce 1990. Počet obyvatel USA v té době činil 248 miliónů, což vypovídá, že L. Sweeney dokázala identifikovat přes 216 milionů obyvatel na základě tří údajů.[36]

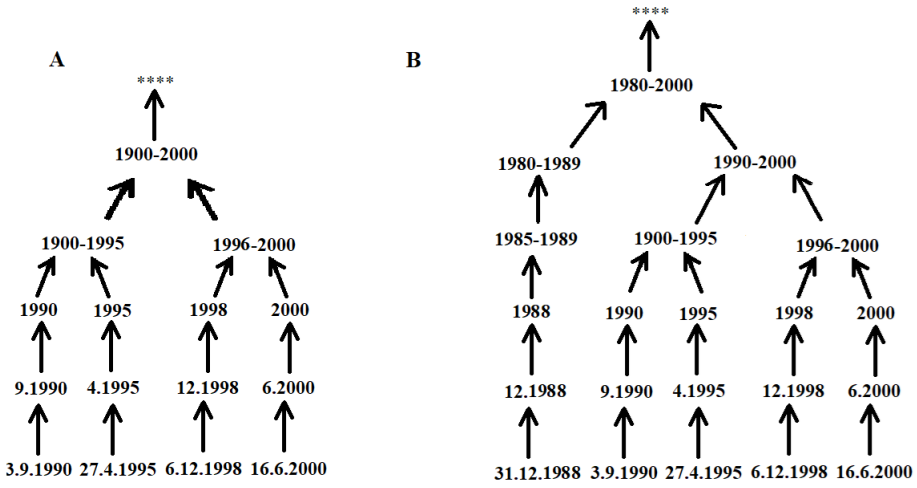
Z těchto důvodů je nutno data generalizovat takovým způsobem, aby byla zachována určitá informační hodnota, avšak nebylo možné na základě jejich zveřejnění identifikovat daný subjekt. K nalezení bilance mezi těmito stranami slouží DGHD vykreslující bijektivní zobrazení[9] PII a jeho postupnou generalizaci. Pro vykreslení DGHD platí pravidla:

1. Absolutní hodnota PII musí být rovna absolutní hodnotě generalizované tabulky. V tomto kontextu platí, že na základě generalizace nemohou být vyjmuta jakákoliv data.

[8]V originále domain generalization hierarchy.

[9]Bijektivní zobrazení lze charakterizovat vlastností, která zaručuje, že každý prvek z prvního zobrazení je přiřazen právě jednomu prvku v druhém zobrazení.

2. Pro jednotlivé stupně generalizace g musí platit: g + 1 má vždy vyšší stupeň anonymizace než stupeň g.



Obr. 2.3: Postup generalizace data narození vycházející z tabulky 2.6.

Hierarchie DGHD je charakteristická vlastnostmi - nejnižší prvek je v originální podobě, obsahuje neanonymizovaný identifikátor a zároveň má nejvyšší informativní hodnotu z celé hierarchie. Nejvyšší stupeň pak zobrazuje nejobecnější tvar identifikátoru, jak je ukázáno na příkladu výše. Zároveň je natolik obecný, že nemá téměř žádný informativní přínos. Z tohoto důvodu je nutno vybrat do procesu anonymizace stupně mezi těmito dvěma body splňující dostatečnou, ovšem ne přílišnou anonymizaci.[33]

Jednotlivé logické kroky jsou voleny způsobem, aby zapříčinily vznik stavu k-anonymity. Za bezpečnou úroveň generalizace se bere stav, kdy každý prvek je zastoupen prvkem obecnějším, pro který pak platí nemožnost určení původních dat a zároveň skladba minimálně dvou prvků[10]. V uvedeném příkladu HDGH B se za bezpečnou úroveň považuje pátá spolu s šestou. Úrovně nula až čtyři nesplňují podmínku skladby minimálně dvou prvků, tím pádem nezaručují k-anonymitu.[35]

Podobným způsobem lze generalizovat:

- PSČ - každá úroveň generalizace nahrazuje dané číslo za nulu. Pozice jsou postupně vybírány od konce zprava doleva. Jestliže jsou všechny pozice nahrazeny nulou, posledním bodem generalizace je nahrazení čísel neurčitými znaky (kupříkladu pomlčkami).
- Pohlaví: hierarchie má pouze dva stupně generalizace, a to na neurčeno následně na vybrané znaky[11].

[10] Záleží na zvolené konstantě k.

[11] Detailní popis popsaného postupu lze nalézt v kapitole 3.2.3 Třetí fáze běhu.

Generalizace výše je provedena dvěma způsoby s odlišnými nejvyššími body hierarchie. Příklad A vynechává subjekt s jménem Zdeněk W. pro přílišnou odlišnost data narození. Jestliže by bylo toto datum připočítáváno i nadále, může při nesprávné volbě poměru metod dojít k celkovému poklesu informativní hodnoty. L. Sweeney ve své práci Achieving k-Anonymity Privacy Protection Using Generalization and Suppression uvádí postup výpočtu pro metriku přesnosti dat[34]:

$$Prec(RT) = 1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{h}{|DGH_{Ai}|}}{|\mathbf{PT}| \cdot |\mathbf{N_A}|} \tag{2.3}$$

Výsledkem výpočtu této metriky jsou procenta dat, která nebyla ztracena krokem zobecnění. Význam jednotlivých značek je uveden na konci práce (Seznam symbolů, veličin a zkratek). K celkové ztrátě dat však lze dojít na základě jednodušší matematiky bez použití výše uvedeného vzorce. Pro zjednodušení celého příkladu bude tabulka složena pouze z roku narození. V případě příkladu 2.3 B by byl zůstatek informací vypočítán[34]:

- Procentuální význam libovolné charakteristiky je roven podílu celkového informativního přínosu původní tabulky PII (vždy 100 procent), počtem druhů přímých identifikátorů:
$$Význam = \frac{1}{1} = 1 \tag{2.4}$$
- Výpočet ztráty informace na základě jednoho stupně vybrané generalizace v hierarchii DGHD je proveden podílem procentuálního významu libovolné charakteristiky vypočítané v předchozím kroku a celkovým počtem stupňů v hierarchii:
$$Ztráta_{dílčí} = \frac{1}{7} = 0,14 \tag{2.5}$$
- Následně je nutno zvolit bezpečný stupeň anonymizace, v tomto případě pátý stupeň.
- Celková ztráta informací je rovna násobku ztráty informace na základě jednoho stupně vybrané generalizace a součtu všech stupňů generalizace, které budou ztraceny (je zde připočítán i nultý stupeň):
$$Ztráta = 0,14 \cdot 5 = 0,7 \tag{2.6}$$

  Informativní přínos lze pak získat odečtením celkové ztráty informace od čísla jedna:
$$Přínos = 1 - 0,7 = 0,3 \tag{2.7}$$

V uvedeném příkladu datům zůstává přesnost 30 procent. Ztráta informační hodnoty činí 70 procent. Pokud by tabulka obsahovala další subjekt s rokem narození v intervalu 1980 až 1989, existovala by zde možná generalizace těchto subjektů v daném intervalu. Míra přesnost by nabyla 57 procent. Aby byly informační ztráty

co nejvíce potlačeny, je pro data dobré přepočítávat informační hodnoty s různými generalizačními postupy a pokusit se tak vybalancovat kontrast přesnosti informací spolu s ochranou údajů.[37]

Obrázek 2.3 A ukazuje druhý způsob generalizace, kdy je využita i metoda zvaná suprese.

