VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

BIG DATA ANALÝZY A STATISTICKÉ ZPRACOVÁNÍ METADAT V ARCHIVU OBRAZOVÉ ZDRAVOTNICKÉ DOKUMENTACE

BIG DATA ANALYSIS AND METADATA STATISTICS IN MEDICAL IMAGES ARCHIVES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. Michal Pšurný

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. Vratislav Harabiš, Ph.D.

BRNO 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Biomedicínské inženýrství a bioinformatika**
Ústav biomedicínského inženýrství

Student: Bc. Michal Pšurný ***ID:*** 147474
Ročník: 2 ***Akademický rok:*** 2016/17

NÁZEV TÉMATU:

Big data analýzy a statistické zpracování metadat v archivu obrazové zdravotnické dokumentace

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši metod big data analýz a statistických metod používaných pro zpracování zdravotnických dat. 2) Navrhněte vhodné využití těchto metod k vytěžení informací z archivu DICOM. 3) Navrhněte, které informace z DICOM archívu jsou vhodné pro big data analýzy. 4) Navrhněte statistické soubory dat, které jsou získatelné z DICOM archivu a mohou mít podstatný význam pro procesy obrazové diagnostiky ve zdravotnickém zařízení. 5) Navrhněte blokové schéma a algoritmy pro zpracování dat identifikovaných v předchozích krocích. 6) Navrhněte a implementujte software (MATLAB nebo Java - implementace přímo do DICOM serveru MARIE SERVER Express) podle navrženého algoritmu. 7) Statisticky vyhodnoťte úspěšnost navrženého řešení a proveďte diskuzi dosažených výsledků. Zadání práce je vytvořeno ve spolupráci s firmou Medical Solutions OR-CZ spol. s r.o. a v rámci práce je možné uznání povinné odborné praxe.

DOPORUČENÁ LITERATURA:

[1] WANG, Baoying, Ruowang LI a W. PERRIZO. Big data analytics in bioinformatics and healthcare. ISBN 9781466666146.

[2] PIANYKH, Oleg S. Digital imaging and communications in medicine (DICOM): a practical introduction and survival guide. Berlin: Springer, c2008. ISBN 9783540745709.

Termín zadání: 6.2.2017 ***Termín odevzdání:*** 19.5.2017

Vedoucí práce: Ing. Vratislav Harabiš, Ph.D.
Konzultant: Ing. Svatopluk Beneš

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

Abstrakt

Diplomová práce se zabývá problematikou big dat ve zdravotnictví. Zaměřuje se data z archivu obrazové zdravotnické dokumentace, konkrétně na hlavičky DICOM souborů. Spolu s obrazovou informací, je do DICOM formátu ukládáno velké množství dalších dat související s pořízením obrazu. Práce mapuje tyto data na 1215 studiích.

Klíčová slova

dicom, big data, big data ve zdravotnictví, metadata, dicom tag, informace z dicom, big data analýza

Abstract

This Diploma thesis describes issues of big data in healthcare focus on picture archiving and communication system. DICOM format are store images with header where it could be other valuable information. This thesis mapping data from 1215 studies.

Keywords

dicom, big data, big data in healthcare, big healthcare data, big data analysis, big data analysis in healthcare, metadata, dicom tag, dicom metadata, information from dicom

Bibliografická citace

PŠURNÝ, M. *Big data analýzy a statistické zpracování metadat v archivu obrazové zdravotnické dokumentace* . Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2017. 56 s. Vedoucí semestrální práce Ing. Vratislav Harabiš, Ph.D..

Prohlášení

Prohlašuji, že jsem svou diplomovou práci, na téma *Big data analýzy a statistické zpracování metadat v archivu obrazové zdravotnické dokumentace*, vypracoval samostatně pod vedením vedoucího a konzultanta semestrálního projektu a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s vytvořením této práce neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 19. května 2017 **……**

podpis autora

Poděkování

Mnohokrát děkuji mému vedoucímu diplomové práce panu Ing. Vratislavu Harabišovi, PhD, taktéž svému konzultantovi Ing. Svatopluku Benešovi a firmě OR-CZ spol. s.r.o. za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce. Dále bych chtěl poděkovat panu Ing. Petru Čačíkovi za poskytnutí dat z archivu FN Brno.

V Brně dne 19. května 2017 **…….**

podpis autora

Obsah

1 Úvod

Problém vzniku a ukládání velkého množství dat nastává i ve zdravotnictví. V souvislosti s takto velkým množstvím dat vzniká otázka, jestli můžeme z ukládaných dat zjistit další informace nebo souvislosti, které nám běžnou analýzou a zpracováním mohou unikat.

Diplomová práce mapuje data, která ukládáme spolu se vznikem obrazových informací ve zdravotnictví. Standardizovaný DICOM formát, který je využíván celosvětově, definuje principy ukládání a přeposílání těchto dat. Spolu s každým jednotlivým snímkem, který ve zdravotnictví vznikne, je ukládáno velké množství dalších, doplňujících dat. Jedná se o informace související s pacientem, modalitou, technickými parametry nebo informacemi o průběhu vyšetření. Zatím neexistují studie, které by ve větším měřítku mapovaly, jaké informace a v jaké kvalitě se do DICOM souborů ukládají.

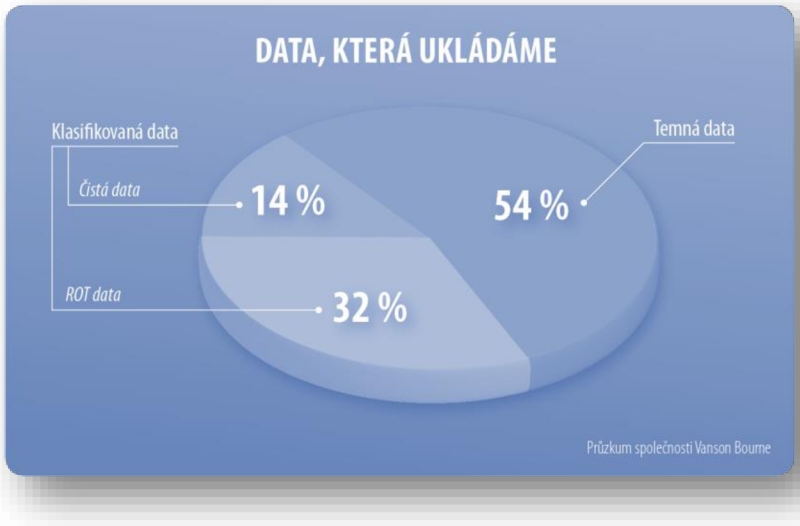
DICOM standard definuje jednoznačné číselné identifikátory, tzv. „tagy“, pomocí kterých se dá ke každé konkrétní informaci přistoupit. Centrum informatiky Fakultní nemocnice Brno poskytlo 1215 studií, na kterých mohla být provedena analýza, která data a v jaké kvalitě se do DICOM souborů spolu s obrazovou informací ukládají.

Diplomová práce vznikla za spolupráce firmy OR-CZ spol. s.r.o., která je předním dodavatelem PACS systémů v České republice.

Diplomová práce v teoretické části vysvětluje pojem big data, zabývá se vznikem a analýzou big dat ve zdravotnictví, popisuje DICOM formát a vytváří tak ucelený přehled k praktické části. V praktické části jsou pak popsány data, která byla k dispozici, separaci dat, mapuje informace obsažené v datech, a nakonec se zabývá jejich analýzou.

2 Big data

S příchodem digitálního světa si postupně začínáme uvědomovat, jak velké objemy dat ukládáme a máme k dispozici. Dle průzkumu provedeného společnosti *Vanson Bourne,* až 54 % uložených dat jsou data, o kterých nic nevíme („*temná data*“). Ze zbylých 46 % tvoří jenom 14 % data, která jsou užitečná a mají informační hodnotu. 32 % pak tvoří tzv. „*ROT data*“, která nemá smysl dále ukládat. Tento poměr je zobrazen na obrázku 1. Průzkum byl proveden na 1475 respondentech v nejrůznějších odvětvích, ve 14 zemích Evropy, středního východu a Afriky [9].



Obrázek 1. Procentuální zastoupení vznikajících dat [9].

Produkovaná data není jenom obtížné technicky spravovat, často bývají zašuměná, nemají strukturu a v této surové podobě ani žádnou informační hodnotu. Uvádí se, že okolo 80 % všech dat se ukládá nestrukturovaně a nejvíce se jich nachází v textových souborech. Pokud se tato data nespravují, rostou každoročně náklady na ukládání těchto dat. Taktéž může být pro organizace velký problém dohledávat informace v případě auditu. Další problém je, že produkce dat exponenciálně roste. Uvádí se, že 90 % všech existujících dat bylo vyprodukováno v posledních dvou letech. Problém se dá řešit tak, že čistá data řádně uložíme a zpřístupníme, ROT data smažeme a na „temná“ data si „posvítíme“. Taktéž vzniká otázka, jestli se z těchto dat dají vyčíst nové informace nebo souvislosti. Problematika big dat se začala nejdříve řešit hlavně ve firmách v souvislosti s náklady na ukládání dat a hledáním nových obchodních a marketingových modelů. Postupně se ale začíná řešit ve všech odvětvích. Zdravotnictví, které je v této problematice technologicky opožděné, se postupně aklimatizuje na dnešní digitální věk [3] [5] [9] [11] [16].

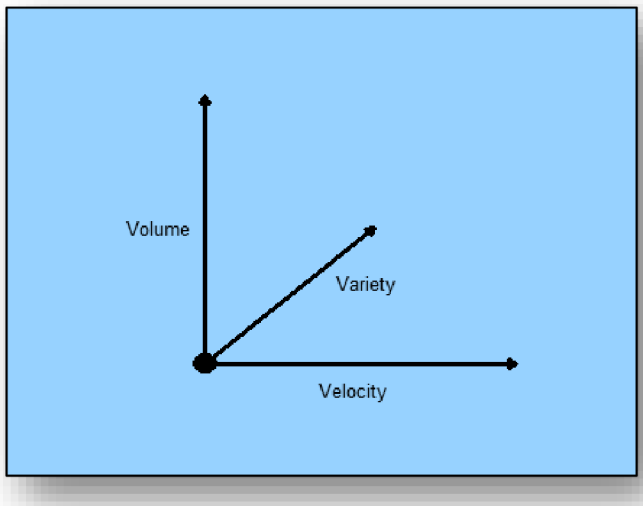
2.1 Definice pojmu big data

Big data není nový pojem, význam se však v čase konstantně mění. Vžil se zejména v souvislosti se vznikem obrovského množství dat v digitálním světě. Charakterizuje se jako soubor datových prvků, které svou velikostí, rychlostí vzniku, heterogenitou a složitostí vyžadují nové hardwarové a softwarové přístupy tak, aby mohly být úspěšně uloženy, analyzovány a vizualizovány [3] [5].

Termín není spjatý s žádným konkrétním množstvím dat, většinou se však začíná mluvit o big datech ve chvíli, kdy se jedná řádově od petabajtů dat (peta = 10^{15}) [5].

Přesná definice pojmu big data se neustále vyvíjí, nicméně se vžila definice, kterou poprvé přinesl analytik společnosti *META Group, Dough Laney*. Ten jej definoval jako trojdimenzionální problém *„3V“*. Viz obrázek 2, což jsou počáteční písmena anglických slov:

- **objem** (*angl.* Volume), čili obrovské množství dat,
- **rychlost** (*angl.* Velocity), kterou data každou vteřinou narůstají,
- **různorodost** (*angl.* Variety) a nestrukturovanost dat, která vznikají.



Obrázek 2. Big data z pohledu 3 V.

V důsledku nejasnosti, heterogenity a neúplnosti dat se v poslední době často přidává další, čtvrté *„V“*, které v překladu znamená věrohodnost (*angl.* **Veracity**) [6].

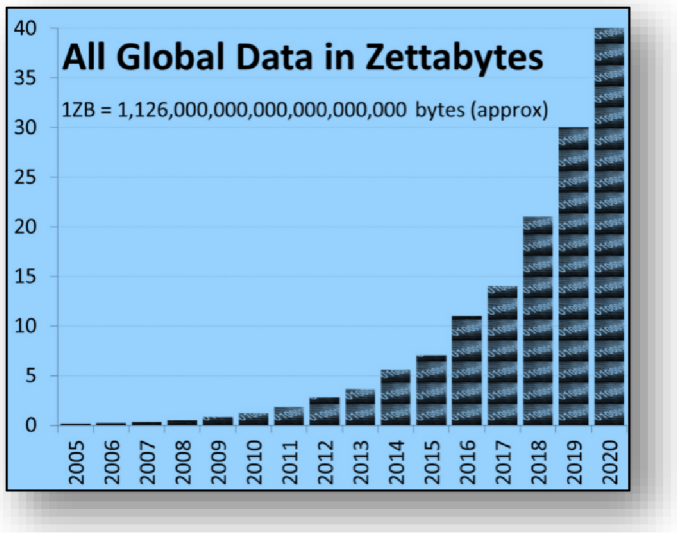
Pokud chceme využít běžnou výpočetní analýzu dat, potřebujeme data ve strukturované podobě. Big data však tuto podmínku většinou nesplňují a nelze k těmto datům přistupovat známými konvenčními metodami a technicky je to velmi obtížné. Proces vyžaduje sběr dat z různých datových skladů a tyto data dál třídit a strukturovat. Rozdíl mezi tradičními daty a big daty je znázorněn v tabulce 1 [11].

Tabulka 1. Rozdíl mezi tradičními daty a big daty [11].

