Posudek oponenta bakalářské práce

Autor/autorka práce: **Jaroslav Malát**

Název práce: **Zabezpečené zpracování medicínských obrazových dat**

Předložená bakalářská práce se snaží řešit anonymizaci medicínských obrazových dat, přesněji řečeno se snaží v obrázku najít „citlivé" texty zakrýt je.

Struktura textu bakalářské práce víceméně kopíruje zadání práce.

První část se věnuje právním aspektům problematiky; činí tak buď doslovnými citacemi právních norem, nebo jejich volným převyprávěním. Jako laik nedokážu posoudit, zda je přehled dostatečný. Z laického pohledu mi ale text připadal poněkud zbytečný. Text je plný jednak banálních sdělení (např. „měli bychom znát potenciální slabá místa", viz strana 4), jednak nesmyslně podrobných výčtů (např. odstavec 4 kapitoly 1 nebo popis HIPAA v kapitole 1.2.1). Dále se text snaží věnovat i mnoha tématům mimo hlavní zaměření práce. Tím se – vzhledem k rozsahu práce – vlastně čtenář dozví jen jakési útržkovité informace bez patřičného kontextu. V této kapitole bych spíš očekával jasné vymezení pojmu „citlivé údaje" a jaké hrozí důsledky při jejich úniku.

Další část se věnuje popisu standardu DICOM. Popis mi opět připadá zbytečně rozsáhlý, zejména proto, že se s ním dále nijak nepracuje. Pro potřeby bakalářské práce by stačilo říct, že DICOM obsahuje obrazová data a jejich popis (tagy). Text je navíc prakticky doslovně převzatý z přednášky doc. Münze o standardu DICOM v předmětu „Informační systémy ve zdravotnictví" na ČVUT (viz http://webzam.fbmi.cvut.cz/szabozol/ISZ/), viz např. popis části PS 3.3 v příloze tohoto posudku.

Část věnující se PACS je prakticky doslovným překladem zdroje [16], viz příloha tohoto posudku. Podobně jako nepříliš kvalitní anglický zdroj, i studentův překlad je jakýmsi roztříštěných souhrnem frází, text nemá hlavu ani patu. Navíc není jasné, proč student neuvádí do souvislosti PACS a anonymizaci dat.

V kapitole 2 se práce zabývá existujícími nástroji pro anonymizaci medicínských dat. Popis je značně neuspokojivý. Text se věnuje především podružným záležitostem (podporované operační systémy, použitý programovací jazyk apod.), ale vynechává podstatné vlastnosti – zda anonymizace probíhá jen na úrovni tagů či rovněž v obrazových datech; jak se ověřuje, že se na odstranění nějakých citlivých údajů nezapomnělo; jak programy spolupracují s hlavním informačním systémem nemocnice; jak se řeší nasazení (deployment) atd. Poněkud zvláštní je studentův názor, že browserové anonymizátory nejsou vhodné kvůli „odesílání dat na cizí server" – za prvé se zdá, že vývoj spíš směřuje k browserovým aplikacím, za druhé bývá server buď spravovaný samotnou nemocnicí, nebo důvěryhodnou organizací. Za úsměvný považuji fakt, že student klade velký důraz na cenu anonymizačního softwaru – pochybuji, že malá nemocnice s ročním obratem cca 500 mil. Kč bude zaskočena cenou licence 1000 USD.

Závěr kapitoly 2 se věnuje požadavkům na anonymizační program, který student dále implementuje. Tyto požadavky jsou většinou obecnými floskulemi – po každém programu bychom chtěli, aby byl přesný, rychlý, konfigurovatelný apod. Požadavek na lokální zpracování je diskutabilní (data stejně bývají uložena centrálně). Naopak zcela chybí diskuse, jak se bude ověřovat úplnost anonymizace, jak bude program napojen na informační systém apod. Student z neznámého důvodu nezdůrazňuje, že jeho cílem je anonymizovat obrazová data (tj. nikoliv jen tagy), a to ve velkém množství obecných obrazových souborů, kde mohou být jakékoliv texty. Zásadně zde chybí diskuse, nakolik je tato úloha realistická – pokud je třeba anonymizovat tisíce snímků z jednoho přístroje (např. sonografu), bude

typicky struktura obrazových dat pevně daná; a pokud je nutné anonymizovat několik obecných snímků, je ruční práce asi nejrychlejší a nejspolehlivější.