Suprese

Suprese, též známá jako potlačení, je využívanou pomocnou metodou pro zajištění stavu k-anonymity. Její princip spočívá v potlačení (odstranění) dat, která jsou příliš odlišná od ostatních skupin až do takové míry, že by mohla vést k identifikaci subjektu (při nedostatečných krocích anonymizace) či k přílišnému stupni generalizace. Je využívána především v případech, kdy generalizace sama o sobě není schopna zajistit „dokonalou“ anonymizaci s udržením určité míry informačního přínosu, tak jak je ukázáno na příkladu 2.3 B.[34]

Zda bude metoda využita, je nutné posuzovat z hlediska co nejmenší informační ztráty na základě posouzení váhy informace. Váhou informace je v tomto kontextu myšlen celkový informativní přínos pro databázi. Je-li subjekt v nějaké charakteristice příliš odlišen od ostatních, ovlivňuje výrazněji celkový výstup databáze než podobné prvky charakteristik jednotlivých subjektů. Odstraněním vybraného subjektu databáze ztratí nejen významný prvek měření, ale vznikne zde i určitá ztráta ovlivňující výpočet celkového přínosu informací.[33]

V uvedené generalizační hierarchii (obrázek 2.3) je tato metoda využita v případu A, odstraněním subjektu s datem narození 31.12.1988. Jak vychází z hierarchie DGHD, není nutno již použít pátý stupeň anonymizace, ale postačí čtvrtý, čímž se navýší informativní přínos zbylých prvků. Výpočet celkové informační ztráty dle postupu uvedeného výše pro tabulku obsahující pouze data narození nabude ztráty 48 procent. Informační přínos bude roven 52 procentům. Výsledky výpočtů však nejsou konečné. K celkové ztrátě je nutno přičíst ztrátu informace (data 31.12.1998). Uvedený příklad níže ukazuje výpočet ztráty[34]:

- Ztráta jedné informace je dána podílem procentuálního významu libovolné charakteristiky (vysvětleno v předchozím příkladě pro výpočet ztráty informací při využití generalizace), celkovým počtem negeneralizovaných informací (počet informací v nultém stupni generalizace):

$$Význam_{informace} = \frac{1}{5} = 0,2 \tag{2.8}$$

- Celková ztráta suprese je dána násobkem ztráty informace a počtu subjektů, které budou odstraněny:

$$Ztráta_{supresi} = 0,2 \cdot 1 = 0,2 \tag{2.9}$$

- Konečná ztráta je dána součtem ztráty způsobené supresí, ztráty způsobené generalizací (postup uveden výše):

$$Ztráta_{konečná} = 0,2 + 0,5 = 0,7 \tag{2.10}$$

Informativní přínos lze pak získat odečtením celkové ztráty informace od čísla jedna.

$$Přínos = 1 - 0,7 = 0,3 \tag{2.11}$$

Z výše uvedeného vyplývá, že suprese není vždy ideálním řeším pro data, která nejsou odlišná ve všech parametrech PII. Pro subjekty, jejichž data lze generalizovat v nízkých úrovních domény, je vhodné využít metodu generalizace a zajistit tak menší informativní ztrátu.[35]

Zhodnocení

K-anonymita je často využívanou metodou, jak zajistit bezpečí subjektům proti zpětné identifikaci. Dojde-li ke shodě neanonymizovaných údajů spolu s anonymizovanými, vždy těmto identifikátorům budou odpovídat nejméně dva prvky anonymního vyobrazení. Obě techniky, kterými lze daného stavu docílit, přinášejí jisté přínosy. Suprese ovlivňuje pouze jedinou n-tici prvků, avšak generalizace pozmění hodnoty všech prvků v dané skupině. Důležitým krokem je zvážení míry informační ztráty a následné posouzení jednotlivých kroků.

Při výběru informací do anonymizačního procesu je nutné uvážit i jednotlivé váhy informací. Nejčastější útoky dedukcí právě začínají na výjimečných (odlišných) subjektech a snaží se odhalit jejich identitu. Kupříkladu osoby trpící vzácným onemocněním bude snadnější re-identifikovat, než subjekty, které onemocněly v posledním měsíci chřipkou. Jestliže se podaří prolomit způsob generalizace, bude i zároveň poodhalena identita všech subjektů. Mezi časté důvody re-identifikace patří:

- neposouzení všech možných ať už přímých, tak i nepřímých identifikátorů
- zvolení nízké hodnoty k
- nevyrovnání váhy informací

Právě tyto chyby spojené s nevyrovnáním váhy informací lze potlačit metodou přidání šumu zmíněnou v kapitole 2.3.

2.2.2 Další metody anonymizace

Výše zmíněné metody však nejsou jedinými způsoby vedoucími k zajištění ať samotné či skupinové anonymizace. Existují další nástavby vycházející z principu k-anonymity. Přestože se k-anonymita může jevit jako zcela „bezchybnou“, postupem času byly vytvořeny různé útoky na prolomení její bezpečnosti.

Homogenní útok, v originále Homogeneity attack, se zakládá na nedostatečném posouzení rizika re-identifikace. Pokud určitý počet prvků sdílí stejnou posloupnost identifikátorů a zároveň obsahují stejné dodatečné informace, lze na základě znalosti identifikátorů provést útok na soukromí vybraného subjektu. Dodatečnými informacemi v tomto případě mohou být myšleny onemocnění, dluhy, jiné citlivé osobní údaje.[38]

Znalostní útok na pozadí, v originále Background Knowledge Attack, naopak zakládá na dedukci a znalosti vedlejších informací. Jestliže je opět zvolena nízká úroveň výskytu dodatečných informací, lze na základě dedukce vyloučit ostatní možnosti a narušit soukromí subjektu. Kupříkladu pokud by v nemocniční databázi hledaný subjekt odpovídal několika zobecněným záznamům, ale dodatečné informace by byly pouze dvojího druhu, na základě znalosti vedlejších informací (negativní test na první onemocnění, nízké procento výskytu nemocí u různých národů) lze subjekt re-identifikovat a dedukovat, jaké citlivé informace se vztahují k vybranému subjektu.[38]

Tab. 2.7: Příklad tabulky splňující podmínky k-anonymity.

Datum narození	**Pohlaví**	**Onemocnění**
1980-2000	Žena	Žloutenka typu C
1940-1959	Muž	Žloutenka typu B
1980-2000	Žena	Zápal plic
1980-2000	Žena	Žloutenka typu B
1940-1959	Muž	Žloutenka typu A

Právě díky těmto útokům byla vytvořena nástavba l-diversita zakládající se na myšlence dostatečné různorodosti citlivých, dodatečných informací u identických zobecněných posloupností atributů. Aby byla zajištěna dostatečná úroveň ochrany před zpětnou identifikací, je nutno při volení rozmanitosti těchto informací brát v potaz i veškeré identifikátory z originální tabulky prvků PII a vyvodit tak potřebnou entropii. Nelze tedy stanovit konstanty l, které by zaručily stav l-diverzity ve všech možných případech. V ideálním případě by každý prvek z této kolekce měl mít vlastní dodatečné informace.

Similarity attack je útok, jehož cílem je prolomit vytvořené bezpečí pomocí l-diverzity. Vychází z nedostatečné entropie citlivých informací - konkrétněji jejich spojitosti. Pokud budou všechny atributy spojeny s jedním zastupujícím problémem (žaludeční, plicní problémy), lze vyvodit, že hledaný subjekt má návaznost s uvedenými problémy. Z uvedené tabulky výše tak vyplývá, že ženy trpí obecným infekčním onemocněním, zatímco muži běžnějším onemocněním.[38]

Tab. 2.8: Příklad tabulky splňující podmínky l-diversity.

Datum narození	Pohlaví	Onemocnění
1980-2000	Žena	Žloutenka typu A
1940-1959	Muž	Chřipka
1980-2000	Žena	Žloutenka typu B
1980-2000	Žena	Černý kašel
1940-1959	Muž	Angína

Poslední představenou metodou řešící nedostatky tentokrát l-diverzity je t-blízkost redukující přílišnou granulitu a snižující dopady l-diversity na data. Využívá prahové hodnoty t, s níž následně poměřuje vzdálenosti v hierarchii domén jednotlivých tříd dodatečných informací s atributy v celé tabulce. Platí pak pravidlo - vzdálenost nesmí být větší než vybrané t.[39]

2.3 Přidání šumu

Šum je využívanou pomocnou metodou pro zajištění stavu anonymity. Funguje na principu zanesení určitých náhodných chyb (šumu) do údajů, které pak zapříčiní, že data nemohou být spojena s konkrétním subjektem, jelikož identifikátory v anonymizovaném zobrazení neodpovídají skutečným identifikátorům vybraného subjektu. Cílem zanesení chyb je převést Poissonovo rozdělení[12] na normální rozdělení[13] a ovlivnit tak skutečné váhy dat. Jestliže je nějaký prvek příliš odchýlen od ostatních, stává se prvkem vzácnějším, nabývá vyšší váhy, ovlivňuje významně celý výstup databáze.[40]

Přidávaná chyba je dělena podle počtu ovlivnění dalších prvků na korelovanou, nekorelovanou.

Korelovaný šum pracuje na principu matic, zachovává skutečnou úměrnost (závislost jednotlivých prvků) mezi maticí zanášených chyb a maticí původních údajů. Tento typ je využíván případy, kdy je nutné zanést do dat šum u více prvků, ale zároveň udržet jejich korelační koeficient[14], jak zobrazuje rovnice níže:

$$\sum{}_z = (1 + \alpha) \sum \tag{2.12}$$

Jinak řečeno, celková hodnota matice s přidanými chybami vychází ze součtu hodnoty původní matice spolu s hodnotou matice přidávaného šumu α. Velikost zkres-

[12]Počet dějů ve vybraném intervalu.