Tradiční data	Big data
Gigabajty až terabajty	Stovky terabajtů, petabajty
Pevná struktura	Nelze zcela strukturovat
Lze jednoduše zjistit souvislosti mezi daty	Nelze vždy zjistit všechny souvislosti mezi daty
Zpracovaná obvykle na jednom počítači	Zpracovaná více počítači zapojenými do sítě
Výsledky dotazů lze vyhodnotit přesně	Dotazy často přesně vyhodnotit nelze *(kvůli rychlím změnám, objemu, času, geografické vzdálenosti úložišť, ...)*

2.2 Big data kolem nás

Problém rychlého a nestrukturovaného ukládání se šíří celou společností – vládní a úřednická data, firemní a průmyslová data, data ve školství, vědě, sportu, médiích apod. Navíc je již přes 3 miliardy lidí připojeno k internetu a sociálním sítím, kteří taktéž produkují objemná data velmi rychle. Na obrázku 3 můžeme vidět statistiku Evropské hospodářské komise OSN (*angl.* United Nations Economic Commission for Europe, UNECE). Na svislé ose vidíme objem dat ve světě v ZB.



Obrázek 3. Celkové množství digitálních dat ve světě v ZB *(dle UNECE)*.

K 9.7.2012 se uvádí, že celkový objem digitálních dat ve světě činilo 2,7 ZB (zetta = 10^{21} bajtů), navíc je z grafu patné, že tento trend výhledově exponenciálně poroste. Jedním z odvětví, kde objemná data přibývají velkou rychlostí je i zdravotnictví [16].

3 Big data ve zdravotnictví

Big data ve zdravotnictví můžeme najít v anglicky psaných článcích také pod pojmem Big Healthcare Data (*zkr.* BHD) nebo Big Data in Healthcare. Zdravotnictví je výborný příklad, kde vzniká velkou rychlostí obrovské množství heterogenních dat. Tato data jsou navíc rozložena mezi klinický provoz, zdravotní pojišťovny, výzkumné pracovníky, vládní a úřední subjekty atd. Data jsou rozeseta po mnoha úložištích, každé z těchto úložišť je navíc zatíženo šumem a neexistuje globální platforma, se kterou bychom k těmto datům přistupovali. V České republice je taktéž velký problém legislativa.

Řízení, zpracování a porozumění big datům ve zdravotní péči je velkou výzvou a existuje zde velký potenciál, na druhou stranu, může být velmi náročné a nákladné. Bez robustní fundamentální teorie pro reprezentaci, analýzu a odvozování závěrů takto složitých dat zůstane cesta k jednotné manipulaci velmi nepřehledná [3] [15].

S těmito problémy se dá definice zdravotnických big dat rozšířit z výše uvedeného „3V“ na definici, která je charakterizovaná šesti problémy:

- velký objem,
- různé zdroje,
- různé stupnice,
- neúplnost,
- nesrovnalost,
- složitost.

Historické přístupy k lékařskému výzkumu se zaměřují především na vyšetření patologických stavů na základě fyziologických změn. Ačkoli je tento přístup k porozumění nemocem nezbytný, je omezený. Vzhledem k množství dat, která máme k dispozici, nedokáže zachytit všechny souvislosti, které by mohly souviset s rozvojem nemocí. Nové technologie umožňují zachytit velké množství informací o jednotlivých pacientech za dlouhou dobu. Mnoho nasbíraných dat zůstalo po dlouhou dobu nedotčeno [3].

Mimo obrazová data, ve zdravotnictví vznikají velké objemy dat mimo jiné v oblastech signálového zpracování, genomice a proteomice. Vznikají databáze, ve kterých je uložen celý genom člověka a sekvenují se další a další DNA sekvence. S rozvojem telemedicíny vzniká nejenom na signálová data požadavek snímat data v reálném čase a tyto data ukládat. Vzniká tak například záznam EKG z celého dne [3].

3.1 Obrazová data ve zdravotnictví

Zobrazovací metody jsou jeden ze základních pilířů v moderním lékařství. Snímky jsou důležitým zdrojem informací převážně k určení diagnózy, plánování a hodnocení terapie. Počítače neslouží pouze k vizualizaci pořízených snímků, taktéž se v nich obrazy vytváří, upravují, rekonstruují a ze série obrazových dat můžeme provádět 3D modely a rekonstrukce [3] [4] [5].

Obrazová data poskytují informace o anatomických strukturách a orgánových funkcích, ve kterých můžeme diagnostikovat patologické stavy. Taktéž se využívají k identifikaci nádorů, segmentaci orgánů, detekci aneurysmatu apod. K těmto aplikacím se nejčastěji využívá strojového učení a technik zpracování obrazu. V současné době je trend spojovat obrazová data s dalšími typy dat, jako jsou signálová nebo genomická data. Toho se využívá například u funkční magnetické rezonance. Integrace lékařských snímků s jinými typy údajů může také zlepšit přesnost a zkrátit dobu potřebnou pro diagnózu [3].

Zobrazovací techniky zahrnují široké spektrum nejrůznějších metod, které se většinou používají pro různé klinické účely. Tato klinická data vznikají na odděleních radiologie, zobrazovacích metod a nukleární medicíny. Přehled nejčastějších zobrazovacích technik vidíme v tabulce 2.

Tabulka 2. Nejčastější zobrazovací techniky ve zdravotnictví.

Zkratka	Význam (anglicky)	Význam (česky)
RTG	*Radiography*	**Konvenční rentgen**
CT	*Computed tomography*	**Počítačová tomografie**
MR	*Magnetic resonance*	**Magnetická rezonance**
MG	*Mamography*	**Mamografie**
US	*Ultrasound*	**Ultrazvuk**
PET	*Positron emissinon tomography*	**Pozitronová emisní tomografie**
SPECT	*Single-Photon Emission Computed Tomography*	**Tomografická scintigrafie**

Nejdéle zaběhnuté je klasické konvenční rentgenové vyšetření, jinak řečeno skiagrafie. Zde se setkáváme s technikami přímé a nepřímé digitalizace. Mimo tyto zobrazovací techniky existují hybridní přístroje, které kombinují více technik dohromady. Například PET-CT nebo PET-MR. Snímky, které jsou pořizované na těchto přístrojích zaznamenávají a ukládají velké množství dalších, doplňujících dat.

Z hlediska dimenze (D) mohou vznikat 2D, 3D i 4D data, v případě 3D rekonstrukce v čase. PET, CT, MR, 3D ultrazvuk jsou považovány za multidimenzionální zobrazovací techniky [3].

Zejména s rozvojem počítačové tomografie začalo vznikat velké množství obrazových dat. Proto musel být zaveden standard, který by tato data uchovával a umožňoval jejich přenos. Standard se nazývá DICOM, viz níže.

4 Analýza zdravotnických big dat (BHD)

Velikost BHD se většinou liší. Studie a analýzy však zahrnují stovky až tisíce jednotlivců, strukturované a nestrukturované datové prvky a metadata, jejichž velikost se může pohybovat od MB až po TB [15].

V uplynulých desetiletích bylo vyvinuto velké úsilí a byla vyvinuta řada datových standardů a slovníků k strukturální a sémantické reprezentaci dat a metadat. Jedním z takovýchto standardů je i DICOM. V analýze BHD se nejčastěji setkáváme se čtyřmi fázemi. S první a poslední fází se setkáváme vždy, druhá a třetí fáze se liší dle případu.

1. rozpoznání složitosti procesu, pochopení struktury dat,
2. vytvoření reprezentativního vzorku dat, na kterém se dají testovat výpočty,
3. modelování dat,
4. interpretace výsledků a vyvození závěrů [15].

Data se z pohledu strukturovanosti mohou ukládat třemi různými způsoby a lze je tedy rozdělit do tří skupin:

- **Data strukturovaná,**
- **Data semi-strukturovaná,**
- **Data nestrukturovaná** [5]**.**

Big data ve zdravotnictví mohou být ukládána všemi třemi možnými způsoby. Nejvíce se však setkáváme s daty nestrukturovanými, což komplikuje tyto data výpočetně zpracovat, protože se jedná pouze o data kvalitativní a nesourodá. Příklady takových dat mohou být hrubá textová data, jako jsou poznámky lékařů, obrázky, video, objemová data, genomické sekvence apod [15].

V souvislosti s problémem nestrukturovaných dat vznikajících na více místech zároveň bylo vyvinuto prostředí pro distribuované zpracování velkých dat – Apache Hadoop [8].

Do DICOM formátu se data ukládají strukturovaně a tato struktura je standardizována. Každá informace uložená do tohoto formátu má svůj jednoznačný identifikátor v podobě tagů. V práci se bude k datům přistupovat pomocí tagů, použití Apache Hadoop tedy není potřeba. Spíše nás bude zajímat které tagy, a v jaké kvalitě přístroje a doktoři zapisují.

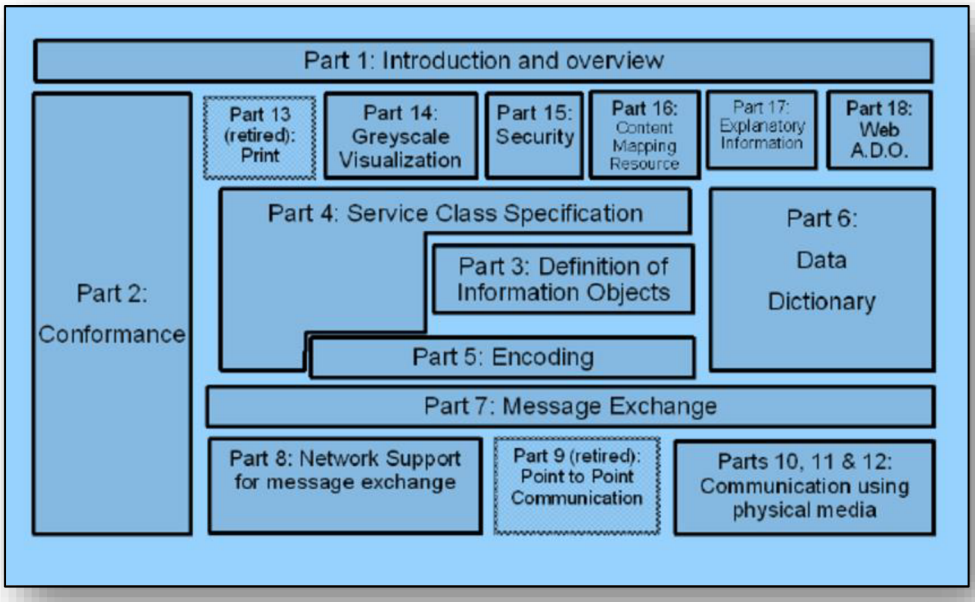
Statistika nabízí velké množství prověřených metod pro analýzu dat. Z pohledu big dat se v analýzách využívají hlavně nelineární statistické metody a vícerozměrné analýzy dat [12].

5 DICOM

DICOM (*angl.* Digital Imaging and Communications in Medicine) je standardizovaný formát pro přenos a uchování obrazové informace ve zdravotnictví, který v roce 1984 definovala a stále vyvíjí americká organizace NEMA (*angl.* National Electrical Manufacturers Association) společně s ACR *(angl.* American College of Radiology) [4]. Pomocí něj se řídí uchování a přenos obrazu ve zdravotnictví po celém světě. Vzniknul z potřeby definovat rozhraní. Díky specifikaci IOD (*angl.* Information Object Definition) si různé aplikace mohou vyměňovat data bez potřeby znalosti aplikací na druhé straně, respektive aby zařízení dokázala přečíst data pořízená z jiných zařízení [7].

Mimo samotnou obrazovou informaci tento formát ukládá mnoho dalších doplňujících informací souvisejících s pořízením obrazu. Některé informace zapisuje sám přístroj, jiné doktor při vyšetření. Předmětem této práce je zmapovat, jaké informace, a v jaké kvalitě se do této hlavičky v České republice ukládají. Tyto data jsou uložené v datové sadě spolu s obrazem, někdy se datová sada uvádí jako hlavička DICOM souboru [2] [7] [17].

DICOM standard je oficiální dokument, který ve 20 kapitolách, na více než 5000 stránkách definuje datové formáty a komunikační protokoly. Je dostupný na oficiálních stránkách DICOM dicom.nema.org, na kterých se dají najít veškeré informace o tomto formátu. Postupně procházel řadou změn a v roce 1992 byla vydána třetí verze tohoto standardu DICOM NEMA PS 3, která je platná dodnes a pořád se vyvíjí [7]. Kapitoly jsou popsány vždy PS 3.X, přičemž X je číslo kapitoly. Na obrázku 4 vidíme jednotlivé části DICOM standardu [13].

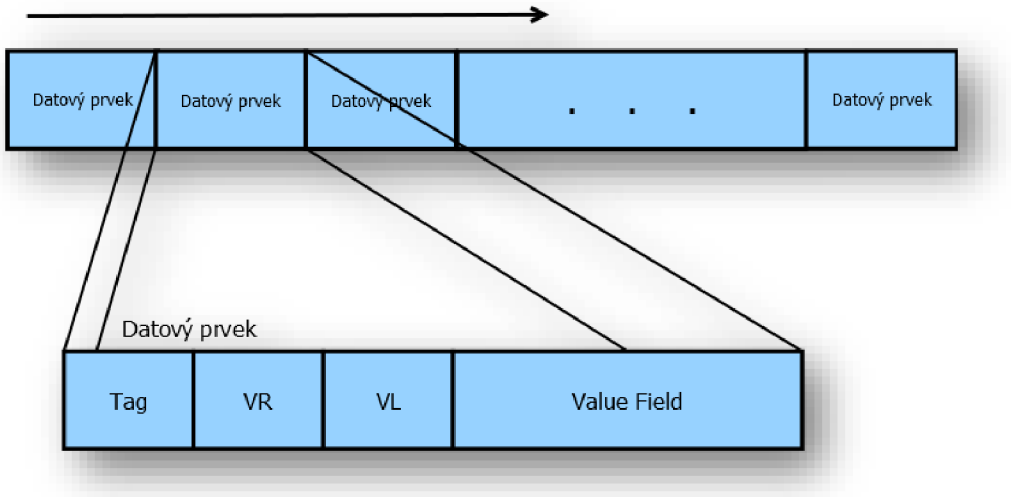


Obrázek 4. Jednotlivé části DICOM standardu.