Kapitola 3 popisuje klasifikaci textů nalezených v obrazových datech, tj. rozhodovací proces, zda text anonymizovat. Za prvé mi připadá velmi zvláštní, že student nepoužívá regulární výrazy nebo gramatiky. Za druhé jsou některá pravidla značně nedotažená, např. mi není jasný požadavek na délku čísla 7–14 znaků, proč nemůže jméno obsahovat např. spojovník atd. Za třetí, a to je nejdůležitější – vůbec se neřeší, jaké texty je nutné v obrazovém souboru zachovat, a jaké je nutné smazat. Pokud se mohou smazat všechny texty, není nutné je klasifikovat; a pokud je nutné některé zachovat, není jasné, jak je poznat.

Kapitola 4 se zabývá samotnou implementací. Většina kapitoly je věnována popisu triviálních metod; netuším, proč jej student nezahrnul jen do Javadoc dokumentace. Mezi ním se tu a tam objeví skutečný popis logiky běhu aplikace, zdůvodnění výběru použitých nástrojů, detaily spouštění pomocných nástrojů apod. Celkově je text značně nepřehledný.

Úvahy stojící za jednotlivými algoritmy i jejich implementace jsou velmi slabé. Například převzorkování obrazu na 5000 pixelů na šířku před procesem OCR je zjevně nesmyslné, přeci musí záležet na velikosti vstupu. Číslo 5000 se sice zadává v konfiguračním souboru properties.properties jako parametr konverze, ale metoda Anonymizer.prepocti() ho zřejmě předpokládá. Implementace metody Anonymizer.zkontrolujCislo() je něco, co by absolvent jakéhokoliv informatického oboru vůbec neměl napsat. O „upravitelnosti“ procesu anonymizace (viz požadavky na program, viz závěr kapitoly 2) vůbec nemůže být řeč, naprostá většina funkčnosti je natvrdo zapsána v kódu; a tak dále. Vedle toho působí jen jako podružné detaily například zbytečná závislost na ImageMagick (převzorkování lze snadno řešit pomocí AWT Graphics), argumenty zdůvodňující výběr OCR Tesseract („jeden ze tří nejlepších engine v UNLV testu přesnosti v roce 1995“, viz str. 35) nebo nejasný účel souboru cetnost.json.

Kapitola 5 se snaží zhodnotit výsledky, ale opět je znát studentova bezradnost. Místo kritického zhodnocení „kolik citlivých údajů se nepodařilo odstranit“, „kolik nerelevantních údajů bylo omylem odstraněno“ a „kolik času zabere uživateli oprava špatné anonymizace“ se dočteme jen nicneříkající údaj „kolik řetězců vyhovujících mým pravidlům bylo odstraněno“. Rovněž chybí názor odborníka, jak je daná implementace anonymizace v praxi použitelná.

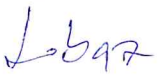
Součástí práce je CD s implementovaným programem a ukázkovými výstupy. Za neomluvitelné považuji, že práce neobsahuje návod na instalaci a spuštění ani ukázková vstupní data. Ačkoliv autor uvádí, že program vyvíjel na platformě Linux, nikde nepíše, že program je na Linux striktně vázaný (Java program volá konkrétní příkazy shellu bez možnosti rekonfigurace), ani neuvádí důvody pro výběr Linuxu. U ukázkových výstupních dat je velmi nešťastné, že proces anonymizace není často patrný (černá přelepka na původně černém pozadí není vidět) a že opět není jasné, kolik řetězců bylo odstraněno omylem.

Závěr: teoretická část textu je velmi slabá, stejně jako implementované řešení. Tím pádem není ani v praktické části textu příliš o čem psát. Ačkoliv má text nadprůměrných 58 stran, jde rozhodně o podprůměrnou práci. Bohužel, kvalita implementovaného řešení neumožňuje nad textem přimhouřit oko.

Navrhuji hodnocení známkou **nevyhověl / nevyhověla** a práci nedoporučuji k obhajobě.