[13]Spojité zobrazení veličin náhodných veličin, které od sebe nejsou přílišně odchýleny.

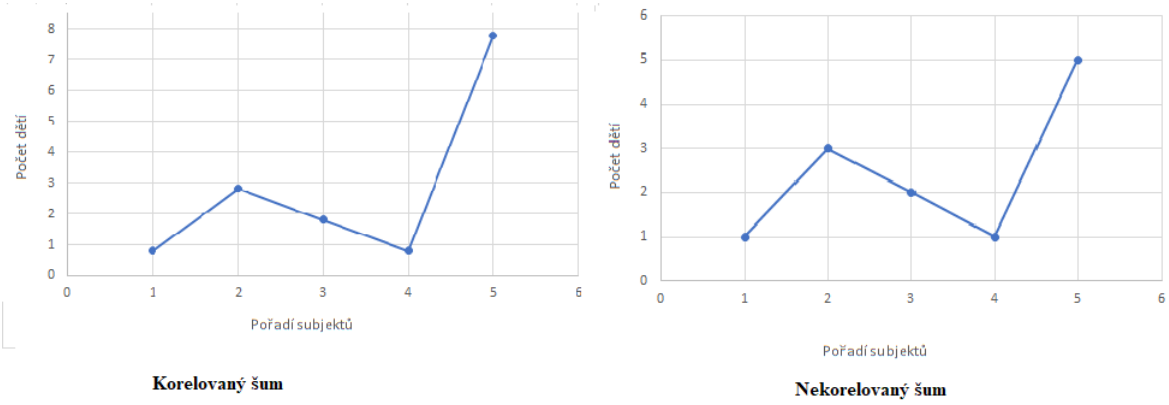
[14]Nabývá hodnot jedna - označuje přímou závislost, až mínus jedna - značí nepřímou závislost.

lení je nutno volit z určitého rozumného intervalu, který přílišně negativně neovlivní výstupy měření.

Druhým typem přidávaných chyb je nekorelovaný šum nezachovávající tentokrát korelační vztahy, pouze kovariance - míru lineární závislosti[15], je charakteristický rovnicí:

$$z_j = x_j + \epsilon_j \tag{2.13}$$

Vždy ovlivňuje pouze prvek, ke kterému se vztahuje. Jestliže tento prvek má lineární závislost s jiným prvkem, bude ovlivněna i jeho hodnota.[41]



Obr. 2.4: Zanesení šumu vycházející z tabulky 2.6.

Pokud by byl zvolen náhodný šum v nedostatečné míře, nezpůsobil by dostatečný efekt na příliš odchýlená data, tak jak je ukázáno v uvedeném grafu výše (korelovaný šum) konkrétně u osoby s číslem pět. Na takto vychýlené body je nutno aplikovat vyšší šum a přiblížit se k normálnímu rozdělení. Naopak z druhého příkladu (nekorelovaný šum) je patrné, že se zvyšujícím se šumem roste významně i ztráta informací. Řešení zanesení chyb do dat není nejbezpečnější volbou, jak odstranit přílišně odchýlená data. Ovšem výsledkem je snížení procent možnosti re-identifikace. Je tedy dobré volit tento postup až na anonymizovaná data, kdy je již zajištěna přibližná rovnováha prvků.[40]

[15]Představuje skutečnost podobnosti prvků, kdy jeden prvek lze vyjádřit pomocí ostatních.

3 Realizace metody anonymizace pro vybranou datovou sadu

Praktickou část bakalářské práce dle zadání tvoří výběr, návrh metodiky pro převedení přímých, nepřímých identifikátorů na informace, které nemohou vést k identifikaci subjektu[1] a následná realizace v libovolném programovacím jazyce. Výsledným výstupem kapitoly je dokumentace programu s návrhem procesu splňujícího požadavky anonymizace.

3.1 Výběr datové sady

Prvním nejdůležitějším bodem celého procesu sběru a zpracování je vybrání vhodných, avšak na druhou stranu i přínosných dat pro vytvoření databáze. Z podmínek stanovených GDPR vychází, že správce vždy, a to bez výjimek, musí stanovit účel vytvořené databáze. Použije-li tento program, musí sám dohlédnout na podmínky vycházející ze zákonného zpracování. Program nekontroluje, zda data byla získána v zákonném intervalu. Jeho cílem je pouze aplikovat na vstupní data princip k-anonymity[2]. Tím bude dosaženo podmínek uvedených v kapitole 1.4 Cíle nařízení GDPR a nově vytvořená data nebudou pod ochranou nařízení.

Aby proces anonymizace byl co nejvíce obecný – nevznikl problém uzpůsobení pro jednu konkrétní datovou sadu, nevychází z této práce žádný datový výzkum. Pro hodnotné výsledky je nutno do programu vybrat vlastní data, která splní níže definovanou strukturu.

Nebudou-li splněny všechny charakteristiky skupiny dat – nenaplnění všech dobrovolných sloupců údajů, program i přesto vyhodnotí výsledky a vytvoří k-anonymitu s minimální konstantou k rovno dvěma[3]. Je zde však podmínka naplnění celého sloupce při zadání jediného údaje v dané dobrovolné charakteristice (sloupci). Jestliže bude nějaký údaj zadán chybně (data budou v jiném tvaru, než program očekává), nebo se uživatel pokusí narušit normální funkcionalitu záměrnou chybou v datech, z důvodu bezpečnosti bude na tento fakt upozorněn a program dále nebude pokračovat ve své činnosti, dokud daná chyba nebude odstraněna. Jestliže by tato možnost nebyla žádným způsobem řešena, nelze předpovídat reakci softwaru na takovouto událost.

Nastane-li jakýkoliv případ z výše zmíněných, uživatel musí spustit celý program od znovu a dovolit opětovnou kontrolu všech dat. Cyklus bude kolovat do doby, než

[1] Tyto informace by však stále měly nést určitou informační hodnotu, je-li to možné.

[2] Detailně popsáno v kapitole 2 Techniky anonymizace

[3] Povinné údaje musí být vyplněny všechny, jinak program ukončí svůj běh.

všechna data budou zadána v očekávaném formátu.

Vybranou datovou sadu tvoří:

- Povinné údaje:
  - Pohlaví – kde "Male", "male", označuje mužské pohlaví, "Female", "female", pak značí ženské pohlaví. Označení jsou vybírány z důvodu jejich univerzálnosti ve světě.
  - Datum narození – z data je nutno vybrat pouze rok narození. Program očekává tuto hodnotu v číselné formě, nejstarším možným zadáním je rok 1900. Hranice pro nejvyšší rok nejsou žádným způsobem nastaveny.
  - Národnost – aby program mohl zpracovávat data, která nepatří pouze občanům České republiky, využít jej k anonymizaci dat napříč různými národy, je nutno vyplnit kolonky různými údaji. Maximální počet národností v databázi není stanoven. Program na vstupu očekává dva nebo tři libovolné znaky[4]. Jestliže uživatel zadá pouze jeden požadovaný znak, bude tato skutečnost vyhodnocena chybně.
  - PSČ – z důvodu různých délek směrových čísel ve světě není program omezen na konkrétní PSČ, ale je uzpůsoben, aby postupně generalizoval jakékoliv směrovací číslo ve velikosti pět (pět číselných znaků).
- Dobrovolné:
  - BMI – pro co nejvyšší obecnost dat, nejširší možnosti využití, je anonymizátor uzpůsoben i k vyhodnocení BMI[5]. Bude-li zadáno, musí být zaokrouhleno na jedno desetinné číslo. Pro jednoduchost hranice nejsou uzpůsobeny různým věkovým skupinám. Dále nerozlišuje různé stupně podvýživy, nadváhy, obezity, všechny tyto případy zaokrouhlí do příslušných intervalů podvýživa, nadváha.
  - Zdravotní stav – údaj je omezen celkovou délkou 20 libovolných znaků (mezery, pomlčky nejsou považovány za chybu, ale jsou započítány do celkové velikosti řetězce).

Všechny charakteristiky uvedené výše byly vybrány pro praktické využití na již existující databáze. Program je přednostně uzpůsoben k anonymizaci menších skupin jedinců – různých sportovních družstev, kroužků, ale zároveň je schopen anonymizovat registr obyvatel, nemocniční registr[6], míru obezity napříč pohlaví, věkem, PSČ či národností. Cílem návrhu je usnadnění zveřejnění, poskytnutí dat

[4]Nekontroluje však korektnost zadání, tím pádem národnosti mohou být vyjádřeny jakýmikoliv znaky.