V naší práci se zabýváme převážně částmi DICOM PS 3.3, kde jsou definovány informační objekty (*angl.* Information Object), tzn. jsou zde popsány významy jednotlivých informací ukládaných do DICOM. PS 3.5, kde je popsaná struktura a zakódování dat a PS 3.6, kde se nachází datový slovník (*angl.* DICOM Dictionary), je zde kompletní výčet tagů, se kterými se v práci dále pracuje.

5.1 Datová sada

Datová sada (*angl.* Data Set) je část DICOM standardu (konkrétně PS 3.5), se kterou budeme v této práci dále pracovat. Jedná se o instanci informačních objektů, skládá se z jednotlivých datových prvků. Na obrázku 5 vidíme, jak takový Data Set vypadá.



Obrázek 5. DICOM Data Set.

5.2 Datové prvky

Datové prvky (*angl.* Data Elements) v DICOM se skládají z těchto polí:

- tag,
- datová reprezentace (*angl.* Value Representation, VR),
- délka dat (*angl.* Value Length, VL),
- pole hodnot (*angl.* Value Field).

Tag představuje jednoznačný identifikátor datového prvku, který se nemůže opakovat v rámci jedné datové sady, VR popisuje formát dat a VL délku dat, v poli hodnot Value Field se nachází samotná uložená data.

5.3 DICOM tagy

Datový prvek je vždy jednoznačně určen svým tagem, jedná se o metadata v Data Set. Metadata jsou data, pomocí kterých můžeme přistupovat k jiným datům. Tag jsou dvě hexadecimální čísla ve formátu viz rovnice 1. Představují jednoznačný identifikátor konkrétní informace, která je pod tímto číslem uložená. V jednom Data Setu může být vždy jen jeden konkrétní tag, ve vložených datových sadách se však mohou opakovat.

$$tag = (xxxx, yyyy) \tag{1}$$

Číslo *xxxx* zařazuje datový prvek do skupiny (tzv. skupinové číslo, *angl.* Group Number). Například *(0010,yyyy)* jsou tagy, které nesou informace o pacientovi. Skupinové číslo taktéž určuje, jestli se jedná o standardizovaný nebo privátní datový prvek. Standardizované datové prvky mají skupinové číslo vždy sudé (výjimku tvoří tagy *(0000,yyyy)*, *(0002,yyyy)*, *(0004,yyyy)*, *(0006,yyyy)*, které jsou rezervované, např. *(0002,yyyy)* jsou MetaInfo). Privátní datové prvky mají skupinové číslo vždy liché, nejsou standardizované, takže z nich nemůžeme určit, jaké informace se pod těmito tagy ukrývají.

Tabulka 3. Skupiny DICOM metadat.

Skupina tagů	Název skupiny
(0080,yyyy)	Informace o studii
(0010,yyyy)	Informace o pacientovi

Číslo *yyyy* (tzv. číslo prvku, *angl.* Element Number) je již konkrétní prvek ve skupině s konkrétní informací a hodnotou. Tedy například tag *(0010,0010)* bude vždy obsahovat informaci o pacientově jménu [7].

V DICOM standardu je tagů definováno více jak 2000. Každý definovaný tag má svoje jméno (*angl.* Tag Name) a DICOM standard taktéž určuje, jaký formát bude datový prvek pod konkrétním tagem mít. Úplný seznam všech definovaných tagů, který nazýváme DICOM Dictonary, respektive DICOM Lookup, najdeme v šesté kapitole standardu (PS 3.6). Některé tagy jsou určeny pro konkrétní typ modality, jiné jsou pro všechny studie stejné.

5.4 Datová reprezentace

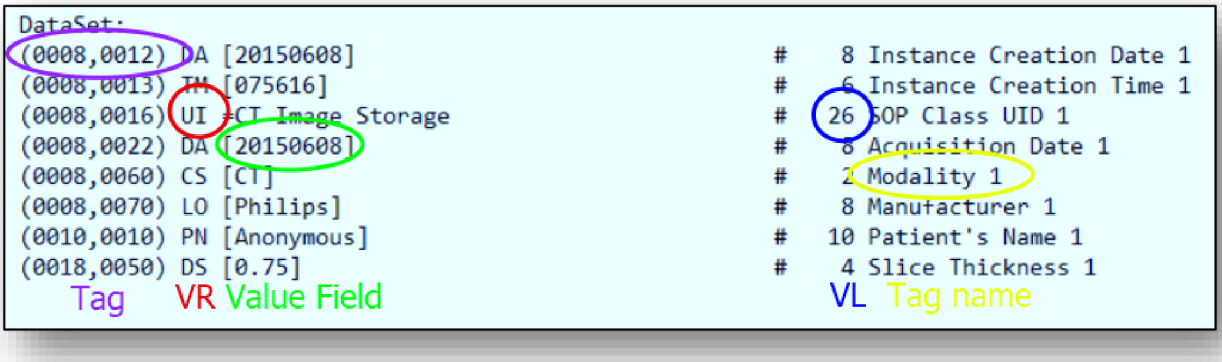
DICOM standard taktéž určuje, jaký formát bude Value Field mít. Formát se nazývá datová reprezentace (*angl.* Value Representation, VR). V datovém prvku se jedná o dvoubajtový znakový řetězec a je implicitně určen tagem. Dle IHE je doporučeno VR explicitně uvádět vždy. V tabulce 4 je výňatek jednotlivých VR hodnot a názvem co znamenají. Celý výpis VR s definicemi a charakteristikami nalezneme ve standardu PS 3.6.1-1.

Tabulka 4: Výňatek VR

VR	Název	VR	Název	VR	Název
AE	Application Entiti	DS	Decimal String	OW	Other Word String
AS	Age Strings	DT	Date Time	TM	Time
AT	Attribute Tag	IS	Integer Strimg	PN	Person Name
CS	Code String	LO	Long String	UI	Unique Identifer
DA	Date	LT	Long Text	UT	Unlimited Text

5.5 Užitečná data v hlavičkách DICOM souborů

Na obrázku 6 můžeme vidět ukázku datové sady, která představuje naše vstupní data k analýze. V tomto dumpu je taktéž uveden vždy název příslušného tagu. Do naší analýzy budou vstupovat jenom tagy a pole hodnot Value Field. Práce se zabývá zmapováním, které Value Field a v jaké kvalitě bývají nejčastěji vyplněna.



Obrázek 6. Příklad našich vstupních DICOM dat.

V hlavičce souboru se mohou ukrývat zajímavá data, která by mohla mít klinickou nebo výzkumnou hodnotu, a to jak z hlediska lékařského, tak provozního. Přístup k těmto datům vede právě přes metadata skrze DICOM tagy [5] [17].

Většina vědeckých prací na téma big data v lékařském zobrazování a radiologii se zabývá diagnostikou artefaktů ve vlastních obrázcích, tedy hledáním patologických tkání apod. Neexistuje studie, která by v České republice mapovala data a kvalitu dat v těchto hlavičkách.

6 DICOM z pohledu big data

Veškerá vyšetření pacientů, u kterých vznikne obrazový materiál, se dále ukládají do archivů obrazové zdravotnické dokumentace. Vzhledem k tomu, že zdravotnická zařízení musí povinně archivovat tato vyšetření, postupně se tyto archivy zaplňují velkým množstvím dat.

6.1 Statistika v České republice

Kolik je v České republice v těchto archivech přesně uloženo dat zmapované není, můžeme však vyjít ze statistik, které zpracovává **Ústav zdravotnických informací a statistiky ČR** (*zkr.* ÚZIS). V publikacích:

1. ***Činnost společných vyšetřovacích a léčebných složek***, která se zpracovává vždy za určité období, je nejaktuálnější statistika z let 2007-2015,
2. ***Činnost zdravotnických zařízení ve vybraných oborech***, která je podrobnější a shrnuje vždy každý rok zvlášť, je nejaktuálnější statistika za rok 2012.

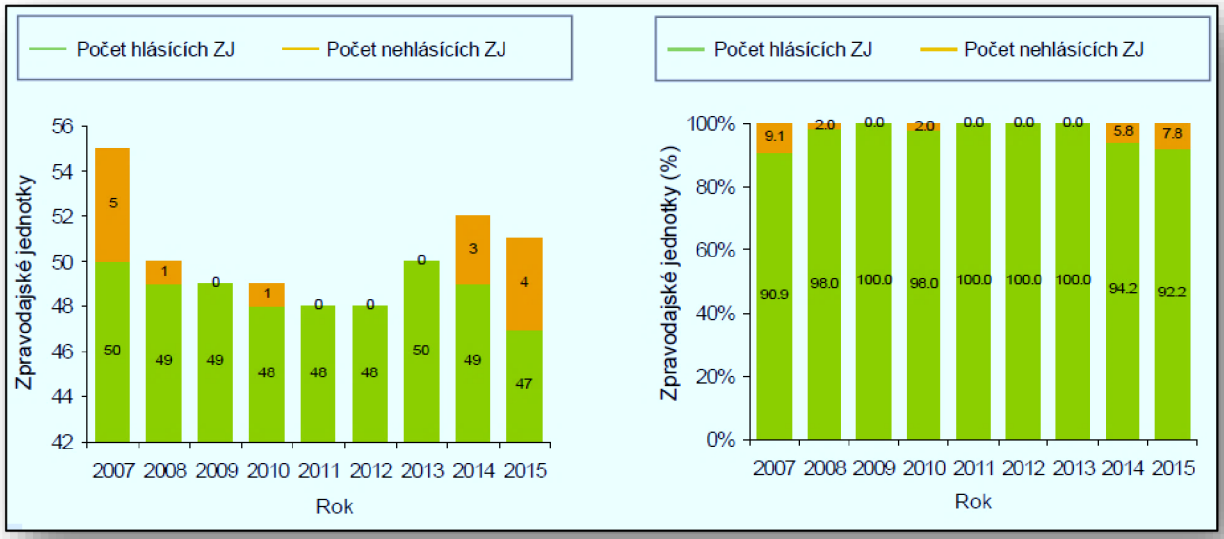
V těchto publikacích jsou kapitoly, které mapují a shrnují mimo jiné i činnosti oborů:

1. radiologie a zobrazovací metody,
2. nukleární medicína.

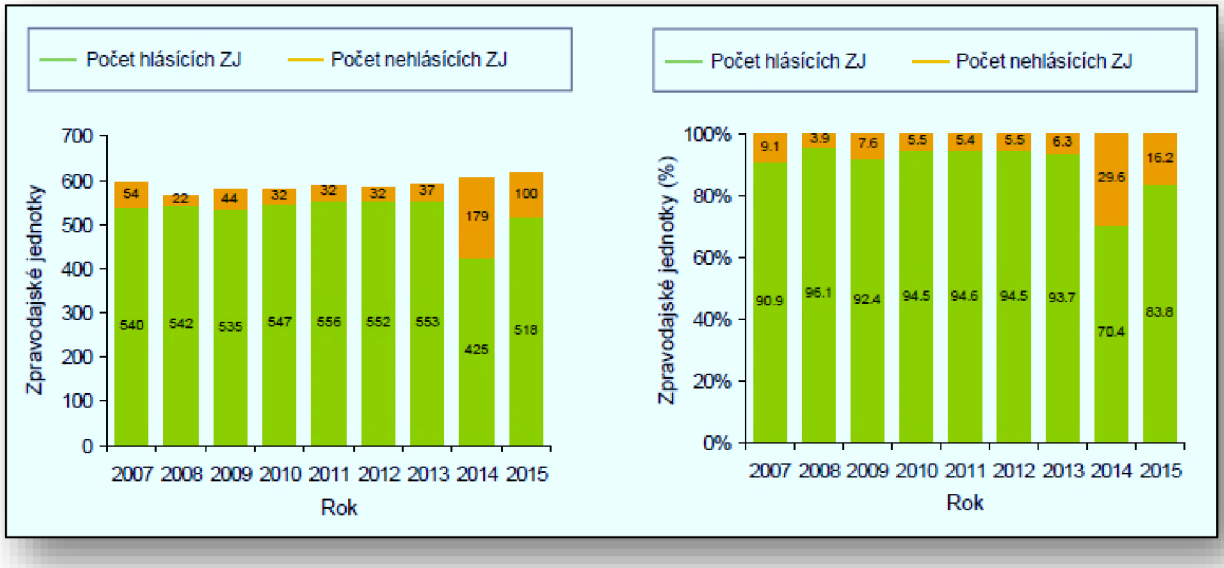
Podklady k těmto statistikám jsou roční výkazy o činnosti zdravotnických zařízení. Vyplnění výkazů však nejsou povinná, takže statistiky nejsou úplné. Na obrázku 7 a 8 můžeme vidět vyplněnost výkazů za období 2007-2015. První graf je vždy uveden v absolutních číslech, druhý graf pak v procentech. Z grafů je patrné, že se vyplněnost výkazů pohybuje většinou nad 90 %, v roce 2014 a 2015 vidíme pokles [10].

