Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Bakalářská práce

Zabezpečené zpracování medicínských obrazových dat

Plzeň, 2016 Jaroslav Malát

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 23. června 2016

Jaroslav Malát

Poděkování

Úvodem bych chtěl poděkovat vedoucí mé bakalářské práce doc. Dr. Ing. Janě Klečkové za důležité připomínky a rady k formální i obsahové stránce práce.

Abstract

Theme: Secure processing of medical image data

The presented bachelor's thesis is focused on the problem of security in the processing of medical image data.

The theoretical part of bachelor thesis is focused on legal regulations and laws for handling private data about patients and presents the most widely used standard for image files, which is DICOM. There is also analysis of available methods for anonymizing image data and the proposed algorithm that improves the process of anonymization.

The practical part is focused on the implementation of algorithm for anonymizing image part of medical data. The program is developed in programming language Java and I operates on system Linux Ubuntu 14.04LTS.

Abstrakt

Téma: Zabezpečené zpracování medicínských obrazových dat

Předkládaná bakalářská práce se zaměřuje na problém bezpečnosti v oblasti zpracování medicínských obrazových dat.

Teoretická část bakalářské práce se zaměřuje na právní předpisy a zákony pro práci s citlivými údaji o pacientech a seznamuje s nejpoužívanějším standardem pro obrazové soubory, kterým je DICOM. Dále je provedena analýza dostupných metod anonymizace obrazových dat a navržení algoritmu, jenž zlepšuje proces anonymizace.

Praktická část je zaměřena na implementaci navrženého algoritmu pro anonymizaci obrazové části medicínských dat. Program je vyvíjen v programovacím jazyce Java a na operačním systému LINUX UBUNTU 14.04LTS.

Obsah

Úvod

Jako téma bakalářské práce jsem si zvolil téma zabezpečené zpracování medicínských dat. Hlavním důvodem byl fakt, že medicínská data jsou součástí běžného života každého jednotlivce.

Práci jsem rozdělil do dvou částí, a to části teoretické a praktické. V teoretické části jsem se snažil o vysvětlení všech pojmů spojených se zvoleným tématem. Na úvod provedu výtah z legislativy České republiky platné k začátku roku 2016, která se dotýká práce s citlivými daty. Dále se věnuji vymezení základních pojmů, jako je DICOM. Praktickou část jsem se rozhodl věnovat vytvoření anonymizačního programu. V poslední části této bakalářské práce jsem se věnoval zhodnocení celého mého snažení a zhodnocení funkčnosti anonymizačního programu.

Mezi cíle této práce patří analýza legislativních a bezpečnostních požadavků pro zpracování medicínských dat. Bude popsán Standardní formát souvisejících obrazových dat DICOM a také technologie PACS, umožňující správu, ukládání a zobrazování obrazové dokumentace. Dále bude provedena analýza dostupných metod anonymizace obrazových dat a navržen algoritmus zlepšující proces anonymizace. Můj anonymizační program vyhodnotí, zda obrazový soubor obsahuje jakákoliv citlivá data, ať už o lékaři nebo o pacientovi, odpovídající předem daným předpisům a následně tato data anonymizuje.

Realizace je rozdělena do následujících bodů, které jsou stejné jako body v zadání bakalářské práce a to:

- seznámit se se současnými legislativními a bezpečnostními požadavky týkajícími se zpracování medicínských dat a standardním formátem souvisejících obrazových dat,
- provést analýzu dostupných metod anonymizace obrazových dat a navrhnout algoritmus zlepšující proces anonymizace,
- navržený algoritmus implementovat a ověřit jeho správnou činnost,
- výsledky zhodnotit.

Práce by měla sloužit jako ucelený popis zabezpečeného zpracování medicínských dat.

1. Úvod do bezpečnostní problematiky a obrazová dokumentace

Na úvod bych rád definoval pojem zdravotnické dokumentace. Zdravotnická dokumentace je souhrn informací o pacientovi (klientovi) zdravotnického zařízení, který může být vedený v jakékoliv podobě.

Tato dokumentace má především sloužit jako pracovní nástroj při léčbě, ale případně i jako doklad či dokonce důkaz v případě forenzního projednávání postupu lékaře při léčení. Nesprávně vedená dokumentace může pomoci v utvrzení o chybném postupu nebo přinejmenším znemožnit dokázání postupu správného.

Většina zdravotnických zařízení ať prvního styku, mezi které patří praktický lékař pro dospělé, zubní lékař resp. stomatologická ambulance nebo praktický lékař pro děti a dorost - pediatrie, gynekologická ambulance, lékařská služba první pomoci; dále ambulantních zařízení, do kterých se řadí oční lékař resp. oftalmologická ambulance, ortopedie, psychiatrie, neurologie, dermatologie, rehabilitace, urologie, klinická psychologie a logopedie, ORL – otorhinolaryngologie, alergologie, dermatovenerologie a zdravotní rehabilitace, nebo hospitalizačních zařízení, kam se řadí nemocnice, porodnice, nemocnice následné péče, fakultní nemocnice, léčebna dlouhodobě nemocných, odborný léčebný ústav, psychiatrická léčebna, nevyjímaje lékárny, laboratoře a lázeňská zdravotní zařízení, dnes využívají informační systémy s údaji o pacientech včetně jména a adresy, rodného čísla, platebních informací a zejména pak citlivé údaje o průběhu léčby.

Tato data jsou nesporně mnohonásobně více ohrožena oproti papírové formě. Slabými místy při procesu nakládání se zdravotnickými informacemi nejsou technologie, ale lidský prvek, který bývá často opomíjen.

Z konkrétních případů lze uvést například hackerský útok na informační síť nejmenovaného zdravotnického zařízení. Pokud půjdeme do krajnosti, mohla by být ohrožena i celková zdravotnická péče. Na druhou stranu můžeme tvrdit, že absolutní bezpečnost informačních systémů je vždy pouze teoretickým pojmem.

Jako argument uvádím překvapující zjištění z kontrolní činnosti ÚOOÚ (Úřadu pro ochranu osobních údajů), více na (1).

- Elektronická podoba zdravotnické dokumentace nebyla totožná s tištěnou.
- Informační systém nemocnice neumožňoval aktivní verzi sledování přístupu do něj, což je ze zákona povinné a musí to být zaznamenáváno.
- Do informačního systému nemocnice vstoupila neoprávněná osoba, přičemž ji nebylo možno ověřit kvůli vypnuté funkci monitorování přístupu.
- Velmi neuváženým počinem bylo zveřejnění záznamu z operačního zákroku na webu nemocnice. Byla zřejmá identifikace osoby a k tomu byl ještě záznam doplněn jménem, částí příjmení a dokonce rodným číslem. I přes doklad o písemném souhlasu musel být záznam okamžitě odstraněn.
- Došlo k úniku citlivých informací o pacientovi. Lékař neznající bezpečnostní prvky hesel komunikoval s pacienty přes e-mail s velmi jednoduchým a nedostatečným heslem, které bylo prolomeno.
- Těžko uvěřitelný je také případ nestátního zdravotnického zařízení, které mělo registrační skříně v čekárně pro pacienty. V ordinaci se střídaly dvě lékařky a nedocházelo k důslednému zamykání skříní. Výsledkem bylo, že kdokoliv v čekárně mohl mít snadný přístup ke zdravotnické dokumentaci ostatních pacientů.

Dalším rizikovým faktorem je bezesporu to, že k citlivým datům přistupují nejen zdravotníci, ale i různé dodavatelské firmy, správci informační sítě apod.

Dále je k datům vyžadován četnější přístup ve srovnání s klasickou papírovou dokumentací.

Jak jsem již řekl, bohužel nelze docílit dokonalé ochrany informačního systému, ale snahou by mělo být dosáhnout optimální úrovně zabezpečení. Zdravotnická zařízení by měla mít vypracován plán kybernetické bezpečnosti a plán krizové připravenosti, který popisuje možné rizikové situace jak vně perimetru (mimo nemocnici – živelná pohroma, teroristický útok, velká dopravní nehoda), tak uvnitř (požár, teroristický útok). Kybernetický útok je nebezpečí hrozící stále většímu množství lidí. Zvenku je to především možnost průniku internetem, WI-FI sítěmi nebo GSM sítí z chytrých zařízení. Zevnitř jde o lidský prvek - nevzdělaného nebo nedisciplinovaného uživatele. Ve většině případů zdravotnických zařízení nejde pouze o snahu ochránit osobní údaje pacientů, ale také o udržení jejich chodu. V tomto ohledu zůstávají zdravotnická zařízení mnohem citlivějším místem

k útoku než například státní instituce či banka. Velmi důležité je školit uživatele IT systémů a eliminovat tak případné chyby lidského faktoru.

Dalšími důvody většího zabezpečení dat ve zdravotnictví lépe než v jiných oborech lidské činnosti je například chybná diagnóza na základě pozměněných údajů, následné pochybení v léčbě a bezprostřední ohrožení zdraví i života samého.

Požadována je dostupnost jak životně důležitých dat a údajů v případě ohrožení zdraví či přímo života pacienta, tak dostupnost údajů pacienta různými odděleními, dále dostupnost pro služby, změny personálu, zástupy a rozlišení různé citlivosti dat pacienta.

Aby ochrana byla účinná, měli bychom znát potenciální slabá místa a možné útočníky (vnitřní i vnější), míru ohrožení, případné postupy, nutné náklady na eliminaci rizik a typ hrozby či útoku, jaká IS vrstva je ohrožena (infrastruktura, OS, DB, aplikace), výstupy mimo informační systém zdravotnického zařízení a způsoby zabezpečení výměny dat mezi zdravotnickými zařízeními, pojišťovnami a dalšími subjekty.

1.1. Právní předpisy

Z hlediska práce se zdravotnickou dokumentací jsou v české legislativě v současnosti tyto důležité platné zákony a vyhlášky:

- zákon č. 96/2001 Sb., o lidských právech a biomedicíně,
- zákon č. 372/2011 Sb., o zdravotních službách,
- vyhláška č. 98/2012 Sb., o zdravotnické dokumentaci,
- zákon č. 101/2000 Sb., o ochraně osobních údajů,
- zákon č. 181/2014 Sb., o kybernetické bezpečnosti,
- vyhláška č. 316/2014 Sb., o kybernetické bezpečnosti,
- vyhláška č. 317/2014 Sb., o významných informačních systémech a jejich určujících kritériích.

Více informací o zákonech na (2) nebo o kybernetickém zákoně na (3).

Z výše uvedených zákonů a vyhlášek se pokusím vybrat či citovat pasáže, které se přímo či okrajově dotýkají tématu bakalářské práce.

1.1.1. Zákon č. 96/2001 Sb., o lidských právech v biomedicíně

Každý má právo na ochranu soukromí ve vztahu k informacím o svém zdraví.

Každý je oprávněn znát veškeré své informace shromažďované o jeho zdravotním stavu a je nutno respektovat i přání každého nebýt takto informován. (4)

1.1.2 Zákon č. 372/2011 Sb., o zdravotních službáchá a jeho novela

Část šestá zákona se věnuje přímo zdravotnické dokumentaci a národnímu zdravotnickému informačnímu systému.

Hlava první určuje zapracování osobních údajů, hlava druhá zdravotnickou dokumentaci, její vedení a nakládání s ní, možnosti do jejího nahlížení či pořizování kopií.

Hlava třetí je pak věnována přímo Národnímu zdravotnickému informačnímu systému (NZIS).

NZIS je jednotný celostátní informační systém veřejné správy určený k:

- zpracování údajů o zdravotním stavu obyvatelstva, o činnosti poskytovatelů, o zdravotnických pracovnicích,
- vedení Národních zdravotních registrů a zpracování údajů v nich uvedených,
- vedení Národních registrů poskytovatelů a zdravotnických pracovníků a zpracování údajů v nich uvedených,
- zpracování údajů pro statistické účely a k zpracování různých šetření.

NZIS je de facto systémem, který sdružuje a spravuje informace obsažené v:

- Národních zdravotních registrech (Národní onkologický registr, Národní registr hospitalizovaných, Národní registr kloubních náhrad a jiné),
- Národním registru poskytovatelů,
- Národním registru zdravotnických pracovníků,
- Národních registrech podle zákona o transplantacích,

- informačních systémech infekčních nemocí vedených podle zákona o ochraně veřejného zdraví.