[5]V celém znění index tělesné hmotnosti je identifikátor umožňující porovnání různých subjektů pouze na základě výšky, váhy. Jiné údaje neovlivňují výsledek.

[6]Správce musí sám posoudit zvýšení možnosti re-identifikace zveřejněním citlivého osobního údaje. Jde-li subjekt na základě tohoto údaje identifikovat, nejedná se o anonymizovaná data.

třetím stranám, další zpracování[7], aby byl naplněn předpoklad GDPR a nedopadala na data nutnost aplikovat podmínky vycházející z tohoto nařízení.

3.2 Návrh funkcionality systému

Podkapitola detailně popisuje jednotlivé fáze běhu anonymizátoru, kterými data postupně prochází.

3.2.1 Všeobecné informace, první fáze běhu

K programu bude připojen soubor s instrukcemi společně s předpřipraveným dokumentem ve formátu excel s koncovkou xlsx[8], jedná se o běžnou koncovku vytvořeného souboru z balíčku Microsoft Office[9]. Formát byl vybrán z důvodu přehlednosti spolu s hojným zástupem alternativních programů, které jsou schopny pracovat se soubory končícími požadovanou koncovkou. Příkladem může být software LibreOffice vyvíjený nadací Document Foundation[10], která podobný kancelářský balíček poskytuje bezplatně. Jeho výhoda spočívá v možnosti spuštění na různých platformách nejen na Windows, ale i Mac, Linux.

Připojený soubor instrukcí má informativní charakter. Jeho cílem je seznámit běžného uživatele s celou funkcionalitou programu. Součástí instrukcí je i část, která informuje uživatele před spuštěním aplikace o nutnosti uložit upravený dokument do koncovky txt[11]. Tento postup je volen z důvodu, aby v případech, kdy se uživateli z nějakého důvodu nepodaří splnit podmínky plynoucí z předchozího odstavce, mohl i přesto využít tento anonymizátor.

Z důvodu usnadnění komunikace mezi uživatelem a programem, není nutno, aby uživatel vytvářel vlastní soubor a kontroloval, zda přepsaná charakteristika odpovídá přesně danému pořadí, které program na vstupu očekává. Mohlo by tak následně dojít k nedorozumění, které by program identifikoval jako chybu nedodržení předepsaného formátu dat. Soubor již bude obsahovat předem seřazenou posloupnost charakteristiky ve tvaru:

1. Národnost
2. Pohlaví

[7] Poskytnou-li se data v neanonymizované podobě třetí straně, vzniká nový správce. Poskytnutí však anonymizovaných informací není nikterak omezeno.[2]

[8] Jedná se o specifikaci Office Open XML (souborový formát) pro aplikaci Microsoft Excel ve verzi Microsoft Office balíčku 2007.

[9] Balíček kancelářského softwaru vyvíjený společností Microsoft.

[10] Nadace se proslavila myšlenkou vývoje svobodného softwaru, který není nikterak omezen úplatou.[42]

[11] Jedná se o běžnou koncovku textového souboru.

3. Datum narození
4. PSČ
5. BMI
6. Zdravotní stav

Jestliže by vyhodnocení, do jaké charakteristiky údaj patří, měl mít na starost program, mohlo by zde docházet k záměnám a výsledky by tak nemusely naplňovat očekávanou funkcionalitu. Příkladem takového chování může být záměna zdravotního stavu spolu s národností, kdy oba řetězce znaků splní podmínku národnosti uvedenou výše[12]. Jestliže by algoritmus rozhodl o záměně těchto údajů, mělo by to kritické dopady na výsledky celého procesu a data by již nemusela být anonymní.

Avšak aby uživatel nebyl limitován, kam se data uloží, či chtěl použít z jakéhokoliv důvodu jiný soubor, bude na začátku programu vyzván k zadání cesty ke složce, kde se očekávané soubory nachází. V případě, že jej program nebude schopen namapovat, vrátí uživateli popis chyby spolu s instrukcemi, které povedou k jeho nahrazení. Je nutné, aby název (pojmenování), spolu s koncovkou (.txt), odpovídal instrukcím v manuálu. Program bude uzpůsoben k hledání konkrétního souboru. V případě nedodržení manuálu nebude schopen zajistit krok načtení dat.

Podaří-li se úspěšné krok namapování, proběhne detailní kontrola formátu dat popsaná výše. Narazí-li na jakoukoliv chybu, do předpřipravené kolekce[13] se nahrají příslušné zadané informace o chybovém subjektu, které správce nahrál do vstupního souboru. Algoritmus vyhodnocuje správnost údajů postupně. Jestliže se zadaný údaj liší ve dvou a více charakteristikách (má dvě a více chyb), bude správce nejprve v prvním běhu upozorněn na první chybu a až při nové kontrole bude upozorněn na další v pořadí. Po dokončení popsaného kroku algoritmus přechází do druhé fáze, avšak pouze v případech, kdy chybová kolekce v sobě nemá nahrány jakékoliv informace. V opačné situaci program ukončí svůj běh a vypíše uživateli veškeré chybové údaje s popisem, který charakterizuje konkrétní problém (důvod, proč je údaj vyhodnocen chybně).

3.2.2 Druhá fáze běhu

Druhá fáze spočívá v samotné anonymizaci dat. Z důvodu existence možnosti netotožných údajů je nutno data oddělit do podobných podskupin. Následně v každé konkrétní podskupině vytvořit anonymní stav pomocí k-anonymity. Praktické řešení využije řetězec[14], do něhož se z vybraných řádků přiřadí charakteristiky subjektů. Pro oddělení jednotlivých konkrétních údajů je řetězec rozkouskován na podřetězce,

[12]Podmínka spočívá v maximální velikosti řetězce rovnu délce dva, nebo tři znaky.
[13]Kolekce odkazuje na místo v paměti, které program vytyčil pro konkrétní datovou sadu.
[14]Jedná se o textový datový typ.

s kterými program dále pracuje[15]. Řádky jsou postupně vybírány od druhého[16] dokud nejsou projity všechny řádky vybraného souboru. Narazí-li se na řádek, jehož celková velikost je rovná nule, bude dle instrukcí ignorován. V této chvíli program do své paměti nahrál všechna příslušná data, která jsou předána dále do procesu anonymizace. Při vkládání malých počtů informací v řádech jednotek je nutné brát v potaz, že program nemusí vždy najít vhodné řešení anonymizace pro příliš odchýlená data. Je uzpůsoben k vytvoření skupinové anonymizace alespoň v řádech desítek informací.

Anonymizace národnosti, pohlaví

Základním stavebním kamenem pro zajištění k-anonymity je počet identických údajů v jednotlivých podskupinách s minimální velikostí výskytu rovnající se číslu dvě. Prvním takto kontrolovaným údajem je charakteristika národnosti. Algoritmus efektivně prochází jednotlivé druhy národností a s každým nalezeným identickým údajem shodujícím se právě v údaji národnosti, je k číselné proměnné přičtena hodnota jedna. Z důvodu využití množiny na porovnávání není možné, aby jedna národnost prošla tímto algoritmem dvakrát. Množina je charakteristická právě výskytem pouze originálních prvků – nedovoluje tedy žádné duplicity. Je-li výše uvedená proměnná rovna číslu jedna, znamená to, že daná národnost nesplňuje podmínku k-anonymity. Z těchto důvodů budou dočasně všechny informace o takových subjektech přesunuty do chybové kolekce sloužící pro ukládání řetězců, které nezapadají do zbývajících dat. K rozlišení, v jakém parametru se objekt liší, je využita knihovna[17]. Její hlavní výhodou je schopnost uchovávat dvojici dat – klíč, hodnota. Klíč slouží k přístupu ke konkrétním hodnotám. V tomto případě klíči budou: národnost, datum narození, pohlaví. Aby objekt již nadále nemohl ovlivnit probíhající anonymizaci, je vymazán z původní kolekce všech objektů[18].