Na obrázku 9 je graf, který shrnuje celkový počet vyšetření na oddělení radiologie a zobrazovacích metod za toto období. Trend počtu vyšetření na oddělení radiologie a zobrazovacích metod mírně stoupá. V roce 2007 proběhlo 13 160 329 vyšetření, v roce 2015 už 15 062 391 [10]. Na obrázku 10 se jedná o stejné porovnání, avšak pro oblast nukleární medicíny.

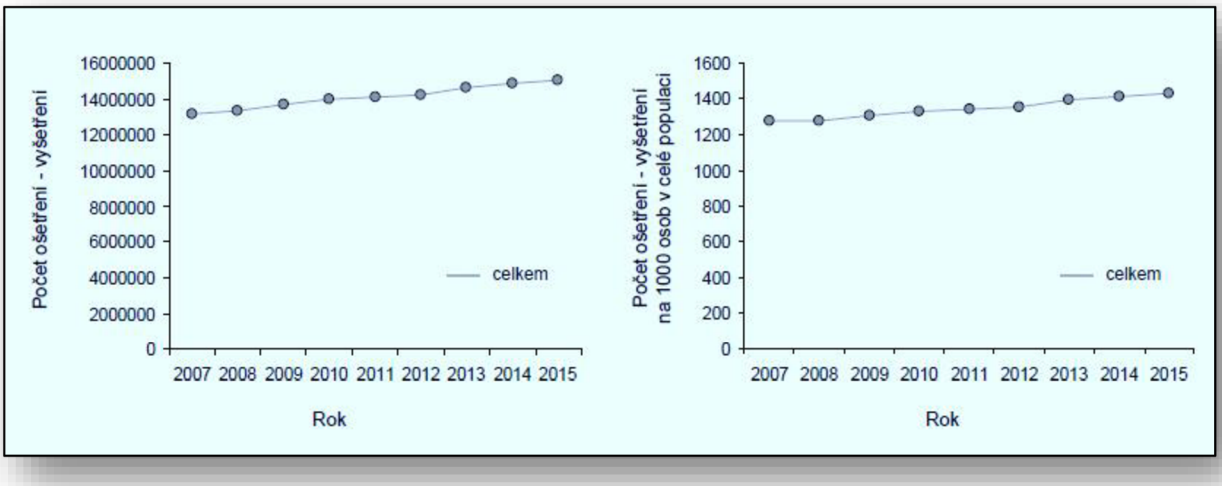
V tabulce 5 vidíme skladbu jednotlivých vyšetření na oddělení radiologie a zobrazovací metody za rok 2012. Tabulku o skladbě výkonů (*in vivo*) na odděleních nukleární medicíny najdeme v příloze 4. Na obrázcích 11 a 12 pak tuto skladbu vidíme znázorněnou graficky v procentech.



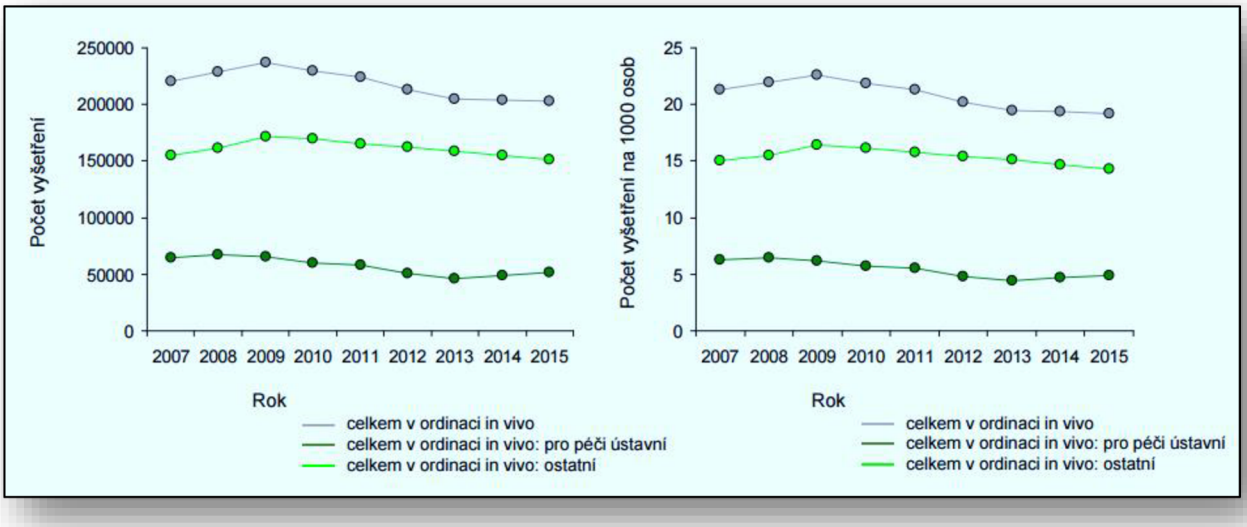
Obrázek 7. Vyplněnost výkazů v oborech radiologie a zobrazovací metody [10].



Obrázek 8. Vyplněnost výkazů v oborech nukleární medicína [10].



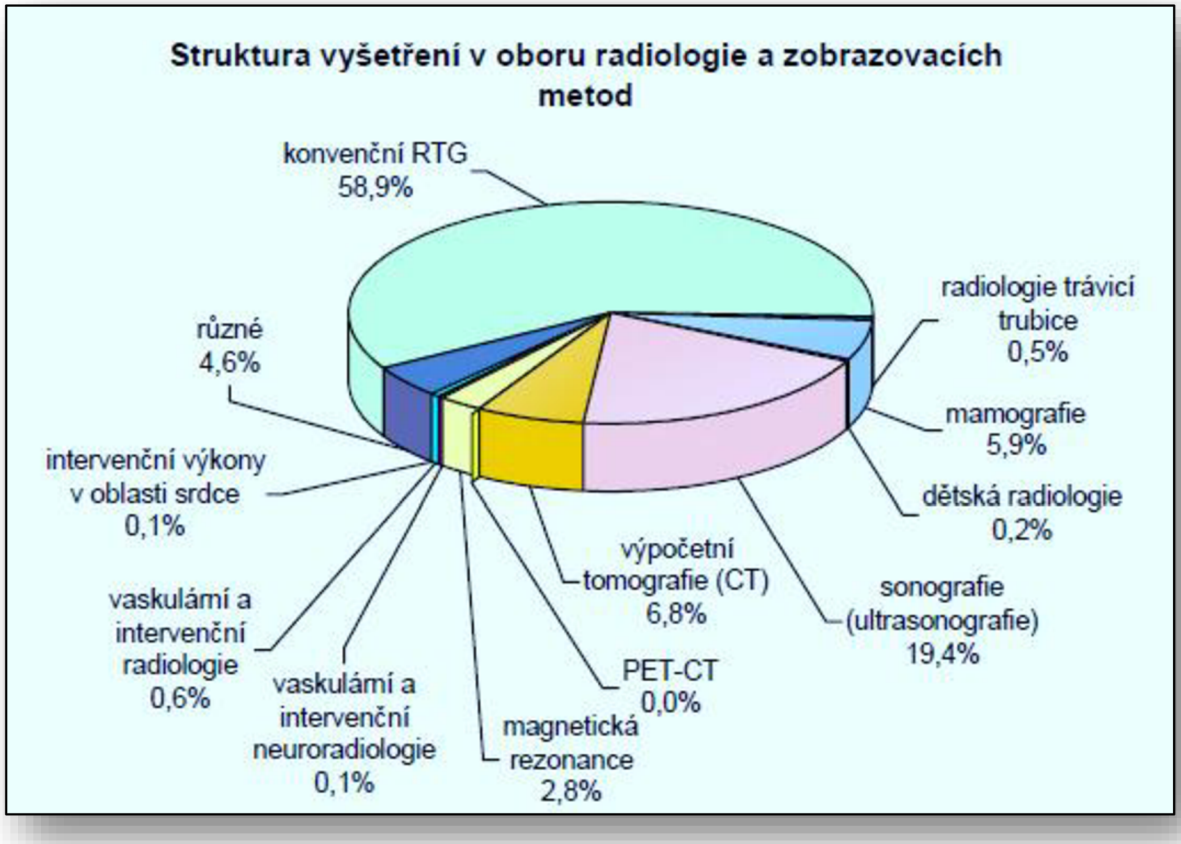
Obrázek 9. Počet vyšetření na oddělená radiologie a zobrazovací metody (2007–2015) [10].



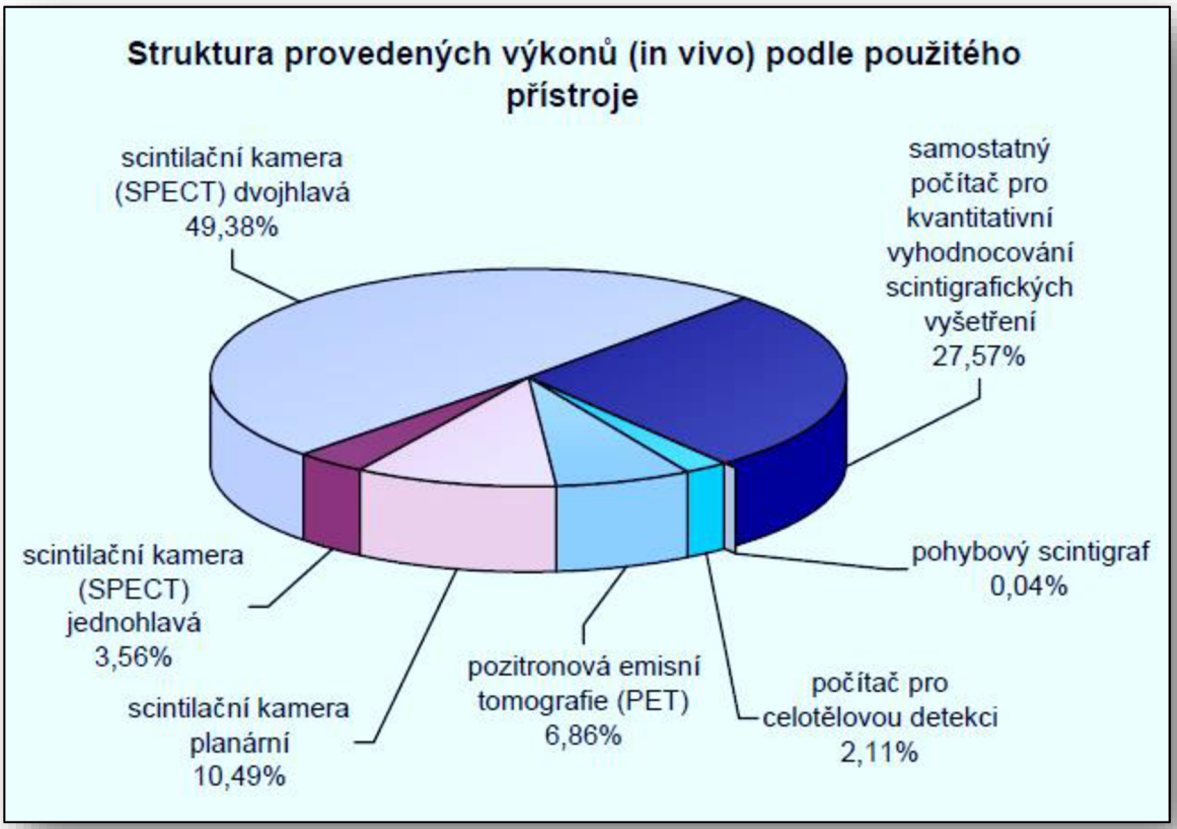
Obrázek 10. Počet diagnostických vyšetření na odděleních nukleární medicíny (2007–2015) [10].

Tabulka 5: Skladba vyšetření na radiologických odděleních a pracovištích

Druh vyšetření	Počet vyšetření				
	celkem		z toho (v %)		
	absolutně	na 1 úvazek pracovníka	v ambulantní části lůžkových zařízení	v samostat. ord. lékařů specialistů	v ostatních zdravotnických zařízeních
Konvenční RTG	8 382 887	1 758,1	70,1	17,4	12,5
Radiologie trávicí trubice	73 072	15,3	90,7	5,7	3,6
Mamografie	842 250	176,6	50,2	35,6	14,1
Dětská radiologie	26 680	5,6	59,9	36,2	3,9
Sonografie (ultrasonografie)	2 765 066	579,9	65,5	25,1	9,4
Výpočetní tomografie (CT)	960 797	201,5	94,3	3,0	2,7
PET-CT	5 583	1,2	-	-	-
Magnetická rezonance	391 198	82,0	80,4	12,6	7,1
Vaskulární a intervenční neuroradiologie	20 701	4,3	100,0	-	-
Vaskulární a intervenční radiologie	87 598	18,4	97,1	2,5	0,4
Intervenční výkony v oblasti srdce	8 224	1,7	100,0	-	-
Různé	659 693	138,4	78,6	13,4	7,9
Celkem	**14 223 749**	**2 983,0**	**70,7**	**18,5**	**10,8**



Obrázek 11. Struktura vyšetření v oboru radiologie a zobrazovacích metod za rok 2012 [14].



Obrázek 12. Struktura vyšetření v oboru radiologie a zobrazovacích metod za rok 2012 [14].

6.2 DICOM z pohledu 3V

Jak již bylo zmíněno, big data se dají charakterizovat pojmem 3V. Data ukládaná do archivů obrazové zdravotnické dokumentace ve formátu DICOM bychom takto mohli taktéž charakterizovat.