Právní úprava po novele rozšiřuje určení NZIS a kromě Národních zdravotních registrů předpokládá i existenci Národního registru poskytovatelů a Národního registru zdravotnických pracovníků.

Dle původní úpravy bylo možné také v registrech zpracovávat bez souhlasu subjektu jejich osobní a další údaje, tyto však byly přesně vymezené. V novém zákoně je oproti tomu rozsah zpracovávaných údajů rozšířen, dle něj se předávají i údaje související se zdravotním stavem pacienta ve vztahu k onemocnění a jeho léčbě, a to zejména:

- údaje socio-demografické a diagnostické,
- údaje o osobní, rodinné a pracovní anamnéze pacienta související s onemocněním včetně posouzení jeho aktuálního zdravotního stavu,
- údaje o poskytovaných zdravotních službách pacientovi,
- údaje o výkonu povolání nebo zaměstnání, popřípadě o výkonu služebního poměru, potřebné pro posouzení zdravotního stavu pacienta.

Jedná se tedy o demonstrativní výčet, který je, oproti úpravě v zákoně o péči o zdraví lidu, rozšiřující a nelze úplně jasně určit, které všechny informace bude možné pod toto ustanovení podřadit. Zákon dále uvádí rozsáhlý seznam subjektů povinných poskytovat informace.

Osoby oprávněné pracovat s informacemi z NZIS jsou určeny přímo v zákoně o zdravotních službách. Jedná se o oprávněné pracovníky správce a zpracovatele registru, zdravotnické pracovníky, oprávněné pracovníky Koordinačního střediska transplantací a oprávněné pracovníky institucí, které jsou zákonem zmocněny k využívání dat z NSIZ.

Strany zastávající úpravu NZIS tvrdí, že tato pouze kodifikuje stav, který už existoval předtím. Odpůrci tvrdí, že rozsah sdělovaných dat překračuje potřebnou míru, zákonná úprava je příliš obecná, a přenáší tak rozhodování o tom, jaká citlivá data budou sdělována na subjekt jiný, než je zákonodárce. Kritici dále tvrdí, že v zákoně chybí účelné a přesně stanovené postupy zpracovávání citlivých údajů

občanů a také řešení otázky zabezpečení údajů. Až následná praxe nejspíš ukáže, jak bude staronový systém fungovat.

Novela tohoto zákona

V souvislosti s anonymizací medicínských dat musím aktuálně zmínit hojně diskutovanou novelu o zdravotních službách, která nabyla účinnosti v červnu tohoto roku (2016).

Dne 17.5.2016 byl ve Sbírce zákonů zveřejněn pod č. 147/2016 Sb. zákon, kterým se mění zákon č. 372/2011 Sb., o zdravotních službách a podmínkách jejich poskytování (zákon o zdravotních službách), ve znění pozdějších předpisů.

Zpracovatelem návrhu novelizačního zákona je Ministerstvo zdravotnictví, které jako hlavní cíle novely označilo jednoznačné definování působnosti Ústavu zdravotnických informací a statistiky ČR, jakožto správce Národního zdravotnického informačního systému s přesně vymezenými kompetencemi a povinnostmi.

Nově je stanoven obsah a funkčnost Národního registru zdravotnických pracovníků, či vymezení Národního registru hrazených zdravotních služeb. Nově jsou rovněž ustanoveny Národní diabetologický registr a Národní registr intenzivní péče.

Zastánci novely tvrdí, že zprůhlední zdravotnictví a umožní zavést plnohodnotný systém hodnocení kvality péče. Odpůrci však poukazují na ochranu shromažďovaných dat o pacientech, která mají být anonymizována až postupně. Další námitkou je, že se bude sbírat více údajů a tudíž systém jejich zabezpečení bude muset být také mohutnější.

Konkrétně se zdravotnickou dokumentací a zdravotnickým informačním systémem zaobírá část šestá zákona, hlava I. - zpracování osobních údajů.

Zde je důležitý zejména § 52.

Při zpracování osobních údajů lze nakládat s rodným číslem pacienta.

Hlava II. řeší zdravotnickou dokumentaci, její vedení a ukončení v různých případech.

Hlava III. řeší Národní zdravotnický informační systém, který je určený mimo jiné pro zpracování údajů o zdravotním stavu obyvatelstva.

Údaji k identifikaci pacienta mohou být např.: číslo pojištěnce, rodné číslo, datum narození, popř. část adresy. (5)

§ 73 říká, že pro statistické a vědecké účely poskytuje statistický ústav z národních zdravotních registrů údaje pouze v podobě, z níž nelze určit konkrétní osobu, ať fyzickou či právnickou.

Zajímavé jsou i změny u Národního registru reprodukčního zdraví. Jsou zde zpracovány osobní údaje, jenž jsou potřebné pro identifikaci těhotné ženy, rodiček, nenarozeného dítěte, ženy s umělým či samovolným přerušením těhotenství. V registru jsou zpracovány údaje o reimplantačních a prenatálních vyšetřeních, taktéž i o potratech.

V národním diabetologickém registru jsou mimo jiné zpracovány i rizikové i prognostické faktory onemocnění, údaje k léčbě, osobní a rodinné anamnézy atd.

Osobní údaje jsou anonymizovány po 25 letech od úmrtí. V národním registru intenzivní péče již po 5 letech.

1.1.3. Vyhláška č. 98/2012 Sb., o zdravotnické dokumentaci

V § 1 se určuje, co má obsahovat zdravotnická dokumentace (identifikační údaje poskytovatele a pacienta, pacientovo pohlaví, data zápisu, razítka, pracovní závěry a konečnou diagnózu, rozsah poskytnutých služeb, aktuální vývoj zdravotního stavu pacienta, návrh léčebných postupů, podání léčivých přípravků, lékařské posudky, záznam o pracovní neschopnosti…)

§ 2 definuje součásti zdravotnické dokumentace, § 3 co musí být uvedeno na každém listu zdravotní dokumentace a kdo je zodpovědný za zápis.

V dalších paragrafech pak nalezneme povinnosti k uchování zdravotní dokumentace a nutnost elektronického podpisu v elektronické ZD.

§ 6 ukládá, že technické prostředky pro vedení zdravotnické dokumentace v elektronické podobě zaručí:

- zabezpečení výpočetní techniky softwarovými a hardwarovými prostředky před přístupem neoprávněných osob ke zdravotnické dokumentaci,
- vedení evidence všech přístupů ke zdravotnické dokumentaci včetně jejich oprav, změn a mazání.

Příloha č. 1 k vyhlášce č. 98/2012 Sb. určuje minimální obsah samostatných částí zdravotnické dokumentace, přílohy 2 a 3 určují zásady pro dobu a samotné uchování a následné zničení ZD. (6)

1.1.4. Zákon č. 101/2000 Sb., o ochraně osobních údajů

Důvodem vzniku zákona o ochraně osobních údajů bylo Listinou lidských práv a svobod zaručené právo na ochranu občana před neoprávněným zasahováním do jeho soukromého a osobního života neoprávněným shromažďováním, zveřejňováním nebo jiným zneužíváním osobních údajů.

Zákon se vztahuje na osobní údaje, které zpracovávají státní orgány, samospráva, fyzické a právnické osoby automatizovaně nebo jinými prostředky. Nevztahuje se na zpracování údajů fyzickou osobou nebo pro osobní potřebu a ve vymezených případech též na zpravodajské služby a policii.

V §13 jsou tyto důležité údaje týkající se ochrany osobních údajů:

- Správce a zpracovatel jsou povinni přijmout taková opatření, aby nemohlo dojít k neoprávněnému nebo nahodilému přístupu k osobním údajům, k jejich změně, zničení či ztrátě, neoprávněným přenosům, k jejich jinému neoprávněnému zpracování, jakož i k jinému zneužití osobních údajů, přičemž tato povinnost platí i po ukončení zpracování osobních údajů.
- Správce nebo zpracovatel je povinen zpracovat a dokumentovat přijatá a provedená technická a organizační opatření k zajištění ochrany osobních údajů v souladu se zákonem a jinými právními předpisy.
- Správce nebo zpracovatel posuzuje rizika týkající se:
    - plnění pokynů pro zpracování osobních údajů osobami, které mají bezprostřední přístup k osobním údajům,
    - zabránění neoprávněným osobám přistupovat k osobním údajům a k prostředkům pro jejich zpracování,

    - zabránění neoprávněnému čtení, vytváření, kopírování, přenosu, úpravě či vymazání záznamů obsahujících osobní údaje opatření, která umožní určit a ověřit, komu byly osobní údaje předány.
- V oblasti automatizovaného zpracování osobních údajů je správce nebo zpracovatel v rámci opatření podle odstavce 1 povinen dále také:
    - zajistit, aby systémy pro automatizovaná zpracování osobních údajů používaly pouze oprávněné osoby,
    - zajistit, aby fyzické osoby oprávněné k používání systémů pro automatizovaná zpracování osobních údajů měly přístup pouze k osobním údajům odpovídajícím oprávnění těchto osob, a to na základě zvláštních uživatelských oprávnění zřízených výlučně pro tyto osoby,
    - pořizovat elektronické záznamy, které umožní určit a ověřit, kdy, kým a z jakého důvodu byly osobní údaje zaznamenány nebo jinak zpracovány,
    - zabránit neoprávněnému přístupu k datovým nosičům. (7)

1.1.5. Zákon č. 181/2014 Sb., o kybernetické bezpečnosti

Předmětem úpravy tohoto zákona jsou práva a povinnosti osob, působnost a pravomoci orgánů veřejné moci v oblasti kybernetické bezpečnosti, nevztahuje se na informační nebo komunikační systémy, jež nakládají s utajovanými informacemi.

Hlava II popisuje systém zajištění kybernetické bezpečnosti. Ten zahrnuje bezpečnostní opatření, kybernetickou bezpečnostní událost a kybernetický bezpečnostní incident a jeho hlášení.

Dále pak popisuje evidence, opatření, varování, reaktivní a ochranná opatření a možné kontaktní údaje. Vymezuje úkoly národního i vládního CERT a jeho provozovatele. CERT (Computer Emergency Response Team) vznikl v roce 1988 na základě aféry s jedním z prvních počítačových červů, kterým byl tzv. Morrisův červ, jenž využil k svému šíření celosvětové sítě internetu. Od té doby CERT monitoruje všechny internetové průlomy, informuje o zranitelných místech

v různých systémech a na základě toho zveřejňuje maximální množství bezpečnostních rad.

Hlava III se pak přímo zaobírá stavem kybernetického nebezpečí. Definuje jej jako stav, kdy je ve velkém rozsahu ohrožena bezpečnost informací v informačních systémech nebo bezpečnost a integrita služeb nebo sítí elektronických komunikací. Nalezneme zde i podmínky pro vyhlášení nouzového stavu.

Všechny prováděcí předpisy k zákonu č. 181/2014 Sb., o kybernetické bezpečnosti, které platí stejně jako zákon od 1. 1. 2015, byly dne 19. 12. 2014 uveřejněny ve Sbírce zákonů v částce 127 pod tímto označením:

317/2014 Vyhláška o významných informačních systémech a jejich určujících kritériích

316/2014 Vyhláška o bezpečnostních opatřeních, kybernetických bezpečnostních incidentech, reaktivních opatřeních a o stanovení náležitostí podání v oblasti kybernetické bezpečnosti (vyhláška o kybernetické bezpečnosti)

315/2014 Nařízení vlády, kterým se mění nařízení vlády č. 432/2010 Sb., o kritériích pro určení prvku kritické infrastruktury. (8)

1.1.6. Vyhláška č. 316/2014 Sb., o kybernetické bezpečnosti

Touto vyhláškou se stanovuje obsah a struktura bezpečnostní dokumentace pro informační systém kritické informační infrastruktury, komunikační systém kritické informační infrastruktury nebo významný informační systém, obsah bezpečnostních opatření, rozsah jejich zavedení, typy a kategorie kybernetických bezpečnostních incidentů, náležitosti a způsob hlášení kybernetického bezpečnostního incidentu, náležitosti oznámení o provedení reaktivního opatření a jeho výsledku a vzor oznámení kontaktních údajů a jeho formu.