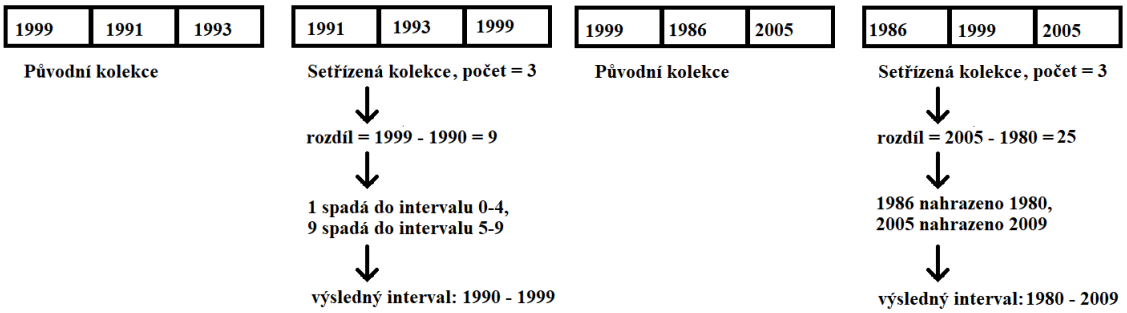
Následuje kontrola četnosti pohlaví příslušných vytvořených podskupin na základě předchozího kroku. Jednotlivé charakteristiky nesmí být anonymizovány bez návaznosti na další. Takovýto krok by přetrhal veškeré vazby a algoritmus by nebyl schopný naplnit podmínku anonymity. Nalezne-li se chyba se souvislostí opět na četnost, aplikuje se identický postup uvedený výše. Hodnoty jsou přidány pod chybový klíč – pohlaví.

[15]Tento krok zajistí možnost přístupu k jednotlivým konkrétním údajům.
[16]V prvním řádku je popis jednotlivých charakteristik.
[17]Jedná se o speciální typ kolekce zajišťující přítomnost každého klíče pouze jednou.
[18]Dále jen anonymizovaná knihovna.

Anonymizace roku narození

První charakteristikou, ve které bude nutno využít metody generalizace, je rok narození. Do předpřipravené kolekce se postupně nahrají prvky shodující se v národnosti, pohlaví. Až se výsledky této kolekce propíšou do anonymizované knihovny, bude promazána a následně opět naplněna další skupinou dat v pořadí. Z důvodu úspory výpočetních prostředků je opět využita množina národností bránící duplicitnímu procházení anonymizované knihovny. Procházení i chybových prvků by v tomto kroku mohlo mít fatální následky pro výpočet anonymizovaného roku. Algoritmus nejprve vyšle seřazenou kolekci jednotlivých roků narození do metody, která vytváří intervaly. Kolekce je seřazena od nejnižšího čísla po nejvyšší. Zde je zkontrolováno, zda jednotlivé roky již v tomto stavu nejsou identické. V případě pozitivní odpovědi program algoritmus přistoupí k další podskupině. V opačném případě pokračuje s následujícími kroky generalizace.

Nyní je potřeba data uzpůsobit do tvaru, který již stanovenou podmínku naplní. Algoritmus se nejprve doptá seřazené kolekce na její celkovou velikost (celkový počet nahraných roků). Zde se funkcionalita rozpadá do dvou větví.



Obr. 3.1: Příklad postupu generalizace data narození při celkovém počtu záznamů rovno třem s různou hodnotou rozdílu.

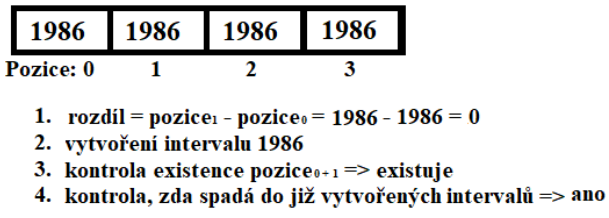
Je-li celkový počet menší než čtyři prvky (tři a méně prvků), vypočítá se rozdíl mezi nejvyšším a nejnižším prvkem. V případě rozdílu menšího než deset (devět a méně), je zkontrolováno, zda se roky shodují v první, druhé, třetí číslici. Tento mezikrok je nutný k zajištění, aby algoritmus rozpoznal, zda se nachází porovnávané roky v stejném desetiletí, století, tisíciletí. V případě kladné odpovědi je nejnižší i nejvyšší číslo zaokrouhleno dle intervalu: od nuly do čtyř – nahrazeno nulou, od pěti do devíti – nahrazeno devítkou. V druhém případě, je-li rozdíl větší než číslo deset nebo se porovnávané roky neshodují v některé z číslic, je potřeba přistoupit k vyššímu stupni generalizace. Ten v tomto případě spočívá v nahrazení poslední číslice (číslice vyjadřující jednotky) nejnižšího roku nulou a jednotkové číslice nejvyššího

roku devítkou. Všechna data této konkrétní kolekce budou nahrazena vypočítaným intervalem, který naplňuje podmínky k-anonymity.

Druhou možností charakteristickou celkovým počtem v kolekci data narození vyšší než tři (čtyři a více), je postup následující. Nejprve je získán rozdíl dvou nejnižších sousedících hodnot. Vyšší rok narození se odečítá od nižšího. Rozdíl pak nabývá hodnot:

- Hodnoty nula – roky narození jsou ponechány.
- Hodnoty v intervalu jedna až devět – zde je nutno nejprve zjistit, u které číslice se roky začínají lišit. To lze zjistit procházením zleva doprava a porovnáváním jednotlivých čísel:
  - Neshoda u třetí číslice – poslední číslice nejnižší hodnoty je nahrazena nulou, nejvyšší pak nahrazena devítkou[19].
  - Neshoda u čtvrté číslice vyjadřující jednotky, shodující se však v první, druhé, třetí číslici – jsou zkontrolovány poslední číslice obou prvků:
    - Obě hodnoty jsou v prvním intervalu – rok nahrazen rozsahem ve tvaru xxx0 – xxx4
    - Každá hodnota je v jiném intervalu – rok nahrazen rozsahem ve tvaru xxx0 – xxx9
    - Obě hodnoty jsou v druhém intervalu – rok nahrazen rozsahem ve tvaru xxx5 – xxx9
- Hodnoty vyšší než deset – poslední číslice nejnižší hodnoty nahrazena nulou, nejvyšší pak nahrazena devítkou.

Tím dojde k anonymizaci prvních hodnot dle principu k-anonymity. Algoritmus zkontroluje, zda v kolekci existuje pozice vypočítaná jako součet aktuální pozice a čísla jedna (z důvodu kontroly vyšších pozic). Pokud by možnost existence nebyla kontrolována, vznikla by zde chyba při přístupu k neexistujícímu prvku, a to vždy v případě porovnávání posledního prvku. Jestliže vypočítaná pozice neexistuje, objekt, který se nachází na aktuální pozici, je přesunut do chybové kolekce. Tento krok je vždy aplikován na poslední prvek seznamu.

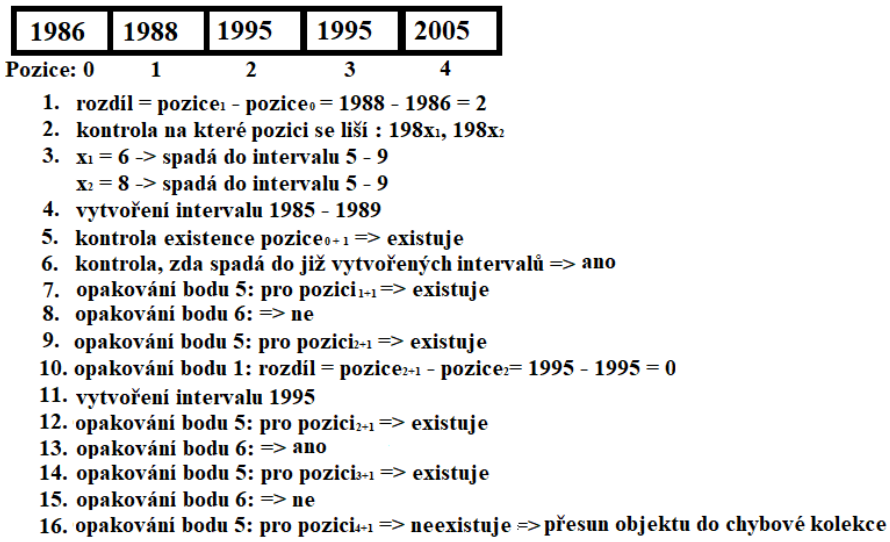


Obr. 3.2: Příklad postupu generalizace data narození v případě identických hodnot.

[19]Vysvětlení procesu nahrazení je identické jako u anonymizace roku narození tří a méně prvků.

Existuje-li prvek na vypočítané pozici, algoritmus ověří, zda nespadá do již vytvořeného intervalu (jeho hodnota je rovna druhému prvku, nebo zapadá do rozsahu již existujícího intervalu). Pokud ano, nastaví si hodnotu na daný interval, rok a opakuje výpočet nové hodnoty, tentokrát však se svojí aktuální pozicí zvýšenou o jedna.