Objem (Volume)

Každý vyprodukovaný DICOM soubor pořízený různou modalitou má různou velikost. Na oficiálních webových stránkách projektu DICOM Library (http://www.dicomlibrary.com), který má za cíl bezplatně sdílet anonymizované informace, obrazy a signály pro vzdělávací a vědecké účely, můžeme najít tabulku (tabulka 6), ve které je vypočítána průměrná velikost snímku v MB z různých modalit. Během 6 let nasbírali od 230 000 uživatelů z 204 zemí přes 321 000 studií, což představuje reprezentativní vzorek.

Tabulka 6. Rozdíly ve snímcích z různých modalit.

Modalita	Popis	Rozlišení (px)	Bitů na px	Velikost (MB)	Počet snímků
CD	Color flow Doppler	768 x 576	8	0,442	
CR	Computed radiography	3520 x 4280	12	30	2
CT	Computed tomography	512 x 512	16	0,524	40-3000
DSA	Digital Subtraction Angiography	512 x 512	8		15-40
DX	Digital Radiography	2048 x 2048	12		2
MG	Mammography	4608 x 5200	14	45,7	1
MR	Magnetic Resonance	256 x 256	16	0,131	60-3000
NM	Nuclear Medicine	256 x 256		0,128	
PET	Positron Emission Tomography	128 x 128		32	
US	Ultrasound	512 x 512	8	0,262	20-240
XA	X-Ray Angiography	512 x 512	16		

Z tabulky je patrné, že se rozlišení snímků liší s různou modalitou. Největší nárok na rozlišení je u mamografických vyšetření. Velikost souborů taktéž bude větší, pokud se studie skládá ze série snímků, jako je tomu u CT a MR vyšetření.

Je známo, že se velikost souborů liší nejenom typem modality, ale taktéž s různým typem vyšetření. Dle Centra informatiky Fakultní nemocnice Brno (FN Brno) se u klasických CR, DX, MG, MR vyšetření velikost pohybuje řádově od jednotek MB až po 2 GB, v případě video smyčky srdce. U hybridních modalit jako jsou PET-MR a PET-CT se velikost pohybuje od 15 GB výš, v závislosti na délce a obsahu záznamu (část těla nebo celé tělo v případě polytraumatu).

V tabulce 5 můžeme vidět skladbu jednotlivých vyšetření na odděleních radiologie a zobrazovací metody. Pokud bychom vzali v úvahu velikosti souborů z tabulky 6 a vynásobili je počtem jednotlivých vyšetření, dostáváme se do řádu petabajtů za jeden rok.

Z grafů taktéž vidíme, že nejvíce vyšetření (více jak 50 %) zaujímá klasické konvenční rentgenové vyšetření.

Rychlost (Velocity)

Jak již bylo výše zmíněno, trend v počtu vyšetření na odděleních radiologie a zobrazovací metody je mírně stoupající. Počet vyšetření se pohybuje okolo 14 – 15 milionů ročně. Nemocnice musí archivovat tato vyšetření několik let. V archivu FN Brno se za období 2002-2017 nastřádalo z 1 711 362 studií 37 TB dat. V archivu se však nachází duplicitní obsah, kvůli používání několika úložišť, čistých dat tedy bude méně. V nekomprimované podobě je měsíční přírůstek dat okolo 250-350 GB, komprese sníží velikost o polovinu až třetinu. Roční přírůstek v tomto archivu je okolo 5-7 TB. Letos do FN Brno přibyla velkokapacitní modalita PET-MR – 20-40GB na studii a to pohne s plněním archivu. Celkově je zaznamenán požadavek na archivaci dalších modalit a zařízení, které dříve data do PACS archivu neukládaly (např. EKG). Navíc, v rámci požadavku na ověřování dokumentace dle standardů EU GDPR (ochrana osobních údajů) a zavedením elektronického podpisu dokumentace, bude požadováno archivovat větší množství dat. Důvodem je, že jakákoliv změna v dokumentaci vede k automatické duplikaci dat. Vytvoří se klon originálu dokumentace, nad kterým jsou následně realizovány všechny změny. Takových vyšetření je a bude poměrně velké množství, zvláště u novorozenců, kde je nutné opravovat identifikaci po přidělení rodného čísla matrikou.

Různorodost (Variety)

Technické řešení ukládání dat je v každém zdravotnickém zařízení jiné. Tato část není předmětem této práce.

Různorodost z pohledu hlavičky souboru DICOM

Jednotlivé informace jsou sice standardizovány pomocí tagů, u většiny ale není nikde řečeno, jak se mají vyplňovat. Například tag *0008,1030*, kam doktoři zapisují popis studie, bývá dosti často vyplněný různě i když se jedná o ten samý typ vyšetření, protože každý doktor jej zapíše trochu jinak. Tag *0008,0080*, kam se zapisuje jméno zdravotnického zařízení, bývá různě vyplněn u různých modalit z té samé nemocnice.

Taktéž není zmapováno, které tagy vyplněné jsou a které ne. Mnoho parametrů zapisuje sám přístroj a to, jaké informace bude zapisovat, se rozhodne při instalaci zařízení.

7 Použitá data

Analýze bylo podrobeno 1215 studií, které poskytlo Centrum informatiky Fakultní nemocnice Brno (FN Brno). Jedná se o anonymizovaná data zachycená projekty pro výměnu obrazových dat mezi zdravotnickými zařízeními během dvou měsíců na Whirpool serveru FN Brno. FN Brno je spádovou oblastí pro Jihomoravský kraj a je komplexním centrem pro dětské pacienty (areál dětské nemocnice) pro výměnu obrazových dat projekty ePacs a ReDiMed (viz níže). Jedná se tedy o data ze zdravotnických zařízení, které jsou účastníky těchto projektů.

7.1 Legislativa

V prvotní fázi diplomové práce byl velký problém s daty. Firma OR-CZ spol. s.r.o. sice spravuje a archivuje velké množství DICOM dat, majitelé dat jsou však nemocnice. U projektu MeDiMed, který vzniknul na Ústavu výpočetní techniky MU se zdálo, že by data být poskytnuta mohla, protože spravují data nemocnic spadající pod stejnou univerzitu, nastal však stejný problém – majitelé dat jsou nemocnice.

Nakonec se v dubnu podařilo data získat od Centra informatiky FN Brno. Do DICOM souboru se však ukládají i citlivá data pacientů. K této problematice se váže následující legislativa:

- Zákon č. 101/2000 Sb., o ochraně osobních údajů a o změně některých zákonů, ve znění pozdějších předpisů
- Zákon č. 372/2011 Sb., o zdravotních službách a podmínkách jejich poskytování, ve znění pozdějších předpisů
    - Vyhláška Ministerstva zdravotnictví č. 98/2012 Sb., o zdravotnické dokumentaci, ve znění pozdějších předpisů

Data byla před poskytnutím anonymizována. V datech chybí jména pacientů a jejich rodné číslo, takže nejsou v rozporu s touto legislativou.

7.2 Projekty ePacs a ReDiMed

Radiologické komunikační centrum ReDiMed a ePacs jsou projekty, které mají za cíl rychlou a zabezpečenou výměnu a přenos obrazových dat mezi zdravotnickými zařízeními.

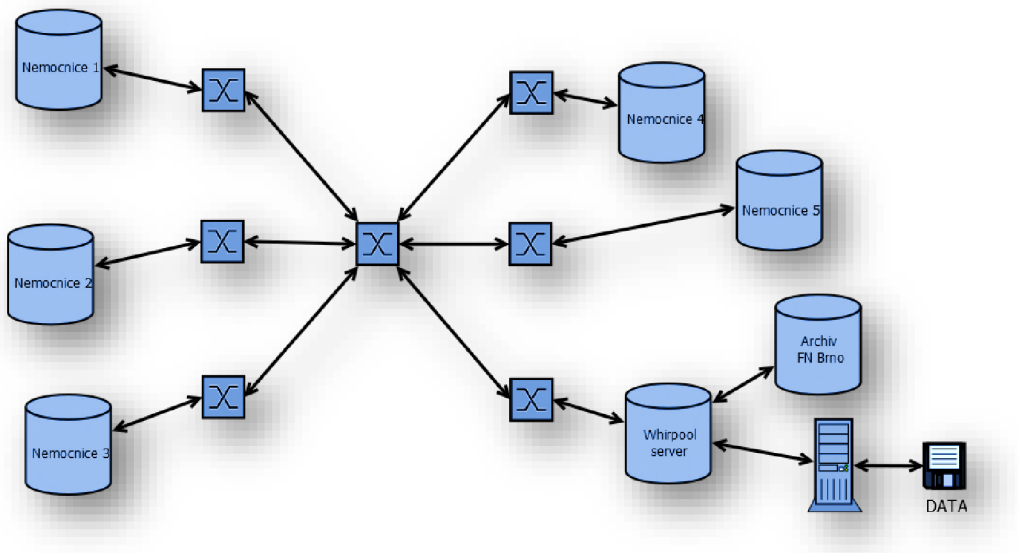
Organizátorem a garantem projektu ePacs je Koordinační středisko pro resortní zdravotnické informační systémy z pověření Ministerstva zdravotnictví ČR, které tuto roli převzalo od Všeobecné fakultní nemocnice v roce 2016. Systém propojení vyvinula firma *ICZ a.s.* Celkový počet připojených zdravotnických zařízení v systému ePacs je 321 *(ke 27.4.2017)* z celé České republiky.

Oproti tomu ReDiMed vzniknul pod záštitou Masarykovy univerzity. Její správu zajišťuje Ústav výpočetní techniky MU a v projektu jsou zapojeny hlavně zdravotnická zařízení z Prahy a Moravy.

Kompletní mapu zdravotnických zařízení, které jsou připojeny na projekty ePacs a ReDiMed, lze najít v příloze 3.

7.3 Whirpool FN Brno

Whirpool FN Brno je server, který ukládá data z projektů ePacs a ReDiMed. Server zajišťuje archivaci zaslaných obrazových dat do hlavního PACS archivu FN Brno v případě, že je dokumentovaný pacient léčen ve FN Brno nebo je zde prováděna konzultace. Whirpool server pro přístup a archivaci využívá aplikaci *Conquest DICOM software*, který je freeware. Pomocí něho se dá vytvářet databáze a k této databázi jednoduše přistupovat. Zařazení Whirpool serveru v celém systému můžeme vidět na obrázku 13.



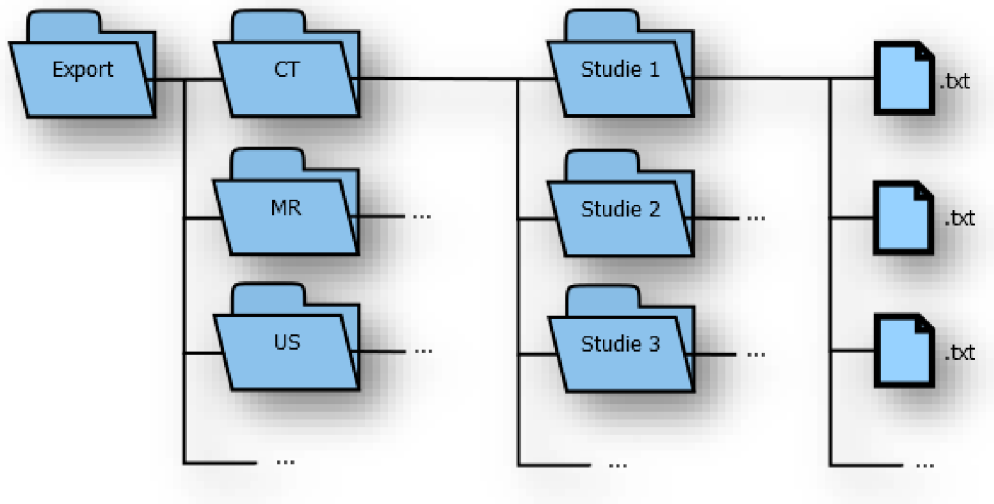
Obrázek 13. Zařazení Whirpool serveru v projektech ePacs a ReDiMed.

7.4 Data

Data jsou hlavičky obrazových DICOM souborů. Shrnuté informace o datech jsou v tabulce 7. Data byla vyexportována do složek, které byly rozděleny dle typu modalit (CR, CT, MR, …). V každé složce byly podsložky s jednotlivými studiemi, v nich pak soubory z dané studie, ve formátu `.dcm`. Množství souborů záviselo na modalitě a studii. Například z běžného CR vyšetření se jednalo o 1 až 3 soubory pro jednu studii. U CT vyšetření se jednalo o stovky až tisíce souborů. Všechny soubory se hromadně přeformátovaly do formátu `.txt`. Struktura vyexportovaných dat z Whirpool serveru je znázorněna na obrázku 14. Vlastnosti komprimované složky se všemi daty nalezneme v příloze 2.

Tabulka 7. Informace o datech.