Při hodnocení rizik se zvažují například tyto hrozby:

- porušení bezpečnostní politiky, provedení neoprávněných činností, zneužití oprávnění ze strany uživatelů a administrátorů,
- poškození nebo selhání technického anebo programového vybavení,
- zneužití identity fyzické osoby,

- nedostatky při poskytování služeb informačního systému kritické informační infrastruktury, komunikačního systému kritické informační infrastruktury nebo významného informačního systému,
- zneužití nebo neoprávněná modifikace údajů,
- odcizení nebo poškození aktiva. (9)

1.1.7. Vyhláška č. 317/2014 Sb., o významných informačních systémech a jejich určujících kritériích

Národní bezpečnostní úřad a Ministerstvo vnitra stanoví podle § 28 odst. 1 zákona č. 181/2014 Sb., o kybernetické bezpečnosti a o změně souvisejících zákonů (zákon o kybernetické bezpečnosti), (dále jen „zákon“):

Touto vyhláškou se stanoví významné informační systémy a jejich určující kritéria, která se člení na dopadová určující kritéria a oblastní určující kritéria.

Dopadovým určujícím kritériem je skutečnost, že

úplná nebo částečná nefunkčnost informačního systému způsobená narušením bezpečnosti informací by mohla mít negativní vliv mimo jiné na provoz jiného významného informačního systému využívajícího služeb hodnoceného informačního systému, který je nefunkční, a dále oběti na životech s mezní hodnotou více než 10 mrtvých nebo 100 zraněných osob vyžadujících lékařské ošetření, s případnou hospitalizací s dobou delší než 24 hodin. (10)

1.2. DICOM

DICOM (Digital Imaging and Communications in Medicine) je mezinárodní standard pro komunikaci a správu obrazových medicínských dat a data s nimi spojená (ISO 12052). Definuje formáty pro obrazová medicínská data, aby mohla být posílána v kvalitě nezbytné pro lékařské účely.

DICOM můžeme najít v každém radiologickém, kardiologickém a i v zařízení pro radioterapii, mezi ně patří například – X-ray, CT, MRI, ultrazvuk atd. Zvyšuje se však využití i v dalších oblastech lékařství, například v očním a zubním lékařství.

S desítkami tisíc zobrazovacích zařízení v provozu a s miliardami lékařských snímků se stává DICOM jedním z nejrozšířenějších zdravotních standardů po celém světě. (11)

1.2.1. Historie vzniku DICOM

Když bylo poprvé představeno CT společně s dalšími digitálními diagnostickými zobrazovacími metodami a se stále rostoucím využíváním počítačů pro klinické aplikace, ACR (American College of Radiology) a NEMA (National Electrical Manufacturers Association), vyvstala nutně potřeba vytvoření standardu pro přenos snímků a s nimi souvisejících informací, a to i mezi zařízeními od různých výrobců.

ACR a NEMA tedy vytvořili společný výbor v roce 1983, aby vytvořili standard, který by:

- podporoval komunikaci digitálních obrazových dat bez ohledu na výrobce zařízení,
- usnadnil rozvoj a rozšíření archivace obrazu a komunikačních systémů (PACS), které mohou také komunikovat s jinými systémy nemocničních informací,
- umožnil vytvoření diagnostických informačních databází, které mohou být čteny velkým rozsahem geograficky distribuovaných zařízení.

Od prvního zveřejnění v roce 1993, způsobil DICOM revoluci v praxi radiologie, kdy bylo možné vyměnit X-ray filmy za plně digitální workflow.

Ať už u oddělení urgentního příjmu, u srdečního zátěžového testování nebo u detekce rakoviny prsu, je DICOM standard, který usnadňuje práci při komunikaci s medicínskými daty a tím usnadňuje práci lékařům.

1.2.2. Základní části standardu DICOM

DICOM standard se původně skládal z 20 základních částí, avšak část PS3.9 a část PS3.13 byly postupem času odstraněny. (12)

PS 3.2 Shoda

V této části standardu jsou definovány principy, které musí zařízení nebo informační systém splňovat, aby dosáhl shody se standardem.

- Požadavky na shodu – část PS3.2 specifikuje obecné požadavky, které musí být v procesu implementace splněny. Konkrétní požadavky pro jednotlivé funkce, data i příkazy jsou pak uvedeny ve specifických částích standardu.
- Prohlášení o shodě – část PS3.2 definuje strukturu dokumentu Prohlášení o shodě. Specifikuje informace, které musí být v dokumentu obsaženy, včetně vazeb na konkrétní požadavky uvedené v ostatních částech standardu.

PS 3.3 Definice informačních objektů

V této části standardu jsou specifikovány třídy informačních objektů (Information Object Classes), které umožňují realizovat abstraktní definici entit reálného světa aplikovatelnou při komunikaci a přenosu medicínských obrazů a informace s nimi spojené (křivky, strukturalizované nálezy, dávky radiační terapie, atd.). Každá definice třídy informačních objektů je tvořena popisem jejího určení a atributů, pomocí kterých je definice realizována.

Standard rozlišuje dva typy tříd informačních objektů:

- Normalizované třídy informačních objektů – obsahují pouze atributy, které jsou vlastní reprezentované entitě reálného světa.
- Kompozitní třídy informačních objektů – mohou obsahovat i atributy, související s entitou reálného světa, které nejsou vlastní (cizorodé).

Kompozitní třídy informačních objektů udávají strukturalizovaný rámec pro realizaci komunikačních požadavků pro zajištění úzké vazby mezi obrazovou informací a informacemi s nimi souvislými.

PS 3.4 Specifikace servisních tříd

Tato část definuje řadu servisních tříd. Servisní třída spojuje jeden nebo víc informačních objektů s jedním nebo více příkazy, které nad těmito informačními objekty mají být vykonány.

Mezi servisní třídy například patří:

- management tisku,
- uložení informací,
- dotaz/opověď,
- základní management worklistu,
- management pacienta,
- management výsledků.

PS 3.5 Datové struktury a kódování.

V této části se specifikuje vytváření a kódování datových souborů (Data set) DICOM aplikací, které vycházejí z užití informačních objektů a servisních tříd. V části se také specifikuje, jaké jsou použité kompresní techniky.

PS 3.6 Datový slovník

V této části se specifikuje centrální registr DICOM datových elementů a jejich definic. Datové elementy představují základní entitu reprezentované informace, včetně jejich unikátní identifikace v rámci standardu DICOM.

Každý datový element je specifikován:

- jednoznačným tagem, tvořeným z čísla skupiny a z čísla elementu,
- jménem,
- hodnotou multiplicity (číslo, udávající kolik hodnot může být zakódováno do datového elementu),
- typem hodnoty (integer číslo, řetězec znaků, atd.),
- každý unikátní identifikátor je specifikován - složením ze dvou částí, které jsou odděleny tečkou.
    - <org root> - unikátní číselná hodnota pro organizaci
    - <suffix> - unikátní číselná hodnota v rámci organizace

Příklad:

UID = <org root>.<suffix>

PS 3.7 Výměna zpráv

Tato část specifikuje služby a protokoly používané aplikacemi medicínských zobrazovacích metod při výměně zpráv v rámci DICOM komunikace. Tyto zprávy jsou složeny z posloupnosti příkazů a z navazujícího datového streamu.

PS 3.7 dále udává:

- operace a informace o stavu (nebo případné změně stavu) entity (DIMSE služby – DICOM Message Service Element), které jsou k dispozici jednotlivým třídám služeb definovaných v části PS 3.4,
- pravidla pro ovládání příkazů realizujících komunikaci na principu požadavek/odezva,
- pravidla pro navázání a ukončení spojení zajišťovaného komunikačními službami,
- kódovací pravidla nezbytná pro tvorbu posloupností příkazů a zpráv.

PS 3.8 Podpora síťové komunikace pro výměnu zpráv

V této části se specifikují komunikační služby a protokoly nejvyšší komunikační vrstvy nezbytné pro komunikaci mezi DICOM aplikacemi, které zajišťují, aby komunikace byla prováděna efektivně a koordinovaně v daném síťovém prostředí. Uvedená specifikace služeb vrchní komunikační vrstvy (Upper Layer Service) je podmnožinou služeb zajišťovaných sedmivrstvovým komunikačním modelem ISO/OSI. Její definice specifikuje použití protokolu DICOM horní vrstvy ve spojení s TCP/IP transportním protokolem.

PS 3.10 Paměťová média a formát souboru pro výměnu médií

Tato část specifikuje obecný model ukládání medicínských obrazových dat na výměnných médiích. Hlavním účelem této části je poskytnout rámec umožňující vzájemnou výměnu různých typů medicínských obrazových dat i s nimi souvisejícími informacemi na různé typy paměťových médií.

PS 3.11 Aplikační profily paměťových médií

V této části se specifikuje aplikační podmnožina DICOM standardu, pro kterou implementace může dosáhnout shody. Takovéto prohlášení shody je aplikováno na funkčnost procesu výměny medicínských obrazových dat a s nimi souvisejícími informacemi na paměťových médiích pro specifické klinické využití.

PS 3.12 Formáty medií a fyzická média pro výměnu medií

Tato část podporuje a usnadňuje výměnu informací mezi medicínskými aplikacemi a specifikuje:

- charakteristiku specifických fyzických medií a jejich formátů,
- strukturu pro popis vzájemných vztahů mezi obecným modelem paměťových medií a specifickými fyzickými médii a jejich formátem.

PS 3.14 Zobrazovací funkce standardní stupnice šedi

V této části se specifikují standardizované zobrazovací funkce, které jsou nezbytné pro konzistentní zobrazování obrazových dat založených na stupnici šedi. Zobrazovací funkce poskytují metody kalibrace konkrétních zobrazovacích systémů, umožňující zajistit konzistentní prezentaci obrazových dat na různých mediích (displeje, tiskárny, atd.). Zobrazovací funkce jsou založeny na lidském vizuálním vnímání (Bartenův model).

PS 3.15 Bezpečnostní a systémové profily managementu

V této části se specifikuje bezpečnost systémů DICOM standardu a pravidla řízení přístupu k datům, která musí být dodržena pro dosažení shody aplikace se standardem. Tu obstarávají obecně uznávané protokoly, jako jsou například DHCP, LDAP, TSL a další.

PS 3.16 Mapování obsahových zdrojů

Tato část DICOM standardu specifikuje, jaké návrhy formátů strukturovaných dokumentů DICOM informačních objektů lze používat. Dále také uvádí množinu kódovaných termínů, které jsou využívány informačními objekty a též překlady kódovaných termínů specifických pro jednotlivé země.

PS 3.17 Vysvětlivky

Část PS 3.14 standardu DICOM obsahuje rozsáhlé dodatečné vysvětlivky k předchozím částem. Ostatní části se na ní taktéž odkazují.

PS 3.18 Webový přístup k DICOM objektům (WADO)

V této části je specifikováno, jaké prostředky umožnují realizaci požadavku na povolené DICOM objekty ve formátu http URL/URI (Uniform Resource Locator/

Uniform Resource Identifier). Požadavek musí obsahovat směrník, který odkazuje na příslušný známý a definovaný DICOM objekt ve formě konkrétního UID.

1.3. PACS

PACS je technologie, která se používá ve zdravotnictví a která umožňuje spravovat, archivovat (ukládat) a zobrazovat obrazovou dokumentaci. Mezi obrazovou dokumentaci řadíme snímky z rentgenu, centrálního tomografu, magnetické rezonance apod.

Dicom (Digital Imaging and Communications in Medicine) se zde používá jako standard a jako univerzální formát komprimovaných obrazových dat.

PACS se skládá ze čtyř částí, kterými jmenovitě jsou: obrazová dokumentace, zabezpečená síť, cílová stanice (terminál, počítač) a úložiště dat.

Systémy Picture Archiving and Communicating System (PACS) se používají ve zdravotnictví jako podsystém nemocničního IS (informačního systému).

PACS jsou typické tím, že zpracovávají velké množství dat. Tahle objemově náročná data vznikají většinou ve specializovaných přístrojích, jako jsou magnetická rezonance – MR (magnetic resonance), centrální tomograf – CT (computed tomography), angiograf nebo arteriograf (XA , X-ray angiography) a další. Data, jež vznikají na těchto modalitách (digitální diagnostická zařízení používaná ve zdravotnictví pro telemedicínské sledování pacienta) mohou mít pro jednoho pacienta velikost řádově až několik gigabytů.