V Plzni 22.8.2017

Ing. Petr Lobaz



Příloha

PS 3.3 Definice informačních objektů

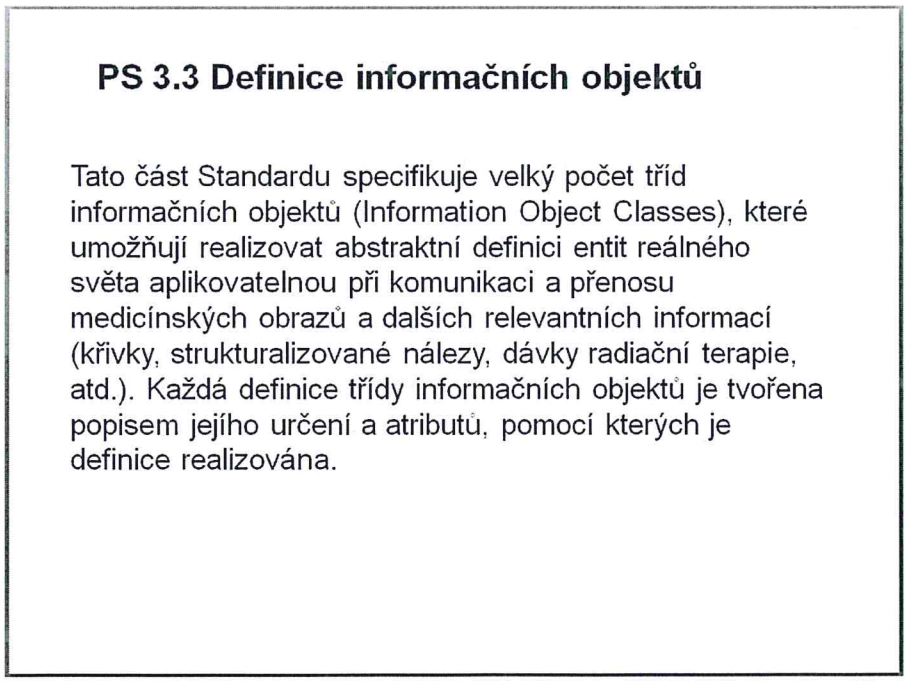
V této části standardu jsou specifikovány třídy informačních objektů (Information Object Classes), které umožňují realizovat abstraktní definici entit reálného světa aplikovatelnou při komunikaci a přenosu medicínských obrazů a informace s nimi spojené (křivky, strukturalizované nálezy, dávky radiační 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.

Výňatek z bakalářské práce, strana 15



Snímek 6 z přednášky DICOM (Digital Image and Communications In Medicine protocol suite), doc. Ing. Jan Münz, CSc. (http://webzam.fbmi.cvut.cz/szabozol/ISZ/Seidl_Munz/DICOM.ppt)

1.4. PACS

PACS (picture archiving and communication system) je technologie ve zdravotnictví pro krátkodobou a dlouhodobou archivaci, vyhledávaní, správu, distribuci a prezentaci medicínské obrazové dokumentace. Mezi obrazovou dokumentaci řadíme snímky z rentgenu, centrálního tomografu, magnetické rezonance apod.

PACS umožňuje zdravotnické organizaci (jako je například nemocnice) zachytit, ukládat, prohlížet a sdílet všechny typy medicínské obrazové dokumentace jak interně, tak i externě. Při nasazování PACS musí zdravotnická organizace zvážit prostředí, v němž bude použito (hospitalizační, ambulantní, nouzové, specializované) a další elektronické systémy, se kterými se bude integrovat.

Interoperabilita (schopnost různých systémů vzájemně spolupracovat, poskytovat si služby, dosáhnout vzájemné součinnosti) snímků v samostatných systémech PACS je obavou pro poskytovatele zdravotní péče, a to i mezi různými poskytovateli v rámci stejného systému zdravotní péče. Přenos lékařských snímků je technologicky možný, pokud nejsou komplikované konkurenčními systémy, které nejsou interoperabilní.

Výňatek z bakalářské práce, strana 19

PACS (picture archiving and communication system) is a healthcare technology for the short- and long-term storage, retrieval, management, distribution and presentation of medical images.

A PACS allows a healthcare organization (such as a hospital) to capture, store, view and share all types of images internally and externally. When deploying a PACS, the organization needs to consider the environment in which it will be used (inpatient, ambulatory, emergency, specialties) and the other electronic systems with which it will integrate.

The interoperability of images in separate PACS is a concern for healthcare providers, even among different providers within the same healthcare system. The transmission of medical images is technologically possible when not complicated by competing, noninteroperable systems.

Výňatek ze zdroje [16],
http://searchhealthit.techtarget.com/definition/picture-archiving-and-communication-system-PACS