V druhém případě, kdy nezapadá do již vytvořeného intervalu, roku a algoritmus ví, že následující objekt existuje, vypočítá se rozdíl hodnot mezi následující a aktuální pozicí[20]. Výsledky jsou vyhodnoceny dle vzniklých rozdílů: nula, jedna až devět, deset a více. Následné reakce jsou identické jako u rozdělení popsaného výše. Cyklus koluje do doby, než program přistoupí k poslednímu prvku kolekce.



Obr. 3.3: Příklad postupu generalizace data narození při rozdílných hodnotách.

3.2.3 Třetí fáze běhu

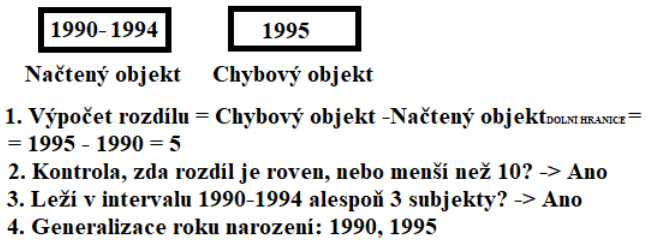
Nyní program musí vyřešit akce s daty, které se na základě odlišnosti dostaly během procesu anonymizace do chybové kolekce. Nejprve tedy přistoupí k objektům pod klíčem data narození. Tato data se nacházela na posledním místě v kolekci a jedná se o nejvyšší rok celé podskupiny[21]. V prvním kroku algoritmus ověří, zda již nespadají do jakéhokoliv intervalu, roku ve své podskupině. Pokud tento rok zapadá, nastaví si příslušný interval, rok na místo aktuálního data narození a zároveň se nahraje zpět do anonymizované knihovny. V opačném případě algoritmus postupně prochází kolekci roků, intervalů a postupně hledá rok, interval, od kterého není odlišen více než devět let[22]. Krok je realizován z důvodu zabránění vzniku velkých intervalů.

[20] Je zde opět pravidlem, že vyšší hodnota je odečtena od nižší.

[21] Podskupina je charakteristická shodou v národnosti, pohlaví.

[22] V případě, že se jedná o interval, algoritmus pro výpočet použije dolní hranici intervalu.

Najde-li takovýto interval, pokračuje s doptáním na četnost jeho výskytu. V situaci, kde je tato četnost vyšší nebo rovna třem, je vybrán jeden údaj a společně projdou generalizací s identickým postupem jako v případě generalizace roku narození pro tři a méně prvků.



Obr. 3.4: Postup generalizace roku narození na základě chybové kolekce.

Objekty přiřazené do chybové kolekce na základě pohlaví zjistí z celkové anonymizované knihovny, zda obsahuje minimálně tři prvky dané národnosti. Bude-li odpověď false[23], subjekt se pokusí vyhledat v této části chybové kolekce podobná data s identickou národností. Podaří-li se nalézt takovou shodu, prvky spolu projdou maximální generalizací charakteristiky pohlaví a následně je na ně aplikován bez ohledu na jejich počet algoritmus pro generalizaci roku narození tří a méně prvků. U druhé možnosti je subjekt vymazán z bezpečnostních důvodů z databáze. Program by nebyl schopen bez dalšího, avšak už nadměrného zásahu, zajistit stav k-anonymity.

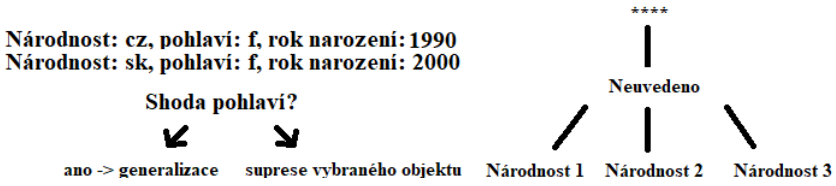
Obr. 3.5: Postup generalizace pohlaví na základě chybové kolekce.

V případě true[24] však pokračuje s doptáváním na rok narození, zda spadá do jakéhokoliv již vytvořeného intervalu. Pokud získá pozitivní odpověď, zjistí, zda v tomto intervalu, roku, leží minimálně tři objekty. V případě splnění i této podmínky je jeden subjekt (splňující výše uvedené podmínky) náhodně vybrán a spolu s chybovým

[23]Označení používané v informatice, vyjadřující logickou nepravdu.
[24]Označení používané v informatice, vyjadřující logickou pravdu.

objektem projdou maximální generalizací pohlaví. Chybový prvek převezme interval roku narození. Následně je nahrán do anonymizované knihovny.

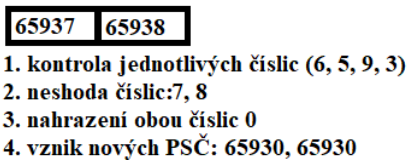
Poslední krokem je vyřešení chybové kolekce s klíčem národnosti. Přiřazením těchto subjektů k jiné národnosti by mohlo dojít k negativnímu ovlivnění celé výstupní databáze. Kupříkladu spojení obéznějšího národu s jiným národem by při malém množství dat zapříčinilo přílišný zásah do výsledků měření. Aby však data nebyla ztracena, program projde ostatní data v této části chybové kolekce a zkusí najít shodu v pohlaví bez ohledu na údaj národnosti. Jestliže shoda nebude nalezena, objekt bude vymazán z celé databáze. V druhém případě projdou tyto objekty krokem maximální generalizace dat národnosti a následně přejdou na generalizaci data narození. Opět je zde aplikován postup vycházející z anonymizace roku narození tří a méně prvků. Na závěr se takto upravená data zařadí do anonymizované knihovny.



Obr. 3.6: Postup generalizace národnosti na základě chybové kolekce.

3.2.4 Čtvrtá fáze běhu

V této fázi všechny prvky téměř splňují podmínky k-anonymity. Poslední charakteristikou, která nebyla zkontrolována, je PSČ.



Obr. 3.7: Postup generalizace PSČ.

Algoritmus postupně projde jednotlivé vytvořené podskupiny na základě stejného intervalu, roku narození, pohlaví, národnosti a zkontroluje, na které pozici porovnávaných údajů PSČ se objekty liší. Jednotlivé pozice jsou vybírány zleva doprava. Narazí-li se na neshodu, je daná číslice nahrazena hvězdou (*). Následně je tento výsledek vyhodnocen způsobem, který kontroluje, na které pozici se nachází

očekávaný znak. Všechny následující pozice včetně aktuální jsou nahrazeny nulou. Z postupu je zřejmé, že zde nedochází k žádnému přesunu do chybové kolekce.

Poslední dobrovolné údaje – BMI, zdravotní stav nejsou žádným způsoben zahrnuty do procesu anonymizace z důvodu složitosti a velké ztráty dat pro malé skupiny vstupních subjektů v řádech desítek. Jednotlivá čísla BMI navíc neodpovídají konkrétnímu typu postavy. Jedná se pouze o obecné ukazatele. Pro hodnotnější a snadno čitelné výsledky budou jednotlivé údaje zastupující číselnou hodnotu nahrazeny následovně:

- 0 – 18,4: podváha
- 18,5 – 25: normální váha
- 25 = 40: nadváha
- nezapadá do žádného intervalu: ****

Zadá-li uživatel druhou možnost – údaje o zdravotním stavu subjektu, není možné, aby program sám vyhodnotil možnosti, kterými lze vybraný problém zastřešit. V těchto případech by bylo nutné využít umělou inteligenci, centralizovanou databázi či uzpůsobit anonymizátor pro konkrétnější datovou sadu. Uživatel zadáním těchto údajů navyšuje možnost re-identifikace, ale zároveň navyšuje i celkový informativní přínos. Je nutné, aby sám již posoudil anonymitu takových dat s přihlédnutím k možným útokům pomocí re-identifikace.

Nyní by všechny údaje anonymizované knihovny měly splňovat k-anonymitu. Program však provede závěrečnou bezpečnostní kontrolu. Jestliže narazí na podskupinu, která stanovené podmínky nesplňuje, je odstraněna z anonymizované knihovny z důvodu možnosti poškození dat. Následně přechází do poslední fáze, a to je zápis výsledků do souboru[25]. Údaje jsou zapisovány postupně, tak jak byly načítány. Jestliže údaj byl vymazán, algoritmus zapíše do souboru prázdný řádek. Uživatel si tak bude moci prohlédnout ztrátu informací zapříčiněnou generalizací, supresí. Tímto krokem program dokončí proces anonymizace.