Můžeme tudíž říci, že PACS je v podstatě systém, který uchová obrazové informace vzniklé na digitálních diagnostických zařízeních, která se využívají ve zdravotnických zařízeních a která pracují na různých principech zobrazovacích metod. Zařízení jsou schopna komunikovat a pracovat na základě standardu DICOM.

Výsledný systém může pracovat s většími objemy dat (řádově GB) a je databázově orientovaný. Aby mohl takový systém fungovat, musí mít vlastnosti, jaké nemají klasické databázové systémy, jelikož je nepotřebují. Na druhou stranu nesmí mít vlastnosti, které by mohly v konečném důsledku (zejména při naplnění daty) databázový systém brzdit. Při vyšetření centrálním tomografem se udělají stovky

až tisíce obrázků o rozlišení 512×512 nebo 1024×1024 bodů a bitové hloubce šedi cca 10-14 bitů. U magnetické rezonance se používá dokonce ještě menší rozlišení (256×256 bodů), avšak při jednom vyšetření může množství vytvořených obrázků dosáhnout množství až několik desítek tisíc.

V angioskopii vznikají dokonce klasické filmy a snímky jsou dynamické. Dočetl jsem se, že v případě konverze (změny) klasického videosignálu do DICOM vznikne z jednotlivých obrázků klasické video s náležitým HD rozlišením. Soubor, který vznikne, může mít i několik gigabytů dat. (13)

DICOM, jak již bylo zmíněno, je standard, který popíše, přenese a uloží informace vzniklé na modalitách. DICOM využívají všechny zobrazovací modality, mimo již uvedených – skiagrafie, mamografie, výpočetní tomografie (CT), magnetická rezonance (MR), ultrazvuk, pozitron emisní tomografie atd.

Pokud jsou obrazová data v analogové formě, musíme je pro přenos do PACSu digitalizovat. Archivace fyzických snímků, která se označuje jako hard-copy je v tomto případě nahrazena tzv. soft-copy (archivace digitálních dat). V praxi to funguje tak, že si lékař otevře určitá data (např. CT řez či MR sken) v pracovní stanici pomocí DICOM prohlížečů, kterých je velký výběr (xVision, OsiriX, MicroDicom, ImageJ, Dicompass, Irfan View…) a které umožňují různě pracovat s daty. Například nám umožní nastavení kontrastu, jasu, spektra barev, měření délek zkoumaného objektu atd. Některé prohlížeče jsou volně stažitelné a jiné placené. Liší se i tím, zda jsou pro Windows, Apple nebo pro unixový systém.

Někdy se můžeme setkat i s názvem teleradiologie, což je obor telemedicíny přímo zaměřený na přenos záznamů CT, rentgenových snímků a magnetické rezonance.

DICOM mimo jiné popisuje, jak má být komprimována (zhuštěna) určitá obrazová informace, jakých má být použito metod a to jak na straně lékařského zařízení (modality), tak na straně stanice.

Každý obrázek má dvě základní části. První část popisuje vlastní bitmapu a blok s DICOM tagy. To jsou textová pole s informacemi o pacientovi, lékařském zařízení (modalitě) a ostatních údajích. Identifikátorem rozumím výraz, který odlišuje jednotlivé entity (objekty reálného světa, které jsou zachyceny v datovém modelu)

a které patří do stejné třídy objektů. Identifikátor se obvykle skládá z čísel, znaků nebo je to kombinace obojího.

Tyto identifikátory popisují ojedinělost každého snímku a rovněž určují jeho příslušnost ve strukturální hierarchii. Jediné, co „reálně“ existuje, jsou vlastně jen jednotlivé snímky, které jsou uspořádávány do určitých sérií, série pak dále do studií nebo vyšetření a vyšetření do pacienta.

Můžeme říct, že informace z vyšetření se tak ukládají v PACS systémech do čtyř úrovňové hierarchie. Teď se pokusím vysvětlit, jak soubory s velkým objemem dat souvisí s databázovými systémy a zda je možné je klasicky používat.

Pojem Binary large object (blob) databáze znají a jsou schopny takový objekt nejenom uložit, ale také ošetřit. Problém je v tom, že různé databáze nakládají se svými daty různě, mají různou vybavovací dobu, různou rychlost obměn dat a podobně. Přístup je odlišný u jednotlivých výrobců PACS systémů.

Někdy je přístup nevhodně limitující neboť může dojít k odstranění položek, které databáze nezná. Extrémem je úplné rozebrání hlaviček DICOM souborů na základní části. Pak dojde k jejich uložení do databázových tabulek s odděleně uloženou bitmapovou částí.

Může dojít i k odstranění privátních tagů (kam jsou ukládány informace od pracovních stanic), které mají sloužit k popisu specializovaných datových obrazovkových výstupů.

Dalším extrémem je uchování celého obrázku v původním tvaru s tím, že v pomocné databázi zůstanou uchovány pouze informace nutné k jeho vyhledání či jen k analýze dat apod.

Každá databáze funguje bezproblémově, když se do ní data vkládají. Při získávání dat nazpět jsou již nutné určité kroky. Například definovat délku odezvy na dotaz podání nebo vybavení dat. A to je hůře realizovatelné zejména pak při velkém objemu dat.

Nejvýhodnější je uchovávat data v nezměněné podobě, ať už z důvodů možnosti změn anebo nemožnosti řídit upgrade konzolí u modalit atd.

DICOM v neposlední řadě umožňuje používání tzv. privátních DICOM tagů či objektů a ty jsou v krajním případě podporovány jen jedním výrobcem. I přesto je nutné, aby byly soubory se soukromými informacemi jak uchovatelné, tak vyvolatelné z klasických DICOM archivů.

U většiny PACS dochází k problému v okamžiku, kdy je jejich kapacita k ukládání dat překročena anebo je archivováno příliš velké množství dat. Aby byl PACS spolehlivý a dostatečně funkční, musí se najít správné parametry jeho vnitřního nastavení a uspořádání.

Určitě budou jiné požadavky na systém pro pracoviště s jednou modalitou a na specializované pracoviště s exponovaným provozem (např. traumatologie). Tam bude třeba uvažovat o jiné digitalizaci.

2. Analýza dostupných metod anonymizace a návrh algoritmu

S rostoucím vlivem sociálních sítí vznikla i poptávka po anonymizaci obrazových dat, z toho se nejčastěji týkala anonymizace fotek. V dnešní době existuje hned několik metod pro anonymizaci. Mezi časté patří tzv. browserové anonymizátory, které umožňují anonymizaci kdekoliv, kde je zajištěn přístup k internetu (tedy bez nutnosti instalace anonymizéru). Jelikož pracuji pouze s obrazovou částí souboru DICOM, nepotřebuji anonymizér formátu DICOM, tudíž jako dostatečné lze považovat pouze anonymizér obrazového souboru, např. fotek.

Facepixelizer

Patří mezi tzv. browserové anonymizátory.

Facepixelizer umožnuje po nahrání fotky na web zvolit jakou část obrazového souboru chceme anonymizovat. Může se tedy jednat, jak již název napovídá, o obličej nebo i obyčejný text. Z hlavní části slouží tato metoda pro anonymizaci obličejů. Jelikož však poskytuje i možnost anonymizace jakékoliv části obrazového souboru (obličeje, texty, předměty, atd.), rozhodl jsem se metodu zmínit.

Metoda nabízí automatickou a manuální anonymizaci obrazových souborů. Automatická anonymizace nalezne všechny obličeje a automaticky je anonymizuje. Druh anonymizace je nastaven na pixelizaci a ta je odlišná podle velikosti anonymizované části.

Manuální metoda nabízí možnost výběru ze tří druhů anonymizace:

- pixelizace,
- rozostření,
- začernění.

Další funkcí, kterou anonymizátor obsahuje, je oříznutí a změna velikosti obrazového souboru. Dále je možné změnu provedenou v anonymizátoru vrátit tlačítkem **Revert**. Anonymizujeme-li část omylem, není třeba zadávat obrazový soubor znova, což je dle mého názoru nespornou výhodou oproti konkurenčním programům.

Mezi výhody této metody patří automatická detekce obličejů a následná anonymizace. Detekce však není dokonalá a nedokáže například rozeznat obličej z profilu nebo nakloněný. Autodetekce, ač je užitečná, však není potřebná pro problém anonymizace osobních údajů v obrazové části souborů DICOM, protože se jedná o snímky z vyšetření, nikoliv o fotky. Dále metoda odstraňuje všechna **Exif** data u **JPG** souborů (místo pořízení, čas, atd.). Cílovým formátem není JPG, ale PNG, takže tato funkce není potřebná.

Facepixelizer je bezplatná služba. Více na (14)**.**

ZORRO Anonymizace

Jedná se o browserovou metodu společnosti Atbon a.s. na anonymizaci dokumentů a začerňování objektů.

Po nahrání souborů do webové aplikace lze přes filtr vyhledat předdefinované vzory (např. rodné číslo). Aplikace nabízí anonymizovat všechny části naráz nebo postupně. Po stisknutí tlačítka Redigovat se zobrazí dokument se šablonou pro anonymizaci obrazového souboru, který byl automaticky vytvořen. Dopředu vyhledané texty pro anonymizaci lze jednoduše zrušit a požadované texty, které nebyly nalezeny automaticky v aplikaci, je poté možné přidat pomocí vykreslení obdélníku.

Pro získání finální verze anonymizovaného obrazového souboru je potřeba stisknout tlačítko Vytvořit redigovaný dokument, které nabídne uložení do počítače nebo zobrazení obrazového souboru v prohlížeči.

Podle referencí je tato metoda hojně využívaná krajskými i městskými úřady České republiky. Více informací zde (15).

PhotoHide

Tato metoda patří mezi další browserové aplikace. Oproti konkurenci nenabízí tolik možností úpravy.

Po nahrání obrazového souboru lze táhnutím myši vyznačit plochu pro anonymizaci. Anonymizace dat je prováděna pixelizací. Velikost pixelizace nelze změnit.

Každý pokus o anonymizaci textu se potvrzuje tlačítkem HIDE. Jedná se o sériový proces, takže nelze anonymizovat více než jeden rámec textu naráz, což je značně nepraktické. Nemá ani funkci vrácení úpravy, tudíž pokud dojde k anonymizaci špatné části, je nutné celý proces s nahráním obrazového souboru opakovat.

Po dokončení anonymizace jednotlivých částí lze anonymizovaný soubor stáhnout tlačítkem Download. Více na (16).

Lionytics Image Anonymizer

Tato metoda obsahuje stažitelný program, který anonymizuje obrazové soubory přímo na lokálním počítači. Program nabízí automatickou i manuální anonymizaci.

Při anonymizaci lze vybrat prefix pro anonymizované obrazové soubory, tvar anonymizačního okénka (elipsa, obdélník, atd.) a také typ výstupního souboru.

Aplikace stojí 99$, což je aktuálně v přepočtu cca 2400Kč. Více na (17).

Vyhodnocení

Ačkoli jsou browserové anonymizátory užitečné, nejsou vhodným řešením pro anonymizaci citlivých dat obsažených v obrazové části DICOM souborů. Obrazové soubory by měly být anonymizovány diskrétně, proto považuji odesílání obrazového souboru na neznámý web za bezpečnostně neakceptovatelné řešení. Z tohoto důvodu jsem se rozhodl vytvořit vlastní anonymizační program, který bude splňovat požadavky ideálního anonymizéru, tj. bude spouštěn lokálně a bude bezplatný.

Návrh algoritmu

Ideální anonymizační program by měl, dle mého názoru, splňovat následující podmínky a vlastnosti:

Být lokální, tj. data by neměla opustit počítač, na kterém je prováděna anonymizace.

Být automatický, tj. anonymizér by neměl vyžadovat příliš mnoho operací přímo od uživatele (ideálně žádné úkony od spuštění), což je velice výhodné, zpracovává- li se velké množství obrazových dat.

Být rychlý, tj. doba strávená anonymizací jednoho obrazového souboru by měla být co nejmenší, což je opět výhodné, zpracovává-li se velké množství obrazových dat.

Být upravitelný, tj. anonymizér by měl nabízet možnosti modifikace, ať už se jedná o možnost přidání a také úpravy usnadňujících funkcí, například grafického rozhraní. Také by měl skýtat např. možnost úpravy formátu zápisu do logovacího souboru. Ve free anonymizéru nebo v placeném takové možnosti nebývají.