Počet studií	1215
Počet souborů	400 717
Celková velikost	258 525 938 666 bajtů (260 GB)
Komprimovaná velikost	26 810 153 625 bajtů (26 GB)
Kompresní poměr	10 %



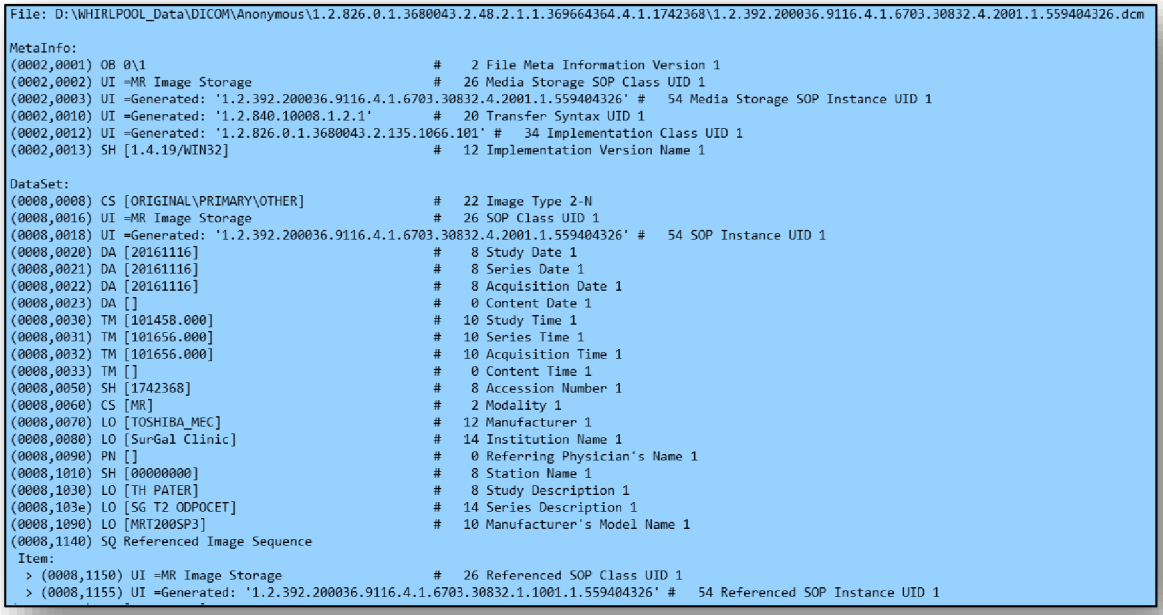
Obrázek 14. Struktura vyexportovaných dat.

8 Zpracování dat

Ke zpracování dat byly využity nástroje MATLAB, Excel a Total Commander. Klíčové tagy, které budeme analyzovat se v rámci jedné studie nemění, proto byla analýza provedena na datech tak, že byl vždy vybrán jeden soubor z každé studie.

8.1 Načítání dat

Vzhledem k povaze dat, která se nacházela v textových souborech, se musely jednotlivé informace rozdělit do skupin. Příklad textového souboru je na obrázku 15.



```
File: D:\WHIRLPOOL_Data\DICOM\Anonymous\1.2.826.0.1.3680043.2.48.2.1.1.369664364.4.1.1742368\1.2.392.200036.9116.4.1.6703.30832.4.2001.1.559404326.dcm

MetaInfo:
(0002,0001) OB 0\1                                  #    2 File Meta Information Version 1
(0002,0002) UI =MR Image Storage                    #   26 Media Storage SOP Class UID 1
(0002,0003) UI =Generated: '1.2.392.200036.9116.4.1.6703.30832.4.2001.1.559404326' #   54 Media Storage SOP Instance UID 1
(0002,0010) UI =Generated: '1.2.840.10008.1.2.1'    #   20 Transfer Syntax UID 1
(0002,0012) UI =Generated: '1.2.826.0.1.3680043.2.135.1066.101' #   34 Implementation Class UID 1
(0002,0013) SH [1.4.19/WIN32]                       #   12 Implementation Version Name 1

DataSet:
(0008,0008) CS [ORIGINAL\PRIMARY\OTHER]             #   22 Image Type 2-N
(0008,0016) UI =MR Image Storage                    #   26 SOP Class UID 1
(0008,0018) UI =Generated: '1.2.392.200036.9116.4.1.6703.30832.4.2001.1.559404326' #   54 SOP Instance UID 1
(0008,0020) DA [20161116]                           #    8 Study Date 1
(0008,0021) DA [20161116]                           #    8 Series Date 1
(0008,0022) DA [20161116]                           #    8 Acquisition Date 1
(0008,0023) DA []                                   #    0 Content Date 1
(0008,0030) TM [101458.000]                         #   10 Study Time 1
(0008,0031) TM [101656.000]                         #   10 Series Time 1
(0008,0032) TM [101656.000]                         #   10 Acquisition Time 1
(0008,0033) TM []                                   #    0 Content Time 1
(0008,0050) SH [1742368]                            #    8 Accession Number 1
(0008,0060) CS [MR]                                 #    2 Modality 1
(0008,0070) LO [TOSHIBA_MEC]                        #   12 Manufacturer 1
(0008,0080) LO [SurGal Clinic]                      #   14 Institution Name 1
(0008,0090) PN []                                   #    0 Referring Physician's Name 1
(0008,1010) SH [00000000]                           #    8 Station Name 1
(0008,1030) LO [TH PATER]                           #    8 Study Description 1
(0008,103e) LO [SG T2 ODPOCET]                      #   14 Series Description 1
(0008,1090) LO [MRT200SP3]                          #   10 Manufacturer's Model Name 1
(0008,1140) SQ Referenced Image Sequence
 Item:
  > (0008,1150) UI =MR Image Storage                #   26 Referenced SOP Class UID 1
  > (0008,1155) UI =Generated: '1.2.392.200036.9116.4.1.6703.30832.1.1001.1.559404326' #   54 Referenced SOP Instance UID 1
```

Obrázek 15: Příklad textového souboru, ze kterého se separovaly data.

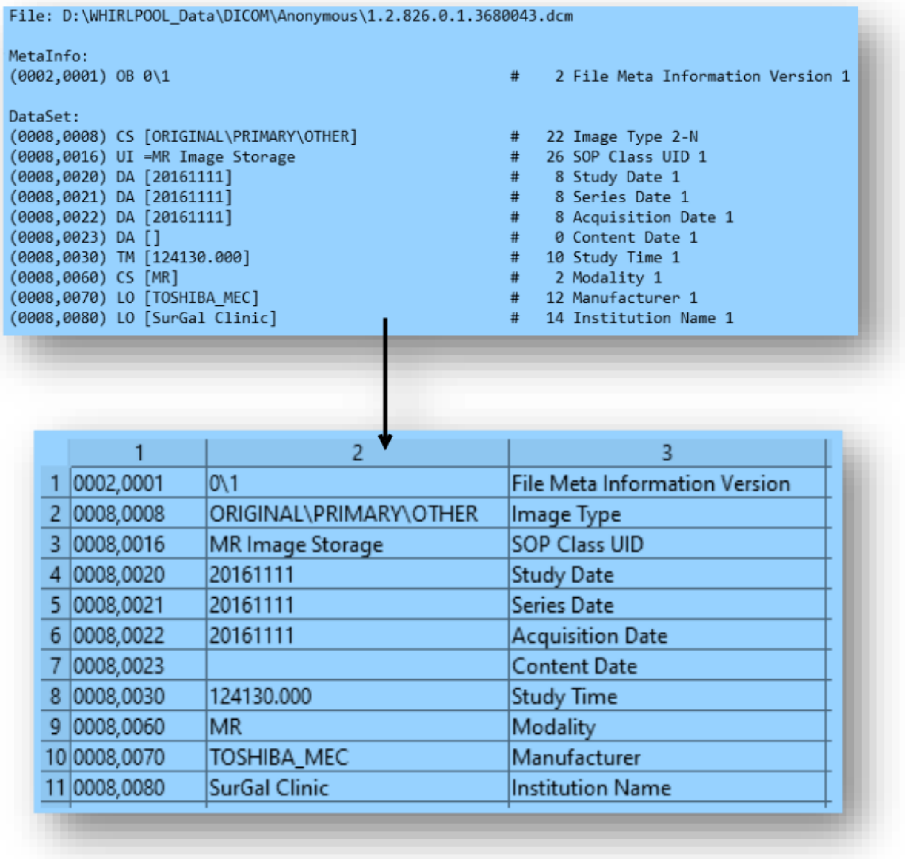
Ze souborů bylo potřeba vyseparovat vždy čísla tagů a k nim příslušné hodnoty. K této problematice se přistoupilo pomocí regulárních výrazů, které představují masku pro data v textovém souboru. Regulérní výraz, který byl využit k separaci dat, je shrnut v tabulce 8.

Funkce `Reading_file`

K separování dat z jednoho souboru byla vytvořená funkce `Reading_file` se vstupní proměnnou - název souboru, který načítáme a cestu ke složce, ve které se soubor nachází. Funkce čte po řádcích daný soubor a postupně separuje data a načítá je do buněk. Výstupní proměnnou jsou pak separovaná data, která potřebujeme. Ukázku, jak funkce pracuje vidíme na obrázku 16.

Tabulka 8. Regulérní výraz použitý k separaci dat a vysvětlení jednotlivých částí výrazu.

Regulérní výraz	
$.*\(([[a-fA-F0-9]\{4\},[a-fA-F0-9]\{4\})\)\s[A-Z]\{2\}\s(.*)\#\s*\d*[\s]*(.*)\s.*$	
Rozdělení výrazu	**Význam ve výrazu**
$.*\($	Definice nepotřebných znaků
$([a-fA-F0-9]\{4\},[a-fA-F0-9]\{4\})$	Separování tagu
$\)\s[A-Z]\{2\}\s$	Definice nepotřebných znaků mezi tagem a Value
$(.*)$	Separování Value
$\#\s*\d*[\s]*$	Definice nepotřebných znaků meziValue a Tag Name
$(.*)$	Separování Tag Name
$\s.*$	Definice nepotřebných znaků



```
File: D:\WHIRLPOOL_Data\DICOM\Anonymous\1.2.826.0.1.3680043.dcm

MetaInfo:
(0002,0001) OB 0\1                                      #    2 File Meta Information Version 1

DataSet:
(0008,0008) CS [ORIGINAL\PRIMARY\OTHER]                 #   22 Image Type 2-N
(0008,0016) UI =MR Image Storage                        #   26 SOP Class UID 1
(0008,0020) DA [20161111]                               #    8 Study Date 1
(0008,0021) DA [20161111]                               #    8 Series Date 1
(0008,0022) DA [20161111]                               #    8 Acquisition Date 1
(0008,0023) DA []                                       #    0 Content Date 1
(0008,0030) TM [124130.000]                             #   10 Study Time 1
(0008,0060) CS [MR]                                     #    2 Modality 1
(0008,0070) LO [TOSHIBA_MEC]                            #   12 Manufacturer 1
(0008,0080) LO [SurGal Clinic]                          #   14 Institution Name 1
```

	1	2	3
1	0002,0001	0\1	File Meta Information Version
2	0008,0008	ORIGINAL\PRIMARY\OTHER	Image Type
3	0008,0016	MR Image Storage	SOP Class UID
4	0008,0020	20161111	Study Date
5	0008,0021	20161111	Series Date
6	0008,0022	20161111	Acquisition Date
7	0008,0023		Content Date
8	0008,0030	124130.000	Study Time
9	0008,0060	MR	Modality
10	0008,0070	TOSHIBA_MEC	Manufacturer
11	0008,0080	SurGal Clinic	Institution Name

Obrázek 16. Ukázka separace dat z textového souboru.

Funkce `Reading_all`

K hromadnému načtení více studií byla vytvořena funkce `Reading_all`, která v cyklu načítá pomocí funkce `Reading_file` všechny soubory ze složky, ve které máme data do struktury `Files`. Vstupní proměnné této funkce jsou cesta do složky, ve které se funkce nachází a cesta do složky, ve které jsou data.

Funkce `tag_filter`

K datům ve struktuře `Files` bylo potřeba nějak přistoupit a dále s nimi pracovat. Proto pro další práci s tagy a jejich hodnotami byla vytvořena univerzální funkce `tag_filter`. Tato funkce má více vstupních proměnných. První proměnnou je název struktury, ze které chceme data dále zpracovávat, v našem případě `Files`. Další dvě proměnné jsou tag a hodnota, kterou má obsahovat. Poslední je pole tagů, které chceme mít ve výstupním poli. Pokud tedy zadáme,

```
array = tag_filter(Files, '0008,0060', 'CT', [{'0008,0070'}, {'0008,0080'}]);
```

funkce ze struktury `Files` načte do pole `array` všechny hodnoty tagů *0008,0070* a *0008,0080* ze studií, kde se tag *0008,0060* rovná CT. Tedy všechny CT studie s hodnotami příslušných tagů. Pro lepší představu je v příloze 6 ukázka výstupu této funkce. Funkce je dynamická, pole výstupních tagů je neomezené, je jedno jestli zadám dva tagy nebo sto. To výrazně usnadňuje další práci s daty.