3.3 Grafické rozhraní programu

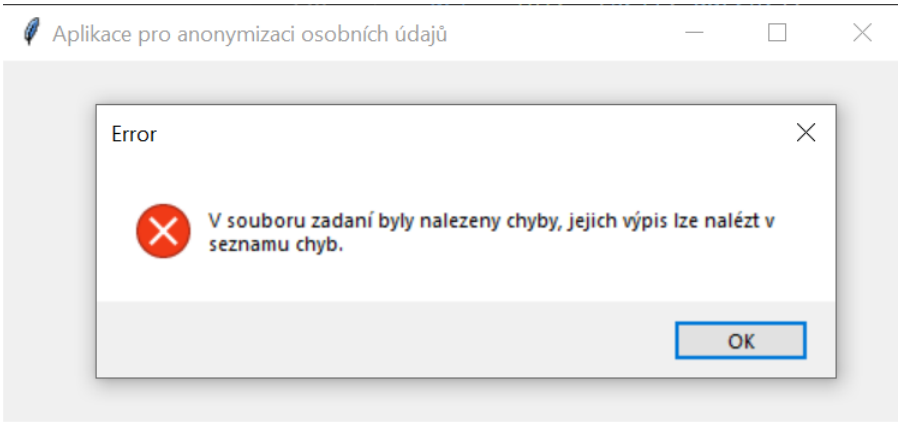
Naprogramovaná aplikace disponuje grafickým uživatelským rozhraním, které má za úkol informovat uživatele o aktuálním stavu programu. K vytvoření této části je využita knihovna tkinter, která je distribuována stejně jako využitý programovací jazyk Python pod Python licens. Tato licence je vlastnostmi postavena na open source licens – ve volném překladu veřejná licence dovolující uživateli určité zásahy do autorsko-právní ochrany[26]. Porušení jakékoliv podmínky zapříčiní neplatnost li-

[25] Jedná se o soubor s koncovkou txt.

[26] Jedná se o druh práva, který zajišťuje autorům ochranu jejich díla. Vztahuje se pouze k dílům, která jsou výsledkem tvůrčí činnosti autora.

cence a zásah do autorského práva.[43, 44, 45]

Program nejprve po spuštění informuje uživatele, zda data byla načtena v pořádku. Jestliže se nepodaří namapovat vstupní, výstupní soubor, samotná data jsou zadána v chybném tvaru, program tuto skutečnost oznámí uživateli chybovým dialogovým oknem. V případě nalezení dat v neočekávaném formátu umožní náhled konkrétních chyb a důvodu, proč pravděpodobně byla takto vyhodnocena.



Obr. 3.8: Ukázka výpisu dialogového okna z přiložené aplikace.

Obdobným způsobem je uživatel informován o dalších již výše popsaných hlavních krocích v průběhu anonymizace. Po ukončení operace zapsání dat do textového souboru umožní v poslední fázi zobrazení anonymizovaných informací. Program běží do té doby, dokud uživatel sám neukončí jeho běh.

Závěr

Teoretickým výstupem závěrečné práce bylo nejen přiblížit hranice vytyčené kolem osobního údaje, ale i přiblížit základní pojmy využívané ve zpracování osobních údajů. První kapitola představila aktuální, platné legislativní dokumenty, případy pochybení nadnárodních společností, práva, povinnosti objektů jako je zpracovatel, správce, subjekt údajů, podmínky zpracování vycházející z Listiny základních práv EU a nařízení GDPR, ale i jeden ze zlomových případů, kdy bylo poprvé rozhodnuto o nepřímých osobních údajích.

Druhá kapitola navázala na pojem zpracování. Specifikovala různé metodiky, kterými lze zajistit, aby identifikátor ztratil vlastnost jedinečnosti a došlo k přetrhání vazeb s konkrétním subjektem údajů. Zhodnotila rozdělení takovýchto vybraných technik dle škály dopadu a ztrát na výsledná data (pseudonymizace, anonymizace). Poukázala však i na možné útoky, kterými „útočník“ dokáže realizovat proces re-identifikace i s anonymizovanými daty. V neposlední řadě upozornila na možné propojení veřejných, anonymizovaných informací.

Praktický výstup práce dle zadání tvoří nejen návrh postupu anonymizačního procesu, ale i jeho samotný přepis do programového prostředí. Vytvořený program dokáže převést původní data s identifikátory na data, která jsou natolik obecná, že by neměla vést k jedinečné identifikaci, pouze však v případech, kdy nebudou zadány dodatečné informace. Program pro co nejširší využití nabízí možnost zadání zdravotního stavu společně s BMI, avšak tyto informace nezahrnuje do samotného procesu anonymizace. V případě BMI pouze dochází k zobecnění zadaných hodnot dle předem zadaných intervalů. Zdůvodnění tohoto konkrétního kroku popisuje poslední část bakalářské práce. Konečné vyhodnocení, zda jsou data opravdu anonymní, tak stále zůstává na správci, zpracovateli osobních údajů.

I přestože je návrh uzpůsoben k co nejmenším možným informačním ztrátám (z důvodu volby nízkého čísla duplicity povinných údajů), stále zde existuje možnost přílišných ztrát dat při malé podobnosti mezi subjekty. Tato ztráta je však brána jako nezbytná pro vytvoření minimální k-anonymity. Při procesu anonymizace správce vždy musí počítat s určitou podobnou informační ztrátou. Zadá-li uživatel větší počet dat, existuje možnost větší podoby identifikátorů a i menší potřeba zobecnění informací.

K přiloženému programu je také vytvořeno základní grafické rozhraní pro komunikaci, které informuje uživatele o stavu běhu, chybách, úspěšném ukončení. Veškeré knihovny využité při tvorbě jsou distribuované pod veřejnou licencí, a tak nebylo zasáhnuto do žádného autorského práva.

Výsledkem bez zadání dodatečných informací jsou vždy data splňující princip k-anonymity. Program však neuvažuje jednotlivé váhy, citlivost informací, a tak

zůstává na správci rozhodnutí, zda jsou dle první kapitoly výsledná data opravdu anonymní, nespadají nadále pod pojem osobní údaj a proto se na ně ochranný rámec právní úpravy ochrany osobních údajů nebude nadále aplikovat.

Literatura

[1] *JEŽEK, Mojmír. Big data a pokročilá analytika datových souborů. Ecovislegal.cz.* [online] Dostupné z URL: <https://www.ecovislegal.cz/aktuality>.

[2] *OHM, Paul. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review. 2009, č. 6, s. 1701–1777. ISSN 0041-5650.*

[3] *Usnesení předsednictva České národní rady č. 2/1993 Sb., o vyhlášení LISTINY ZÁKLADNÍCH PRÁV A SVOBOD jako součásti ústavního pořádku České republiky. In: Sbírka zákonů. 1992.* [online]. Dostupné z URL: <https://www.psp.cz/docs/laws/listina.html>.

[4] *Zákon 40/2009 Sb. In: Trestní zákoník. Ministerstvo vnitra České republiky, 2009, 40/2009.* [online] Dostupné z URL: <https://www.zakonyprolidi.cz/cs/2009-40>.

[5] *Úřad pro ochranu osobních údajů.* [online] Dostupné z URL: <https://www.uoou.cz/>.

[6] *Listina základních práv Evropské unie. In: Úřední věstník, 2012/C, 326/02, 2012* [online] Dostupné z URL: <https://eur-lex.europa.eu/legal-content/CS/ALL/?uri=CELEX%3A12012P%2FTXT>.

[7] *Směrnice Evropského parlamentu a Rady Evropské unie 95/46/ES o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů. In: Úřední věstník, 1995, 95/46/ES.* [online] Dostupné z URL: <https://eur-lex.europa.eu/legal-content/CS/TXT/HTML/?uri=CELEX:31995L0046&from=cs>.

[8] *Základní příručka k ochraně údajů: Obecné nařízení. Úřad pro ochranu osobních údajů. 2017.* [online] Dostupné z URL: <https://www.uoou.cz/1-obecne-narizeni/d-27266/p1=3938>.

[9] *Nařízení Evropského parlamentu a Rady Evropské unie 2016/679 ze dne 27. dubna 2016 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení směrnice 95/46/ES. In: Úřední věstník, 2016, 2016/679.* [online] Dostupné z URL: <https://eur-lex.europa.eu/homepage.html>.

[10] *Co jsou osobní údaje. Česká asociace ochrany osobních údajů.* [online] Dostupné z URL:
<http://www.gdprbezobav.cz/osobni-udaje>.