Být výhodný, mám teď na mysli z finančního pohledu, tedy čím levněji, tím lépe. Nejlepší by byl samozřejmě zadarmo, nikoliv však na úkor kvality. Kvalita by měla být srovnatelná s placenými anonymizéry.

Být přesný, to znamená, že rozeznávání znaků by mělo být co nepřesnější. Přesnosti lze docílit úpravou rozeznávání řetězců (Tesseract-Ocr) i v anynomizaci. Mám na mysli případné trénování Tesseract-Ocr na obrazových souborech určených pro výzkum nebo na vylepšení kódu programu, pro spolehlivější ošetření citlivých dat určených k anonymizaci.

Po pečlivé analýze bodů uvedených výše jsem došel k závěru, že ideálním řešením bude vytvoření vlastního anonymizačního programu, který by měl dodržet všechny podmínky ideálního anonymizéru.

3. Analýza anonymizace obrazových dat

Modernizace způsobu nakládání se zdravotnickými obrazovými daty a neustálé vylepšování dostupné technologie pro jejich pořizování, zapříčinily nutnost dalšího zpracování medicínských dat. V této práci je řešen důležitý úkol a to odstranění zbytkových citlivých dat z obrazových souborů. Po odstranění citlivých dat jsou obrazové části DICOM souborů použitelné i mimo zdravotnická zařízení, např. pro účely výuky nebo statistiky.

V další části mojí bakalářské práce uvedu postupně metody a knihovny, které jsem při anonymizaci použil.

3.1. Metody hledání textových řetězců

Abych mohl citlivá data vymazat, potřebuji je nejprve v souboru najít. Jelikož však předem nevím přesné řetězce, hledám pouze řetězce splňující určitá kritéria.

Algoritmus hrubé síly

Algoritmus postupuje tak, že pro každou pozici v textu kontroluje, jestli v ní nezačíná hledaný řetězec. Složitost hledání v běžném textu je O(m+n).

Jelikož nevíme, jaké řetězce hledáme, musí existovat určitá pravidla, podle kterých se při hledání řetězců řídíme. Algoritmus hrubé síly byl použit při kontrole řetězců, které anonymizovat nepotřebují, avšak splňují podmínky pro anonymizaci (např. všechny znaky jsou velké). Vytvořil jsem tudíž sérii řetězců, které při shodě s nalezeným textem (který splnil podmínky pro anonymizaci) nejsou považovány za citlivé informace nutné k anonymizaci.

3.2 Použité nástroje

Abych docílil splnění zadání své bakalářské práce, potřeboval jsem využít různé škály open source nástrojů. V následujících kapitolách vysvětlím, jaké open source nástroje jsem použil a také je stručně popíši.

Na úpravu obrazových souborů jsem použil **ImageMagick**, distribuovaný pro nejrůznější operační systémy. Po úpravě obrazového souboru jsem použil open

source engine **Tesseract-ocr** pro naskenování dat z obrazového souboru a jejich následné uložení do příslušného formátu.

Jakožto vhodný formát pro výstupní data z **Tesseract-ocr** jsem zvolil formát **HOCR**, který je odnoží hypertextových formátů (**HTML**), konkrétně **XHTML**. Důvodem zvolení **HOCR** formátu bylo, že tento výstupní formát uchovává i pozice nalezeného textu, které jsou nezbytné pro jeho anonymizaci. Textový výstup nebo výstup **PDF** informace o poloze textu neuvádí.

Pro parsování formátu **HOCR** jsem ve svém programu využil Java knihovnu zvanou **JSoup**, díky které jsem byl schopen načtená data uložit do pole stringů, se kterým se mi bude lépe pracovat.

Jako vývojové prostředí pro svůj program jsem použil **NetBeans** pro Linux.

3.2.1. Tesseract-Ocr 3.03

Tesseract je pravděpodobně jedním z nejpřesnějších open source OCR. V kombinaci s Leptonica Image Processing Library je Tesseract schopen přečíst rozsáhlé množství obrazových formátů a dokáže je i převést do více jak 60 různých jazyků. Tesseract byl považován za jeden ze tří nejlepších engine v UNLV testu přesnosti v roce 1995. Od roku 1995 se vývoj Tesseractu zpomalil až do roku 2006, kdy začal být vývoj Tesseractu sponzorovaný společností Google. Nyní je vydán pod Apache Licencí 2.0.

Tesseract podporuje operační systémy:

- Windows
- Linux
- Mac OS X.

Důvodů, proč jsem si vybral Tesseract pro skenování obrazového souboru a upřednostnil ho tak před ostatními Open Source OCR enginy, je hned několik:

- Tesseract je celosvětově považován za jeden z nejlepších OCR enginů.
- Výsledky rozpoznávání znaků byly lepší než u ostatních mnou zkoušených enginů (např. GOCR).
- Tesseract lze trénovat pro lepší výsledky rozpoznávání znaků.

- Možnost výstupu dat ve formátu XHTML.

Výstupním formátem Tesseract může být textový soubor, PDF soubor nebo soubor HOCR. Pro svou bakalářskou práci jsem použil výstupní formát typu HOCR, který mi umožnil získat pozice mnou hledaných slov (stringů).

Přesnost rozlišování znaků se v mém případě pohybovala okolo 80%. Více viz (18).

3.2.2. ImageMagick 6.7.7-10

ImageMagick je balík nástrojů na vytváření, úpravu a zpracování bitmapových obrázků. Dokáže číst a přepisovat rozsáhlou škálu formátů souborů (oficiální stránky udávají přes 200 typů souborů), mezi nimiž jsou i formáty použité při zpracování bakalářské práce, tedy původní formát PNG a mnou převedený formát JPG.

Důvodem použití tohoto nástroje byla nutná úprava obrazových souborů, kde se schopnost enginu Tesseract rozpoznat znaky pohybovala pod hranicí 50%. To jsem považoval za nepřípustné. Viz více na (19).

3.2.3. JSoup

Při programování praktické části své práce jsem použil Java knihovnu JSoup, která slouží k parsování dat z hypertextových typů souborů, v mém případě tedy XHTML – hocr. Více informací je dostupných na (20).

3.2.4. AWT Graphics

K načtení obrazového souboru do programu a k následné anonymizaci citlivých dat jsem se rozhodl použít základní knihovnu AWT Graphics, která je standardně obsažena v Javě.

3.2.5. jTessBoxEditor 1.6

Tesseract-OCR nabízí možnost trénování pro lepší výsledky při rozeznávání textu. Pro trénování jsem použil grafické rozhraní jTessBoxEditor, jež umožňuje export dat zlepšujících proces rozeznávání.

Grafické rozhraní jTessBoxEditor lze spustit téměř na všech operačních systémech. Podporuje formáty TIFF, JPG, PDF, BMP a pro práci důležitý PNG.

Trénování textu v grafickém rozhraní je relativně jednoduché.

Více na (21).

4. Implementace

Práci na anonymizaci jsem rozdělil do dvou částí:

- Úprava obrazových souborů a získání dat pro anonymizaci obrazových souborů.
- Anonymizace obrazových souborů.

Program sám obstarává obě dvě části.

Vývoj programu jsem implementoval v operačním systému Linux.

Program spouštím pomocí příkazu:

java –jar program.jar properties.properties

První parametr označuje properties soubor, ve kterém jsou uloženy všechny potřebné proměnné. Mezi potřebné proměnné patří cesta, název původního obrazového souboru pro anonymizaci, soubor obsahující tzv. „bílá slova“. „Bílá slova“ jsou slova, která mohou splňovat parametry pro anonymizaci, avšak nebudou anonymizována - např. názvy přístrojů (Phillips, SIEMENS, atd.). Dále také cesta k výstupnímu adresáři, kam bude uložen logovací soubor a anonymizovaný obrazový soubor. Do výstupní složky budou taktéž uloženy soubory potřebné pro anonymizaci, ty však budou před ukončením programu smazány.

4.1. Výběr programovacího jazyka

Pro zpracování požadavků bakalářské práce jsem se rozhodl použít programovací jazyk Java a to nejenom proto, že je jedním z nejrozšířenějších programovacích jazyků, ale také proto, že mám s Javou nejvíce zkušeností získaných při studiu.

4.2. Třídy programu

Program jsem rozdělil do 3 tříd:

- Hlavni
- Obrazek
- Anonymizer

4.2.1. Třída Hlavni

Tato třída slouží jako hlavní a spouštěcí třída.

V hlavní metodě je vytvořen objekt anonymizer a načten soubor properties.properties, který obsahuje potřebné údaje pro anonymizaci obrazového souboru (tj. název a cestu obrazového souboru, cestu do adresáře určeného pro zápis a cestu k souboru obsahující tzv. bílá slova).

Po vytvoření objektu anonymizer je pak zavolána metoda Anonymizuj(), která anonymizuje obrazový soubor.

4.2.2. Třída Anonymizer

Třída Anonymizer obsahuje objekt Anonymizer, který je složen z následujících proměnných:

- nazev – název obrázku (může obsahovat i cestu k obrázku),
- cesta – cesta pro výstupní soubory,
- text – cesta k textovému souboru, který obsahuje bílá slova.

Třída obsahuje několik metod, které postupně vysvětlím:

- Anonymizuj()
- ZjistiJmeno()
- ZjistiCislo()
- ZjistiDatum()
- AnonymizujJmeno()
- AnonymizujCislo()
- AnonymizujDatum()
- Prepocti()
- Zapis()
- Smaz()

Metoda Anonymizuj()

Tato metoda je nejdůležitější metodou v třídě Anonymizer. V metodě se vytvoří objekt obrazek s příslušnými parametry. Dále se přiřadí soubor hOCR z metody Ocr

z třídy Obrazek. Soubor hOCR se parsuje pomocí knihoven Jsoup na formát UTF- 8. Je vytvořen Buffered Writter na zapisování do logovacího souboru a také scanner pro načítání tzv. bílých slov z textového souboru.

Metoda postupně načítá každé slovo ze souboru hOCR s jeho souřadnicemi. Nesplňuje-li žádné slovo pravidla pro anonymizaci, výsledek se zapíše do logovacího souboru.

Metoda ZjistiJmeno()

Tato metoda zjišťuje, zdali slovo nalezené v souboru hOCR odpovídá určeným pravidlům. Slovo se také kontroluje se seznamem tzv. bílých slov z textového souboru. Pokud metoda najde slovo, které není v seznamu bílých slov a zároveň odpovídá určeným pravidlům, je zavolána metoda na anonymizaci nalezeného slova, tedy metodu AnonymizujJmeno().

Metoda ZjistiCislo()

Metoda slouží ke zjištění rodného čísla. Je-li nalezen jiný znak než číslo, metoda vyhodnotí, že není potřeba anonymizace. Nalezne-li metoda číslo odpovídající pravidlům, zavolá metodu AnonymizujCislo().

Metoda ZjistiDatum()

Kromě rodného čísla se v obrazovém souboru může také vyskytnout datum narození pacienta. Tato metoda zjišťuje, jestli se taková data nevyskytují v obrazovém souboru. Je-li nalezeno datum narození, metoda zavolá metodu AnonymizujDatum() pro anynomizaci.

Metoda AnonymizujJmeno()

Tato metoda nejdříve zavolá metodu Zapis(), dále zjistí potřebné údaje k anonymizaci, jako jsou výška a šířka obrazového souboru určeného k anonymizaci. Po získání potřebných údajů se zavolá metoda pro vykreslení obdélníku s odpovídajícími souřadnicemi.

Metoda AnonymizujCislo()

V této metodě je zavolána zapisující metoda Zapis() a následně jsou vypočteny souřadnice pro vykreslení obdélníku do obrazového souboru.

Metoda AnonymizujDatum()

Tato metoda zavolá metodu Zapis() pro zapsání souřadnic do logovacího souboru a poté se „vykreslí“ pomocí metody Vykresli() čtverec do obrazového souboru.

Metoda Prepocti()

Tato metoda slouží pro přepočtení souřadnic z upraveného obrazového souboru na souřadnice obrazového souboru určeného k anonymizaci. K přepočtení souřadnic je použita matematická trojčlenka.