Funkce `dicomdict_read`

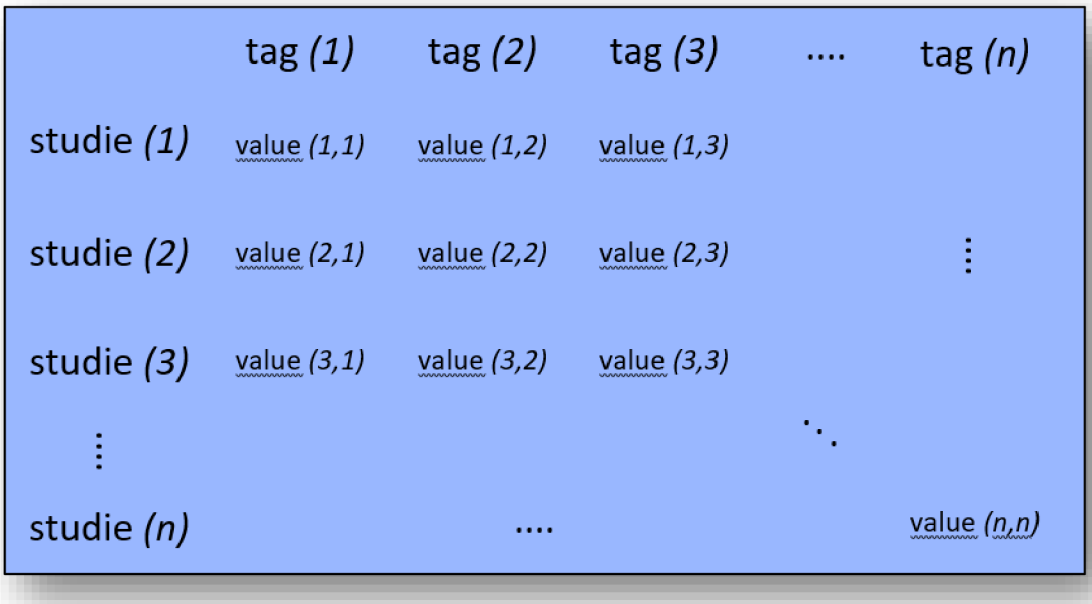
K načtení celé DICOM lookup byla vytvořena funkce `dicomdict_read`. Funkce slouží k načtení všech standardizovaných tagů a jejich popisu z DICOM Dictionary. Vstupní proměnné jsou název textového souboru, ve kterém je uložena DICOM Lookup a složka, ve které se soubor nachází.

Tabulka 9: Funkce pro separování a načítání dat

Načtení jednoho souboru	`[sep_data] = Reading_file (file_name, folder)`
Hromadné načtení	`[Files] = Reading_all (slozka_skript, slozka_data)`
Práce s tagy	`[tagArray] = tag_filter (Files, tag, value, outputTags)`
Načtení DICOM lookup	`[dicom_lookup] = dicomdict_read (file_name, folder)`

Po načtení DICOM Lookup získáme pole všech tagů. Pomocí funkce `tag_filter` tak můžeme načíst všechny studie se všemi tagy. Získáme tím matici hodnot, kde řádky jsou příslušná studie a sloupce příslušné tagy.

Ve studiích nejsou vyplněny všechny tagy, které Dicom Lookup nabízí. Proto sloupce, které představují tagy s nulovými hodnotami, můžeme smazat. Po této aplikaci získáme matici, viz obrázek 17, ve které jsou data přehledně načtena. V případě načtení našich 1215 studií jsme získali matici 1216x579. První řádek je číslo tagu a první sloupec číslo studie.



Obrázek 17. Matice hodnot jednotlivých tagů v rámci studií.

8.2 Četnosti vyplnění tagů

Naše matice má pole s 578 tagy, tedy 578 tagů bylo v našich 1215 studiích alespoň jednou vyplněno hodnotou. Jsou to tagy, se kterými budeme dále pracovat. Výpis všech 578 tagů je k nalezení v přiloženém CD v souboru `tabulky_statistika.xls` v záložce *Výpis tagy*.

Důležitým parametrem je četnost vyplnění jednotlivých tagů. Četnost se vypočítala jako počet tagů s nenulovým obsahem. S tagy, které nebývají častěji vyplněné, nemá smysl dále pracovat. Tagů, které měly relativní četnost vyplnění vyšší než 80 %, bylo pouze 49. V tabulce 10 vidíme nejzajímavější z nich. V příloze 7 pak můžeme nalézt výpis všech 49 tagů s příslušnými četnostmi vyplnění a v přiloženém CD v souboru `tabulky_statistika.xls` v záložce *Relativní a absolutní četnosti* nalezneme relativní a absolutní četnosti všech vyplněných tagů.

Tabulka 10. Nejzajímavější tagy, jejichž relativní četnost vyplnění je vyšší jak 80 %

Tag	Tag Name	AČ*	RČ** [%]
'0008,0060'	Modality	1215	100
'0008,0070'	Manufacturer	1212	99,75308642
'0008,0080'	Institution Name	1197	98,51851852
'0008,1010'	Station Name	1178	96,95473251
'0008,1030'	Study Description	1082	89,05349794
'0008,1090'	Manufacturer"s Model Name	1143	94,07407407
'0010,0030'	Patient"s Birth Date	1198	98,60082305
'0010,0040'	Patient"s Sex	1209	99,50617284
'0010,1010'	Patient"s Age	1058	87,0781893

* absolutní četnost
** relativní četnost

Tag 0008,0060 nám říká, na jakém typu modality obraz vzniknul. Hodnoty, které se v tomto tagu mohou nalézat, jsou standardizované a dá se s nimi velmi dobře pracovat a budou využity při další analýze.

Tag 0008,0070 nám specifikuje od jakého výrobce přístroj pochází. Tento tag vyplňuje sám přístroj a jak se tento tag bude vyplňovat se určí při instalaci přístroje.

Tag 0008,0080 nese informaci o zdravotnickém zařízení, ve kterém byl snímek pořízen.

Tag 0008,1030 vyplňují lékaři nebo jiní pracovníci a specifikuje vyšetření pacienta. Základní informace o pacientovi bývají vyplněny téměř vždy, z nichž zajímavé jsou tagy **0010,0030 / 0010,0040 / 0010,1010**, tedy pacientovo datum narození, pohlaví a věk.

Zajímavé jsou i další tagy, které mají ale relativní četnost vyplnění menší než 80 % a budou rozebrány níže.

8.3 Rozdělení studií dle typu modality

V této analýze bylo potřeba studie nejdříve rozdělit podle jednotlivých modalit. Tag *0008, 0060*, který je standardizovaný, určuje typ modality a tím i typ souboru, protože na základě toho se budou, respektive nebudou vyplňovat další tagy. Relativní četnost vyplnění tohoto tagu je 100 %, takže mohly být rozděleny všechny studie. Mimo samotné obrazové vyšetření mohou existovat i soubory, které později doplnily nějakou studii, například doktor později do studie něco dopsal. Jedná se například o Stuctured Report (SR) [7].

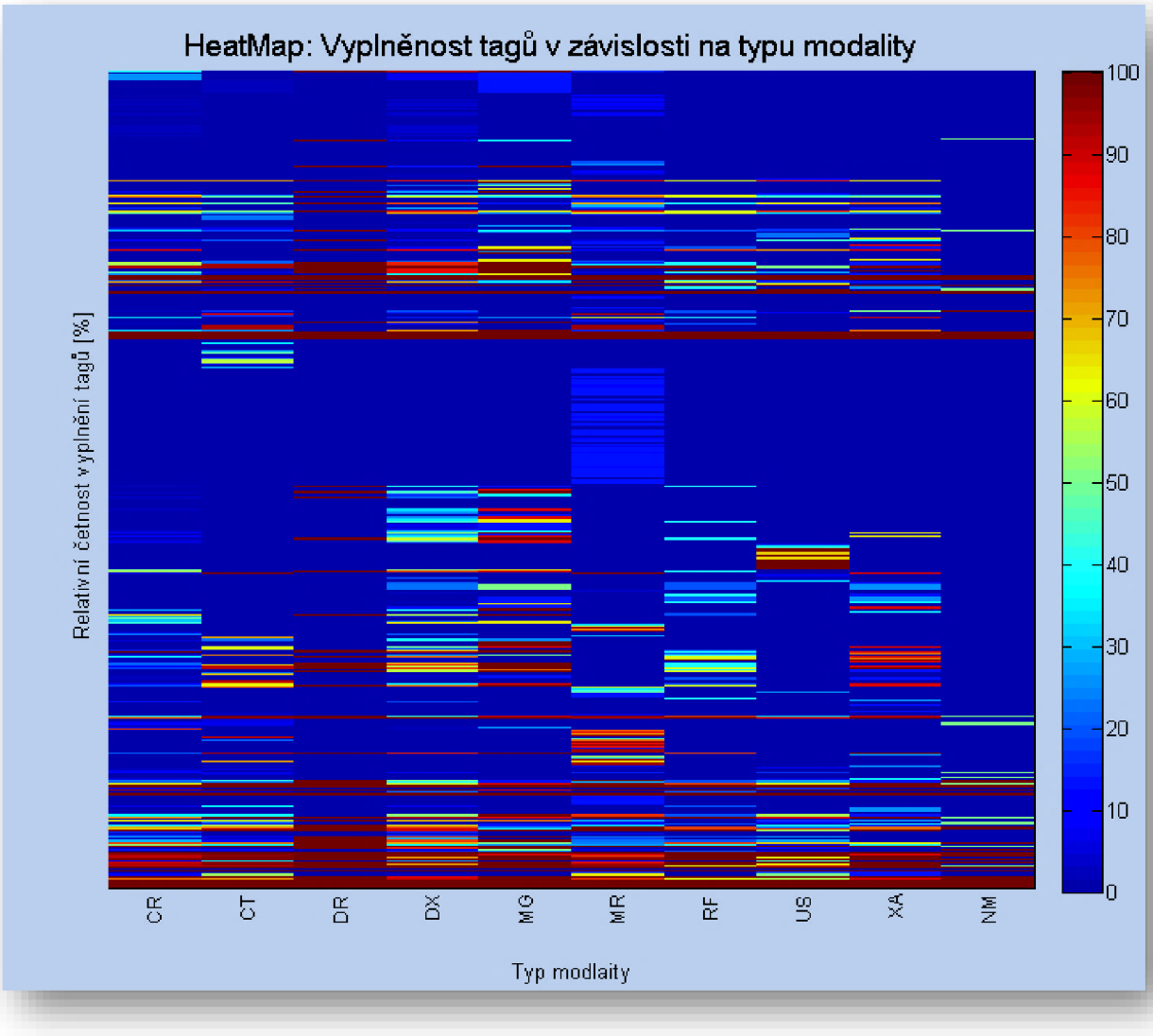
Tabulka 11. Přehled studií.

Přehled studií			
Zkratka	Typ vyšetření (angl.)	Typ vyšetření	Počet studií v datech
CR	Computed Radiography	Skiagrafie nepřímá digitalizace	206
CT	Computed Tomography	Počítačová tomografie	384
DR	Digital Radiography	Skiagrafie přímá digitalizace	2
DX	Digital Radiography	Skiagrafie přímá digitalizace	132
MG	Digital Mammography	Mamografie	8
MR	Magnetic Resonance	Magnetická rezonance	429
NM	Nuclear Medicine	Nukleární medicína	2
RF	X-Ray Radiofluroscopic	Fluoroskopie	5
US	Ultrasound	Ultrazvuk	18
XA	X-Ray Angiographic	Angiografie	8
KO	Key Object Selection		2
PR	Presentation State		4
SR	SR Document		13
Špatně vyplněný tag			2
Suma			1215

V další analýze se budeme zabývat pouze studiemi přímo z modalit. Studie KO, PR a SR proto byly odfiltrovány. U dvou studií, konkrétně Nemocnice Třinec, p.o., tento tag není vyplněný, přesněji řečeno, tag je vyplněný pouze stejným tagem.

Pro přehlednost byla vytvořena heat mapa, viz obrázek 18, která zobrazuje relativní četnosti vyplnění tagů v závislosti na typu modality. Na levé ose jsou jednotlivé tagy. Heat mapa je přiložená na CD a dá se zobrazit v prostředí MATLAB příkazem `view`. Po přiblížení určité oblasti zájmu se zobrazí i jednotlivé tagy.

Z obrázku si můžeme všimnout rozdílného vyplnění tagů u CR a DX. Tedy rentgenových vyšetření s přímou a nepřímou digitalizací. Na první pohled je patrné, že rentgeny s přímou digitalizací ukládají větší množství informací, ať už z hlediska počtu tagů nebo jejich četnosti vyplnění. Pokud si přiblížíme některé oblasti zájmů (viz příloha 9) abychom zjistili, které tagy více vyplňují přístroje přímé digitalizace, zjistíme, že se jedná převážně o technické parametry. To je logické, protože přístroje přímé digitalizace jsou novější a novější přístroje zapisují více technických parametrů.