[11] *Předpis 110/2019 Sb. – Zákon o zpracování osobních údajů. In: Poslanecká sněmovna Parlamentu České republiky: Poslanecká sněmovna Parlamentu České republiky, 2019, ročník 2019, 110/2019.* [online] [cit. 16. 10. 2020]. Dostupné z URL:
<https://www.psp.cz/sqw/hp.sqw?akk=7>.

[12] *BARBARO, Michael; ZELLER, Tom. A Face Is Exposed for AOL Searcher No. 4417749. The New York Times, 2006.* [online] Dostupné z URL:
<https://www.nytimes.com/2006/08/09/technology/09aol.html>.

[13] *MILLS, Elinor. AOL sued over Web search data release. CNET, 2013.* [online] Dostupné z URL:
<https://www.cnet.com/news/aol-sued-over-web-search-data-release/>.

[14] *AOL Is Sued Over Privacy Breach. Los Angeles Times, 2006.* [online] Dostupné z URL:
<https://www.latimes.com/archives/la-xpm-2006-sep-26-fi-aol26-story.html>.

[15] *Nonnemann, František. Objektivní, či subjektivní pojetí osobních údajů? Právní rozhledy. 2015, č. 12, s. 425–431.*

[16] *Rozsudek Soudního dvora Evropské unie ze dne 19. 10. 2016 ve věci č. C-582/14, Breyer. Soudní dvůr EU, 2016.* [online] Dostupné z URL:
<http://curia.europa.eu/juris/document/document.jsf?docid=184668&doclang=CS>.

[17] *MATYSOVÁ, Monika; NEŠPŮREK, Robert; OTEVŘEL Richard. Rozhodnutí Breyer a dynamická IP adresa jako osobní údaj. Pravniprostor.cz, 2017.* [online] Dostupné z URL:
<https://www.pravniprostor.cz/clanky/obcanske-pravo/rozhodnuti-breyer-a-dynamicka-ip-adresa-jako-osobni-udaj>.

[18] *Základní příručka k ochraně údajů: Nejdůležitější pojmy. Úřad pro ochranu osobních údajů. 2017.* [online] Dostupné z URL:
<https://www.uoou.cz/1-obecne-narizeni/d-27266/p1=3938>.

[19] *Maastrichtská smlouva o Evropské unii. In: Úřední věstník, C 191, 1992, s. 1–112.* [online] Dostupné z URL:

<https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=celex%3A11992M%2FTXT>.

[20] *Směrnice o ochraně fyzických osob 2016/680. In: Úřední věstník, 2016, 2016/680.* [online] Dostupné z URL: <https://eur-lex.europa.eu/legal-content/CS/TXT/PDF/?uri=CELEX:32016R0679&from=CS>.

[21] *Účel a právní základ zpracování. Medium. 2017.* [online] Dostupné z URL: <https://medium.com/ochr%C3%A1nce-%C3%BAdaj%C5%AF/%C3%BA%C4%8Del-a-pr%C3%A1vn%C3%AD-z%C3%A1klad-zpracov%C3%A1n%C3%AD-79718122d08b>.

[22] *Zásady zpracování osobních údajů. Ministerstvo vnitra České republiky.* [online]. Dostupné z URL: <https://www.mvcr.cz/gdpr/clanek/zasady-zpracovani-osobnich-udaju.aspx>.

[23] *PAVLOVIČOVÁ, Jana. GDPR – Správce osobních údajů. GDPR.* [online] Dostupné z URL: <https://www.gdpr.cz/gdpr/heslo/spravce-osobnich-udaju/>.

[24] *PAVLOVIČOVÁ, Jana. Definice osobních údajů podle GDPR. 22Hlav. 2017.* [online] Dostupné z URL: <https://www.bvmaudit.cz/definice-osobnich-udaju-podle-gdpr>.

[25] *Základní příručka k ochraně údajů: Práva subjektu údajů. Úřad pro ochranu osobních údajů. 2017.* [online] Dostupné z URL: <https://www.uoou.cz/1-obecne-narizeni/d-27266/p1=3938>.

[26] *KOHÚTOVÁ, Zuzana. Anonymizace, pseudonymizace a šifrování osobních údajů jako bezpečnostní opatření dle GDPR. Fly-eye.cz. 2017.* [online] Dostupné z URL: <https://fly-eye.cz/blog-detail-1.html>.

[27] *Guidance on Anonymisation and Pseudonymisation. Data Protection Commission, 2019.* [online] Dostupné z URL: <https://www.dataprotection.ie/en/guidance-landing/anonymisation-and-pseudonymisation>.

[28] *Data masking: Anonymisation or pseudonymisation. PrivSec Report, 2017.* [online] Dostupné z URL: <https://gdpr.report>.

[29] *Guide to Basic Data Anonymisation Techniques. Personal Data Protection Commission Singapore, 2018.* [online] Dostupné z URL: <https://iapp.org/media/pdf/resource_center/Guide_to_Anonymisation.pdf>.

[30] *LEVICKÝ, Dušan. Kryptografia v informačnej bezpečnosti. Košice: Elfa, 2005, s. 266, ISBN 978-80-8086-022-6.*

[31] *Pseudonymisation techniques and best practices. European Union Agency, 2019. ISBN 978-92-9204-307-0.*

[32] *SPALDING, Hugo. Anonymous and pseudonymous data: Are they actually important? Data Marketing Association, 2018.* [online] Dostupné z URL: <https://dma.org.uk>.

[33] *SAMARATI, Pierangela; SWEENEY, Latanya. Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression. 1998.*

[34] *SWEENEY, Latanya. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledgebased Systems. 10 (5), s. 571–588. 2002*

[35] *SAMARATI, Pierangela; SWEENEY, Latanya. Generalizing data to provide anonymity when disclosing information. Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. 1998.*

[36] *SWEENEY, Latanya. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh, 2000.*

[37] *EL EMAM, Khaled; DANKAR, Fida Kama. Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association. 15 (5), s. 627–637. 2008.*

[38] *MACHANAVAJJHALA, Ashwin; GEHRKE, Johannes; KIFER Daniel; VENKITASUBRAMANIAM, Muthuramakrishnan. -Diversity: Privacy Beyond k-Anonymity. Department of Computer Science. Cornell University, 2007.*

[39] *LI, Ninghui; LI, Tiancheng; VENKATASUBRAMANIAN Suresh. t-Closeness: Privacy beyond k-anonymity and l-diversity. IEEE 23rd International Conference on Data Engineering. Institute of Electrical and Electronic Engineers. 2007. ISBN 1-4244-0802-4.*

[40] *BRAND, Ruth. Microdata protection through noise addition. Lecture Notes in Computer Science, vol. 2316, pp. 97-116, 2002. Volume title Inference Control in Statistical Databases, ed. J. Domingo-Ferrer. Berlin: Springer-Verlag, 2002. ISBN 978-3-540-43614-0.*

[41] *MIVULE, Kato. Utilizing Noise Addition for Data Privacy, An Overview. Proceedings of the International Conference on Information and Knowledge Engineering, (pp.65-71). USA: Las Vegas, 2013.*

[42] *LibreOffice Document Foundation. The Document Foundation.* [online] Dostupné z URL:
<https://www.documentfoundation.org/>.

[43] *History and License. Python.* [online] Dostupné z URL:
<https://python.org/3/license.html>.

[44] *TIŠNOVSKÝ, Pavel. Grafické uživatelské rozhraní v Pythonu: knihovna Tkinter. ROOT.CZ. 2017.* [online] Dostupné z URL:
<https://www.root.cz/autori/pavel-tisnovsky/?pi=9>.

[45] *TIŠNOVSKÝ, Pavel. Grafické uživatelské rozhraní v Pythonu: další možnosti nabízené widgety Text a ScrolledText. ROOT.CZ. 2017.* [online] Dostupné z URL:
<https://www.root.cz/autori/pavel-tisnovsky/?pi=9>.

Seznam symbolů a zkratek

K	Klíč využívající se v symetrické kryptografii.
Prec(RT)	Míra přesnosti.
h	Úroveň generalizace.
PT	Původní data s identifikátory.
ϵ_j	Distribuované chyby náhodné proměnné.

A Popis přiloženého média

Součástí řešení problematiky, kterými se zaobírá bakalářská práce, je i přenesení celého teoretického návrhu anonymizátoru do programového prostředí. Program je zkompilován do spustitelné aplikace pro jednoduché využití. Součástí celého řešení jsou dále instrukce, předpřipravené soubory, informace o licenci, pod kterou je program distribuován a v neposlední řadě tři soubory (NacteniDat.py, Anonymizace.py, Zapis.py) obsahující samotný kód programu ke dni 31.5.2021.