Metoda Zapis()

V této metodě se zapíše nalezené slovo do logovacího souboru. Souřadnice nalezeného slova určeného k anonymizaci jsou vytaženy z pole. Metoda po získání souřadnic slova, určeného k anonymizaci, zavolá metodu Prepocti(). Po přepočtení původních souřadnic jsou nové souřadnice zapsány do logovacího souboru.

Metoda Smaz()

Tato metoda slouží pouze ke smazání souborů vytvořených v průběhu anonymizace. Jedná se o zvětšený obrazový soubor **JPG** a soubor typu **hOCR**.

4.2.3.Třída Obrazek

Třída Obrazek obsahuje objekt Obrazek, který se skládá z proměnných:

- nazev – název obrázku (může obsahovat i cestu k obrázku),
- cesta – cesta pro výstupní soubory.

Třída obsahuje tři metody, které jsou následně vysvětleny:

- Ocr()
- Vykresli()
- Uloz()

Metoda Ocr()

Tato metoda obsahuje dva procesy, které jsou důležité pro anonymizaci obrazových dat. První proces je úprava obrazových souborů a druhý je spuštění Tesseractu za

účelem vytvoření hOCR souboru. Jde o metodu s návratovou hodnotou File, tedy vrací zpátky soubor hOCR do třídy Anonymizer.

Metoda Vykresli()

Tato metoda vykreslí do grafiky **g** černý obdélník podle parametrů, které se předají při zavolání metody.

Metoda Uloz()

Metoda Uloz() slouží jako finální úprava obrazového souboru. V této metodě je vytvořený nový soubor s předponou „anon_“ což značí, že se jedná o již anonymizovaný soubor. K souboru je pak přiřazen upravený obrazový soubor ve formátu PNG.

4.3. Postup při tvorbě programu

V této kapitole budou podrobněji popsány metody, proměnné a také postup při tvorbě anonymizačního programu.

4.3.1. Úprava obrazových souborů

V bakalářské práci byl ImageMagick použit pro změnu rozlišení původního obrazového souboru. Důvodem této změny byla příprava souboru pro práci s programem Tesseract. Zjistil jsem totiž, že na obrazových souborech, které mně byly pro práci přiděleny, není úspěšnost rozeznávání písmen a číslic zdaleka tak uspokojivá. Na oficiálních stránkách Tesseractu jsem se dočetl, že změnou velikosti rozlišení se zlepší i kvalita načítání textu, viz tabulka č. 1. K dosažení svého cíle jsem tedy použil **– resize** z knihovny ImageMagick.

Pomocí příkazu – resize jsem zkoušel zadat několik hodnot a porovnával jsem, z jaké hodnoty dostanu nejlepší výsledek – tedy poměr velikosti souboru a schopnosti Tesseractu rozeznat vyžadované znaky.

Tabulka 1: Porovnání velikostí a úspěšnosti rozeznávání textu pro formát PNG

Rozlišení v pixelech	Velikost	Ukázka textu
Originální (980x980)	500 kB	Fakunm nemacmce men

5000x5000	5 000 kB	Fakuitni nemocnice Plzen
10000x10000	14 000 kB	Fakuitni nemoonioe Plzen

Z výše uvedené tabulky je jednoznačně vidět velký rozdíl mezi originálním obrazovým souborem a mnou upraveným souborem.

Rozdíl mezi rozlišením 5000 a 10000 není už tak razantní, avšak lepší hodnoty jsem dostával z rozlišení 5000. Při rozhodování záleželo také na velikosti souboru a rychlosti jeho zpracování. Rozhodl jsem se tedy každý upravovaný obrazový soubor nejdříve změnit na 5000x5000 pixelů.

Formát PNG však ve velkém rozlišení nabývá též velké velikosti (viz tabulka výše) která zpomalovala chod programu až o několik vteřin. Rozhodl jsem se tedy při změně rozlišení také změnit formát souboru, v mém případě jsem zvolil formát JPG. Dále jsem testoval, zdali změna formátu neovlivní úspěšnost načítání znaků (viz tabulka č. 2).

Tabulka 2: Porovnání velikosti a úspěšnosti rozeznávání textu pro formát JPG

Rozlišení v pixelech	Velikost	Ukázka textu
Originální (980x980)	500 kB	Fakunm nemacmce men
5000x5000	2 000 kB	Fakultni nemocnice Plzen
10000x10000	5 000 kB	Fakuttni nemocnioe Plzen

Výše uvedená tabulka potvrzuje, že se mé obavy ohledně snížení rozpoznatelnosti znaků nepotvrdily, ba naopak se v určitých pasážích textu rozpoznávání znaků dokonce zlepšilo a formát JPG dosahoval poloviční velikosti, než formát PNG.

Pro změnu formátu a rozlišení obrazového souboru tedy použiji následující příkaz:

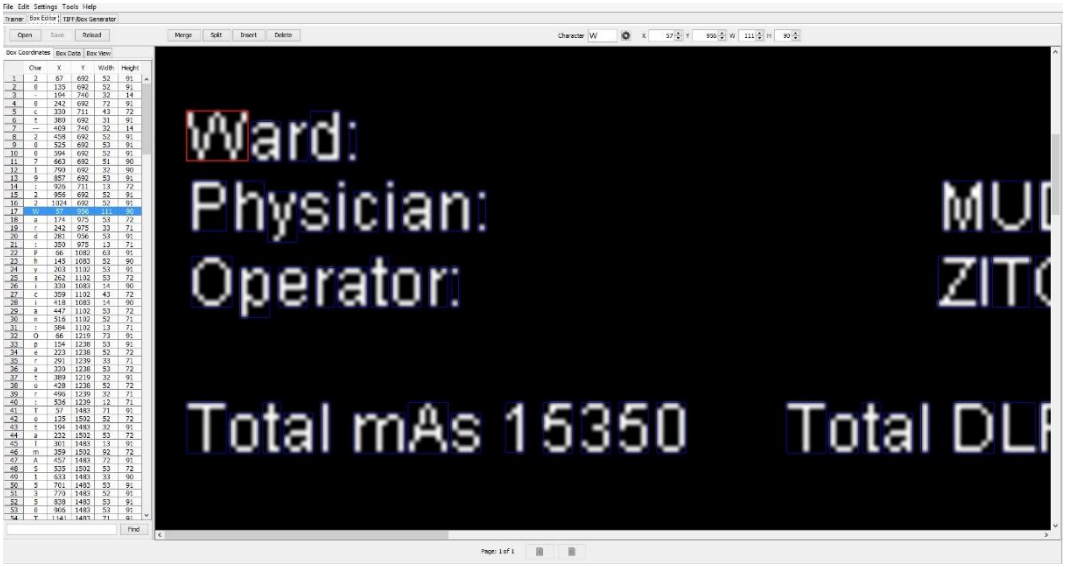
convert " + nazev + " -resize 5000 " + cesta + "pred.jpg –l eng

kde proměnná „nazev“ obsahuje název a cestu k původnímu obrazovému souboru a proměnná „cesta“ obsahuje cestu k adresáři, kam lze zapisovat.

4.3.2. Trénování Tesseract-Ocr

Tesseract-Ocr umožňuje pro přesnější rozeznávání textu možnost trénování na obrazových souborech. Trénování jsem se rozhodl provádět na zvětšených obrazových souborech (pomocí resize), protože program převádí originální obrazové soubory na rozlišení 5000x5000 (kvůli lepšímu rozeznávání). Po načtení obrazových souborů lze zkontroloval a opravit popř. přidat tzv. boxy obsahující rozeznaný znak. K úpravě jsem použil program jTessBoxEditor.

Obrázek 1: Příklad kontroly boxů v jTessBoxEditor viz Obrázek 2



Po zkontrolování a případné opravě boxů, jsem schopen vytvořit trénovaná data.

Program Tesseract očekává soubory pojmenované podle určitých pravidel. Jedno z těchto pravidel je:

[jazyk].[název fontu].exp[číslo]

Mnou vytvořená trénovací data se mohou například jmenovat takto:

eng1.meddata.exp0.png

eng1.meddata.exp0.box

název eng1 značí, že již existuje jazyk eng.

Pro vytvoření trénování je poté nutné zadat posloupnost příkazů:

tesseract eng1.meddata.exp0.png eng1.meddata.exp0 nobatch box.train

unicharset_extractor eng1.meddata.exp0.box

Je také potřeba definovat přidaný jazyk, tedy:

echo "meddata 0 0 0 0 0" > font_properties

kde první hodnota je název jazyku a další jsou binární hodnoty vlastností fontu – italic, bold, fixed, serif, fraktur.

Po definování jazyku je potřeba vytvořit soubou shapetable, v doslovných překladu tabulku tvarů, ten je však pouze potřeba, jedná-li se o Indické jazyky, proto tento ve trénování přeskočím.

Následují příkazy:

mftraining -F font_properties -U unicharset -O eng1.unicharset

eng1.meddata.exp0.tr

a příkaz:

cntraining eng1.meddata.exp0.tr

Nyní jsem získal všechny potřebné soubory pro vytvoření trénovaných dat v Tesseractu (traineddata), avšak předtím než se vytvoří trénovaná data, je potřeba přidat vytvořeným souborům prefix jazyku. Tuto operaci lze provést ručně nebo pomocí série příkazů:

mv inttemp eng1.inttemp

mv normproto eng1.normproto

mv pffmtable eng1.pffmtable

v případě použití tabulky tvarů také (mv shapetable eng1.shapetable).

Po přidání prefixu zbývá už jen dva poslední kroky. První krok je vytvoření trénovaných dat pomocí příkazu:

combine_tessdata eng1.

a druhým krokem je přesunutí trénovaných dat do příslušného adresáře s daty pro spuštění, např.:

sudo cp eng1.traineddata /usr/local/share/tessdata/

Tesseract-Ocr také umožňuje spuštění s více jazyky najednou:

tesseract obrazek.png output -l eng+eng1

kdy bude použit jazyk eng i jazyk eng1.

4.3.3. Spuštění Tesseract a získání hOCR souboru

Po převedení obrazového souboru je dalším krokem zjištění citlivých dat. Program Tesseract naskenuje obrazový soubor a uloží jej do mnou zvoleného výstupního formátu. Program Tesseract používá jako výchozí výstupní formát typ TXT, ten je však pro mojí práci nedostačující, jelikož neobsahuje pozice nalezeného textu.

Zvolil jsem tedy výstup ve formátu hOCR. Jedná se o soubor typu XHTML.

hOCR soubor jsem získal pomocí příkazu:

tesseract " + cesta + "pred.jpg " + cesta + nazevObrazku + "
hocr –l eng

kde proměnná „cesta“ značí adresář k upravenému JPG obrazovému souboru. Proměnná „nazevObrazku“ obsahuje název originálního obrazového souboru, který je určen k anonymizaci.

Výstupní soubor je ve formátu hOCR. Soubor obsahuje veškerý nalezený text v upraveném obrazovém souboru a také pozice nalezeného textu.

Příklad celého procesu níže.

tesseract /tmp/prej.jpg /tmp/obraz.png hocr

Příkaz **convert** i příkaz **tesseract** si program sám volá automaticky při spuštění.

Výstupem jsou dva soubory:

- pred.jpg
- data.hocr

Oba soubory jsou ukládány do předem zvoleného adresáře ze souboru properties a po skončení programu jsou automaticky smazány.

4.3.4. Nalezení citlivých dat v obrazových souborech

Po získání hOCR souboru a upraveného obrazového souboru jsem připraven na anonymizaci citlivých dat v původním obrazovém souboru.

Nejdříve je však potřeba dostat načtené údaje uložené v souboru hOCR a abych toho docílil, použil jsem následující metody z knihoven JSoup.

```
org.jsoup.nodes.Document doc = Jsoup.parse(input, "UTF-8");
```

Tato metoda uložila celý obsah HOCR souboru do dokumentu **doc**, ze kterého pak budu načítat mnou požadovaná data.

Načítání dat probíhá v cyklu **for** a pro uložení získaných řetězců textu jsem vytvořil dvě pole. Do prvního pole budu ukládat nalezené řetězce textu a do druhého pak pozice nalezeného textu.

Parsování pak vypadá takto:

```
for (Element ocrxWord : doc.select(".ocrx_word")) {

    jmena[i] = ocrxWord.text();  //JMENO, PRIJMENI,
                                 CISLA

    pozice[i] = ocrxWord.attr("title");  //bbox 250 192
                1606 375; x_wconf 70
```

Cyklus hledá pouze text v souboru hOCR, tedy hodnoty „**.ocrx_word**“.

Do pole **jmena[]** se pomocí příkazu **ocrxWord.text();** uloží všechny stringy z dokumentu, které byly naskenovány pomocí Tesseractu.

ocrxWord.attr(“title“); uloží do pole **pozice[]** všechny hodnoty atributu **title**, mezi nimiž jsou i námi vyžadované pozice načtených stringů.

Každý string je po uložení kontrolován, zdali splňuje předem daná pravidla pro anonymizaci dat. Pokud splňuje kontrolovaný textový řetězec pravidla, lze tvrdit,

že řetězec obsahuje citlivá data, jako je například jméno pacienta, rodné číslo nebo datum narození.

Pro zjišťování citlivých dat jsem vytvořil tři metody:

- **zjistiJmeno();**
- **zjistiDatum();**
- **zjistiCislo()**;

Index **i** odkazuje na hodnotu momentálně testovaného řetězce. Všechny metody jsou **void**, tudíž nemají žádnou návratovou hodnotu.

Metoda **zjistiJmeno();**

Tato metoda, jak již název napovídá, je určena pro nalezení jména nebo příjmení v získaném textovém řetězci.

Při vývoji programu jsem při prohlížení obrazových souborů zjistil, že téměř vždy je jméno nebo příjmení napsáno velkými písmeny. Proto jsem se rozhodl kontrolovat řetězce, ve kterých jsou pouze velká písmena.

Pro procházení řetězce jsem použil cyklus **for**, který kontroluje každý char v řetězci funkcí:

Character.isUpperCase(jmena[i].charAt(k));

Hodnota **k** udává pozici znaku v testovaném řetězci.

Návratová hodnota funkce je typu boolean, tedy true nebo false. Pokud je nalezeno číslo nebo jiný znak, cyklus **for** se přeruší a přechází se na další nalezený řetězec.

V této metodě jsem musel ošetřit případy, kdy se z Tesseractu načetl řetězec, který obsahoval na konci jména znak tečky nebo čárky viz tabulka č. 3.

Tabulka 3: Příklad ukládání řetězců do pole stringů

Původní text	**První řetězec**	**Druhý řetězec**
PŘIJMENÍ, JMÉNO	PŘIJMENÍ,	JMÉNO

Metoda dokáže akceptovat tečku nebo čárku, vyskytují-li se na konci řetězce a zároveň, pokud je řetězec delší než tři znaky. Je tomu tak z důvodu, aby nebyl akceptován řetězec složený ze dvou čárek jako jméno.

Pokud splňuje nalezený řetězec všechna pravidla, program rozhodne, že je nutné tento řetězec anonymizovat. Je zavolána příslušná metoda **anonymizujJmeno()**, která anonymizaci provede. Jak tato metoda funguje a jak probíhá anonymizace, vysvětlím v další části.

Metoda **zjistiDatum();**

Při prohlížení dodaných materiálů jsem objevil další pravidlo a to, že datum narození vždy začíná znakem hvězdičky „*****“. Metoda proto při nalezení znaku hvězdičky předpokládá, že dále bude následovat datum narození. Opět platí pravidlo, že řetězec musí být delší než tři znaky.

Po splnění kritérií pro anonymizaci se zavolá metoda **anonymizujDatum()**.

Metoda **zjistiCislo();**

Tato metoda je určená pro zjištění rodného čísla. Funguje na principu metody **zjistiJmeno()** s tím rozdílem, že místo kontroly, zda je znak velké písmeno (UpperCase), kontroluje shodu znaku s číslem, viz funkce:

Character.isDigit(jmena[i].charAt(k))

Hodnota **k** udává pozici charu v testovaném řetězci.

V metodě jsem ošetřil případy, kdy načtená hodnota z Tesseractu nemusela odpovídat originální hodnotě, viz tabulka č. 4.

Tabulka 4: Příklad nepřesně načteného řetězce z Tesseractu

Původní text	**Načtený text**
12/23/4567	1272374567
123456789	123456?89

Metoda dokáže akceptovat i takto nepovedeně načtené řetězce. Dokáže stejně jako metoda **zjistiJmeno()** ošetřit tečku nebo čárku na konci řetězce.

Zavedl jsem pravidlo, že řetězec musí být delší než 7 znaků. Zmíněné pravidlo je zapotřebí, nastane-li podobný případ jako ve výše uvedené tabulce, kdy Tesseract zamění při načítání znak lomítka „/“ za číslici **7**. Z řetězce o 8 číslicích se pak rázem stane řetězec obsahující 10 číslic.

Po nalezení citlivých dat se zavolá funkce **anonymizujCislo()**. Zavolanou funkci vysvětlím v další části.

4.3.5. Anonymizace nalezených citlivých dat

Pro anonymizaci nalezených citlivých dat jsem vytvořil tři metody:

- **anonymizujJmeno();**
- **anonymizujDatum();**
- **anonymizujCislo();**

Metody fungují na podobném principu. Jedná se o metody typu void, takže nevrací žádnou návratovou hodnotu.

V každé metodě se zavolá vnořená metoda **Zapis();** a vypočte se šířka (width) a výška (height) obdélníku pro vykreslení.

Následně se zavolá metoda **Vykresli();**, která použije vypočtenou šířku a výšku a předá jí společně s přepočtenými souřadnicemi anonymizovaného textu.

Metoda Zapis();

Abych v programu zamezil zbytečnému opakování stejných příkazů, vytvořil jsem metodu **Zapis();** Metoda obsahuje pole **rozlozeniPozic[]**, do kterého se pomocí funkce split rozdělí řetězec na podřetězce.

Ukázka kódu:

```
rozlozeniPozic = pozice[i].split(" |\\;|\\,|\\^");
```

Tento proces je nezbytný k získání pozic řetězce, neboť původní text obsažený v poli **pozice[]** nemá správnou formu (viz část Nalezení citlivých dat – parsování).

Příklad řetězce pole **pozice[i]**:

```
bbox 250 192 1606 375; x_wconf 70
```

Pomocí funkce split tedy získáme hodnoty, viz tabulka č. 5.

Tabulka 5: Rozdělení hodnot pomocí funkce split

k = 0	**k = 1**	**k = 2**	**k = 3**	**k = 4**	**k = 5**	**k = 6**
bbox	250	192	1606	375	x_wconf	70

Kde hodnota **k** značí index pole **rozlozeniPozic[]**.

Pro anonymizaci jsou důležité pouze souřadnice nalezeného textu, tedy **X1**, **Y1**, **X2** a **Y2**. Získaná hodnota z funkce split je textový řetězec, proto je nutné pomocí parseru hodnoty převést na číselné.

Po převedení dosadím hodnoty k=1 až k=4 do proměnných **X1**, **Y1**, **X2** a **Y2**.

Získání souřadnic ovšem není poslední krok, protože tyto souřadnice jsou určené pro upravený obrazový soubor, tedy ten s rozlišením 5000x5000. Dalším krokem je převedení souřadnic pro anonymizaci původního obrazového souboru.

Na to jsem vytvořil metodu **Prepocti();**.

Po získání potřebných souřadnic pak zavolám metodu **Vykresli();**

Vykresli(int X1, int Y1, int X2, int Y2);

kde **X1**, **Y1**, **X2**, **Y2** jsou pozice nalezeného textu.

V průběhu chodu program vypisuje na obrazovku veškeré úkony, které provádí a zároveň zapisuje do logovacího souboru důležité informace o anonymizaci.

Na začátku logovacího souboru je zapsán obrázek a informace, zda obsahuje citlivá data či nikoliv (Private data/Clean).

Přiklad zápisu logovacího souboru:

obrazek.png Private data

obrazek.png 25 19 160 37 JMENO

Jednotlivé údaje jsou odděleny tabulátorem.

Metoda Prepocti();

Metoda funguje na jednoduchém principu trojčlenky, kdy vím, že upravený soubor bude vždy v rozlišení 5000x5000, takže stačí jen procentuálně přepočítat souřadnice pro původní obrazový soubor.

5. Zhodnocení výsledků

V této poslední kapitole bych rád zhodnotil dosažené výsledky. Cílem bakalářské práce bylo vytvořit algoritmus zlepšující anonymizaci dat. Vytvořené metody splňují požadavek zvýšené bezpečnosti citlivých údajů při práci s obrazovou částí. Program dokáže identifikovat a odstranit citlivé osobní údaje z obrazových souborů.

Z původních dodaných souborů bylo pro program použitelných pouze 41 z 86, protože zbývajících 45 souborů neobsahovalo žádná data či žádný text. Při zpracování programu jsem použil několik dostupných aplikací, které jsem uváděl v praktické části a které byly nezbytné pro dosažení optimálních výsledků. Dále byla vytvořena databáze tzv. bílých slov, což jsou slova, která při shodě nalezení v textu se nepovažují za citlivá data. Omezujícím prvkem při tvorbě programu byl OCR program Tesseract verze 3.03, který přes vysokou úspěšnost rozeznávání znaků (přibližně okolo 80%), nebyl perfektní. Tuto vadu jsem se snažil do jisté míry v programu opravit zvětšením obrazového souboru a také pomocí trénování programu Tesseract. Také Tesseract rozděloval textové řetězce, kde si myslel, že je mezera a začátek nového řetězce. Program byl vyvíjen a testován na platformě Linux Ubuntu 14.04LTS.

Výsledkem mé práce je program, který anonymizuje medicínská data, aniž by se musela nahrávat na neznámé webové stránky, které by mohly ohrozit soukromí dat a tím i následné soukromí pacientů či případných dalších zájmových skupin.

Na závěr bych chtěl podotknout, že výsledky programu, který úspěšně anonymizuje citlivá medicínská data nalezená v dodaných obrazových souborech části DICOM, mě samotného překvapily.

V dalších verzích programu by bylo možné rozšířit funkčnost programu, např. vytvořením učenlivé databáze pro jména a příjmení, která by usnadnila anonymizaci dat, např. nebude-li jméno nebo příjmení velkými písmeny. Dále by bylo možné vytvořit grafického rozhraní s možností prohlížení upravených obrazových souborů. Také by bylo možné lépe trénovat OCR engine Tesseract, který by pak dával lepší výsledky.

Závěr

Tato bakalářská práce byla zaměřena na analýzu legislativních bezpečnostních požadavků pro zpracování medicínských dat a vytvoření anonymizačního programu pro zpracování citlivých dat obsažených v obrazové části DICOM souborů.

V této práci jsem nejprve osvětlil úvod do bezpečnostní problematiky a poté jsem rozvedl právní předpisy a zákony, které s touto problematikou přímo či okrajově souvisejí.

V dalším bodě teoretické části jsem představil nejpoužívanější formát pro medicínská data, jímž je formát DICOM. Zde jsem uvedl i něco z historie, nejenom v České republice, a popsal jeho základní části. Mimo jiné jsem zhodnotil dostupné metody anonymizace. Ač byly browserové aplikace užitečné, vyšlo mi z této analýzy najevo, že nejlepším řešením zůstává zpracování vlastního programu.

Tomu je de facto věnována celá praktická část, v níž jsem se věnoval vytvoření souborů hOCR pomocí programu Tesseract-Ocr, ze kterých jsem poté mohl citlivá data získávat. Za pomoci knihoven JSoup jsem z těchto souborů mohl postupně parsovat citlivá data. Po vyparsování zjistí mnou vytvořené metody, zda se jedná o data citlivá. Pokud ano, jsou nalezená data okamžitě anonymizována. V opačném případě program tato data ignoruje. Nalezená data jsou automaticky zapisována do logovacího souboru a průběžný stav programu je vypisován na obrazovce počítače. Po kontrole všech získaných řetězců ze souboru hOCR se všechny pomocné soubory, vytvořené pro anonymizaci, smažou. Mezi tyto pomocné soubory patří hOCR a pomocný zvětšený obrazový soubor.

Jsem vděčný za možnost zpracování tohoto tématu a to z důvodu, že mně bylo umožněno nahlédnout i do fungování zdravotnictví. Jelikož velká část materiálů, které jsem obdržel či si vyhledával, byly v anglickém jazyce, měl jsem také možnost si prověřit své jazykové znalosti a mohl jsem aplikovat dosažené teoretické znalosti v oboru Informační systémy.

Literatura a prameny

1. **Konference ICT ve zdravotnictví | Inflow.** ***Inflow | magazín nejen pro knihovníky.*** **[Online] [Citace: 24. duben 2016.] http://www.inflow.cz/konference-ict-ve-zdravotnictvi.**

2. ***Zákony pro lidi - Sbírka zákonů ČR v aktuálním konsolidovaném znění.*** **[Online] [Citace: 24. duben 2016.] http://www.zakonyprolidi.cz/.**

3. Kybernetický zákon. ***Kybernetický zákon.*** **[Online] [Citace: 24. duben 2016.] http://kybernetickyzakon.cz/.**

4. WWW.CLK.CZ: Úmluva o lidských právech a biomedicíně:. ***WWW.CLK.CZ: Česká lékařská komora - OS Děčín (Index):.*** **[Online] [Citace: 24. duben 2016.] http://www.clk.cz/oldweb/zakpred/Uml096-2001_EtikaBiomed.html.**

5. Zákon o zdravotních službách - č. 372/2011 Sb. - Aktuální znění:. ***Zákony pro lidi - Sbírka zákonů ČR v aktuálním konsolidovaném znění:.*** **[Online] [Citace: 24. duben 2016.] https://www.zakonyprolidi.cz/cs/2011-372.**

6. Vyhláška o zdravotnické dokumentaci - č. 98/2012 Sb. - Aktuální znění:. ***Zákony pro lidi - Sbírka zákonů ČR v aktuálním konsolidovaném znění:.*** **[Online] [Citace: 24. duben 2016.] https://www.zakonyprolidi.cz/cs/2012-98.**

7. Zákon o ochraně osobních údajů - č. 101/2000 Sb. - Aktuální znění. ***Zákony pro lidi - Sbírka zákonů ČR v aktuálním konsolidovaném znění.*** **[Online] [Citace: 24. duben 2016.] https://www.zakonyprolidi.cz/cs/2000-101.**

8. Zákon o kybernetické bezpečnosti a o změně souvisejících zákonů (zákon o kybernetické bezpečnosti) - č. 181/2014 Sb. - Aktuální znění:. ***Zákony pro lidi - Sbírka zákonů ČR v aktuálním konsolidovaném znění:.*** **[Online] [Citace: 24. duben 2016.] https://www.zakonyprolidi.cz/cs/2014-181.**

9. Předpis č. 316/2014 Sb. ***https://www.zakonyprolidi.cz.*** **[Online] [Citace: 24. duben 2016.] https://www.zakonyprolidi.cz/cs/2014-316.**

10. Sb., Předpis č. 317/2014. Vyhláška o významných informačních systémech a jejich určujících kritériích - č. 317/2014 Sb. - Aktuální znění:. ***Zákony pro lidi***

- *Sbírka zákonů ČR v aktuálním konsolidovaném znění.* [Online] [Citace: 24. duben 2016.] https://www.zakonyprolidi.cz/cs/2014-316.

11. DICOM: About DICOM. *DICOM Homepage.* [Online] [Citace: 24. duben 2016.] http://medical.nema.org/Dicom/about-DICOM.html.

12. DICOM Homepage. *DICOM Homepage.* [Online] [Citace: 24. duben 2016.] http://medical.nema.org/standard.html.

13. Systémy PACS z hlediska databázových informačních systémů:. *SystemOnLine.cz - ekonomické a informační systémy v praxi:.* [Online] [Citace: 24. duben 2016.] https://www.systemonline.cz/clanky/systemy-pacs-z-hlediska-databazovych-systemu.htm.

14. Facepixelizer | Pixelate - Blur - Anonymize | Free Online Image Editor:. [Online] [Citace: 24. duben 2016.] http://www.facepixelizer.com/.

15. ATBON. Anonymizace a úprava dokumentů. [Online] Atbon, a.s. [Citace: 24. duben 2016.] http://www.atbon.cz/redacting.ph.

16. PhotoHide. PhotoHide.com - hide the face on your personal photos to ensure your privacy:. *PhotoHide.com.* [Online] [Citace: 24. duben 2016.] http://www.photohide.com/.

17. Lionytics™. Lionytics Image Anonymizer. *Lionytics Image Anonymizer.* [Online] [Citace: 24. duben 2016.] http://www.lionytics.com/lionytics/image-anonymizer/.

18. *tesseract-ocr An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google. - Google Project Hosting.* [Online] [Citace: 24. duben 2016.] https://code.google.com/p/tesseract-ocr/.

19. ImageMagick: Convert, Edit, Or Compose Bitmap Images. *ImageMagick: Convert, Edit, Or Compose Bitmap Images.* [Online] [Citace: 24. duben 2016.] http://www.imagemagick.org/script/index.php.

20. jsoup Java HTML Parser, with best of DOM, CSS, and jquery. *jsoup Java HTML Parser, with best of DOM, CSS, and jquery.* [Online] [Citace: 24. duben 2016.] http://jsoup.org/.

21. VietOCR. Tesseract box editor & trainer. *VietOCR.* [Online] [Citace: 24. duben 2016.] http://vietocr.sourceforge.net/training.html.

22. O projektu. *ePACS - DICOM komunikace mezi zdravotnickými zařízeními.* [Online] [Citace: 24. duben 2016.] http://www.epacs.cz/faces/pages/o-projektu.xhtml.

Seznam zkratek

DICOM ..... Digital Imaging and Communications in Medicine

OCR ..... Optical Character Recognition

CT ..... Computed Tomography

ACR ..... American College of Radiolog

NEMA..... National Electrical Manufacturers Association

PACS ..... Picture Archiving and Communication System

HTML ..... HyperText Markup Language

XHTML ..... eXtensible HyperText Markup Language

HOCR ..... Optical Character Recognition

CERT ..... Computer Emergency Response Team

UNLV ..... University of Nevada-Las Vegas

MRI..... Magnetic Resonance Imaging

ÚOOÚ..... Úřad pro ochranu osobních údajů

Seznam tabulek

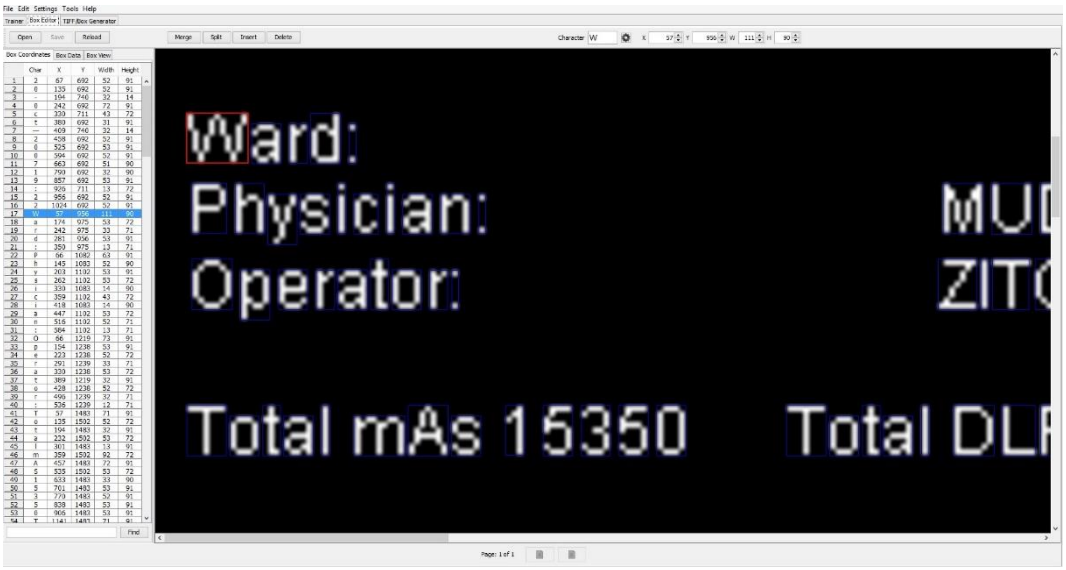
Seznam obrázků

Seznam příloh

- CD obsahující tuto dokumentaci ve formátu PDF a XDOC, zdrojové soubory a podpůrné knihovny.
- Příloha A
- Příloha B

Příloha A – Upravování boxů v jTessBoxEditor

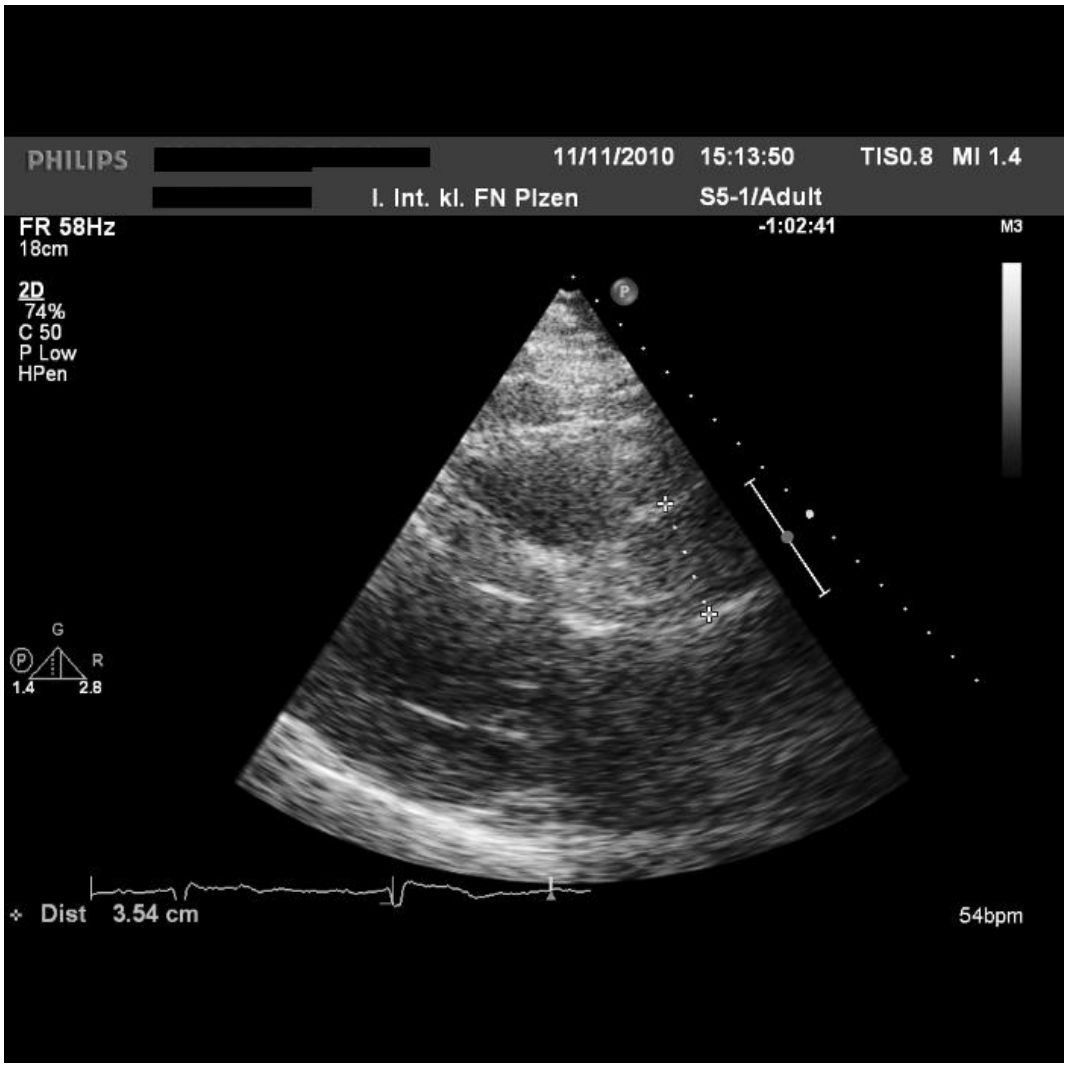
Obrázek 2: Upravování boxů v jTessBoxEditor



Zdroj vlastní.

Příloha B – Anonymizovaná medicínská data

Obrázek 3: Příklad anonymizace obrazového souboru



Zdroj vlastní.