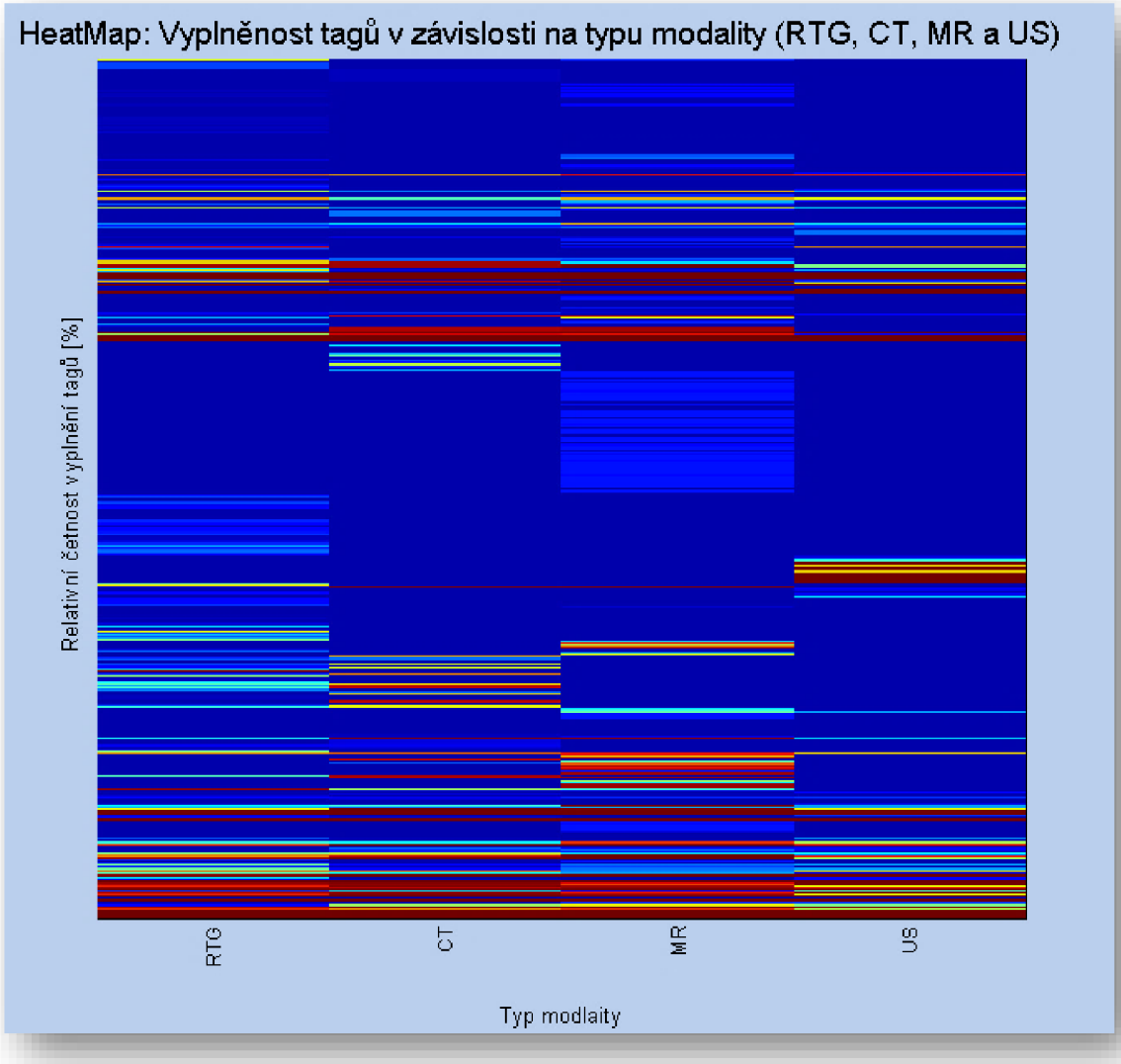


Obrázek 18. Heat mapa.

U modality DR vidíme, že vyplněné tagy jsou vyplněné 100%. To je způsobeno tím, že máme k dispozici pouze dvě studie. Stejně tak modalit RF a XA nemá analýza z našeho pohledu smysl. Proto modality CR, DR a DX byly dále sjednoceny do jednoho typu vyšetření RTG. Dále ponecháme modality CT, MR a US, kde máme k dispozici větší množství studií. Tuto redukci shrnuje tabulka 12.

Tabulka 12. Zredukované typy modalit a počet vzorků.

Typ modality	Počet vzorků
RTG	341
CT	384
MR	429
US	18



Obrázek 19. Heat mapa relativní četnost vyplnění tagů v závislosti na typu modality.

Z heat mapy na obrázku 19, která je taktéž přiložená na CD, vidíme, že u různých modalit jsou vyplněny různé tagy. Ve spodní části heat mapy jsou tagy u všech modalit vyplněny více. Tyto tagy patří převážně MetaInfu, které je vyplněno vždy, dále jsou to informace o zdravotnickém zařízení a o pacientovi.

8.4 Kvalita dat

Většina tagů nemá standardizované hodnoty, kterými se má vyplnit, jako tomu je například u typu modality. Proto vzniká značná nejednotnost ve vyplněných tazích.

Tag *0008,0080* – Institution Name

Tag *0008,0080*, ve kterém je uložen název zdravotnického zařízení, je vyplněn téměř vždy. Jednotnost ale chybí. V tabulce 13. můžeme vidět dva příklady různého vyplnění tohoto tagu, i když se jedná o stejnou instituci. Tento tag nastavuje servisní technik při instalaci. Ve Fakultní nemocnici u sv. Anny v Brně je velké množství přístrojů a každý tento tag vyplňuje různě.

Pokud bychom chtěli sjednotit do jedné množiny přístroje z jedné instituce, bylo by to značně obtížné. Při velkém vzorku dat by se dalo využít strojového učení nebo bychom museli vytvořit slovníček, který by ale nebylo jednoduché vytvořit. V prvním případě se nedá využít slovíčka *brno*, ve kterém se instituce nachází, dalo by se jedině využít slovo *anny* s úspěšností 90 %. V druhém případě by slovo *zlin*, tedy město, ve kterém se instituce nachází použít dalo a úspěšnost by byla 100 %.

Tabulka 13. Příklad různého vyplnění tagu Institution Name

'FN U Sv. Anny'	'KNTB ZLIN'
'FN U Sv.Anny Ortopedie'	'KNTB Zlin'
'FN U sv. Anny'	'KNTB Zlin a.s.'
'FN U sv. Anny v Brne'	'KNTB Zlin, OZM'
'FN USA ICRC'	'KNTB a.s. Zlin'
'FN u sv. Anny'	'KNTB, Zlin'
'FN u sv. Anny v Brne 77697907'	
'FNUSA'	
'Fakultni nemocnice U sv. Anny - Brno'	
'Fakultni nemocnice u sv. Anny v Brne'	

Tag *0008,0070* – Manofacturer

Tag *0008,0070* zaznamenává výrobce přístroje. Jak se tento tag vyplňuje se taktéž zadává při instalaci přístroje. Z tabulky 14 je patrné, že každý servisní technik zadá výrobce v trošku jiném formátu, ale základ zůstává stejný. Pokud bychom, tedy filtrovali výrobce Siemens přes slovíčko *siemens*, úspěšnost by byla 100 % a jinak tomu není ani u jiných výrobců.

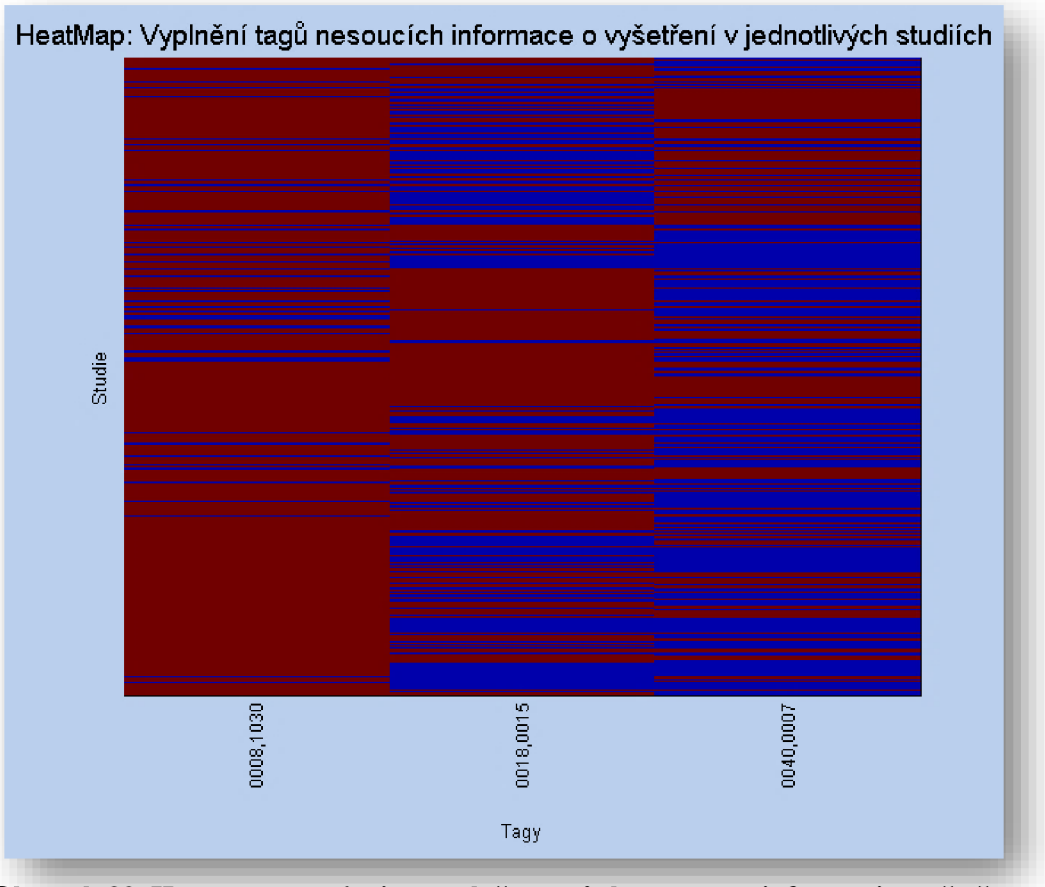
Tabulka 14. Příklad různého označení výrobců v tagu 0008,0070

'SIEMENS'	'Philips'
'SIEMENS NM'	'Philips Healthcare'
'Siemens'	'Philips Medical Systems'
'Siemens Healthcare'	

Tagy nesoucí informace o vyšetření

Tag *0008,1030* zaznamenává informace o vyšetření. Jedná se o data, která do počítače zanese vyšetřující lékař. Tato data mohou být cenná, jsou akorát velmi heterogenní. Každý lékař zaznamená vyšetření trochu jinak.

Další zajímavý tag nesoucí informace o vyšetření je tag *0018,0015*. Do něj se zaznamenává část těla, která byla zobrazována. Relativní četnost vyplněnosti tohoto tagu je 60,3 %, tedy více jak polovina studií tuto informaci obsahuje. Poslední tag je *0040,0007*, Scheduled Procedure Step Description, ve kterém se taktéž zaznamenává popis vyšetření nebo část těla.



Obrázek 20. Heat mapa popisující vyplnění tagů, které nesou informaci o vyšetření

Na obrázku 20 můžeme vidět heat mapu absolutních četností vyplněnosti všech tří tagů v rámci všech studií (červená - je vyplněn, modrá - není vyplněn). Z obrázku je vidět, že se vyplněnost v rámci studií překrývá. Když tedy není informace uložená v prvním tagu nebo je nečitelná, můžeme se podívat do zbylých dvou tagů. Tato heat mapa je taktéž v přiloženém CD.

Nejdůležitější je tag *0008,1030*, Study Description. V tomto tagu jsou data ale silně heterogenní. Pokud bychom chtěli vymyslet algoritmus, který by tato data dokázal přečíst, bylo by dobré do něj zakomponovat i tag *0018, 0015* (nesoucí informaci o části těla), který je daleko lépe čitelný. Tento tag však nebývá vyplněn tak často. Ukázka, jak jsou tyto tagy vyplněné je v tabulce 15. Vidíme, že se jedná o velmi heterogenní data.

Tabulka 15. Ukázka hodnot tagů, ve kterých se nachází informace o vyšetření

Soubor	**Tagy**		
	0008,1030'	'0018,0015'	'0040,0007'
'Studie1.txt'	'L PATER'	'TLSPINE'	[]
'Studie10.txt'	'Hrudnik'	'CHEST'	[]
'Studie100.txt'	'Head^01_HeadNeuro Adult'	'HEAD'	'ct oblicej.skelet'
'Studie1017.txt'	'C PATER'	'CSPINE'	[]
'Studie1018.txt'	'MOZEK'	'HEAD'	[]
'Studie102.txt'	'Thorax^01_ThoraxRoutine Adult'	'CHEST'	'CT plic nativ'
'Studie1103.txt'	'CT bricha'	[]	[]
'Studie1104.txt'	'Head^02_MOZEK_NEURO Adult'	'HEAD'	'CT mozku-nativ'
'Studie1140.txt'	'Thorax^1_Plice_mediastinum Adult'	'CHEST'	'CT plice nedias'
'Studie1141.txt'	'Head^1HeadRoutine Adult'	'HEAD'	'CT mozku'
'Studie1142.txt'	'Head^1HeadRoutine Adult'	'HEAD'	'CT mozku nativ,'

Expoziční informace

Do tagů se ukládá mnoho technických parametrů. Mohou to být například poziční souřadnice. Tyto informace pro nás nemají velký význam. Jediné parametry, které by nás mohly zajímat, jsou tagy, do kterých se ukládají data, ze kterých bychom mohli odhadnout dávku ozáření.

Každá modalita má vyčleněné jiné tagy, do kterých se tyto informace ukládají a taktéž nesou jinou informační hodnotu. U rentgenu je to tag *0018,1405*, Relative X-Ray Exposure. U CT pak například tagy *0018,1150*, Exposure Time a *0018,1151*, X-Ray Tube Current.

9 Diskuze

1215 studií je malý vzorek dat, abychom mohli tvrzení z této práce vztáhnout na celou Českou republiku.

Dá se nepřímo tvrdit, že starší přístroje zapisují méně technických informací než přístroje novější.

Vzhledem k tomu, že tag *Institution Name*, tedy název zdravotnického zařízení, ze kterého snímek pochází, bývá vyplněný téměř vždy, dalo by se s ním dále pracovat. Za předpokladu, že bychom zlepšili kvalitu těchto dat (třeba je sjednotili s Národním registrem poskytovatelů zdravotních služeb) nebo vymysleli algoritmus na základě filtrování slov nebo strojového učení, mohli bychom vytvořit mapu migrace dat mezi zdravotnickými zařízeními. Věděli bychom, z jakého archivu zdravotnického zařízení jsme data exportovali, a v datech bychom dohledali původ snímku. Z takovýchto dat bychom mohli vytvořit mapu odkud kam obrazová data putují. Z takovéto mapy bychom pak například viděli, z jakých okresních nemocnic posílají pacienty do větších nemocnic a na jaká vyšetření. Takováto data zatím nikde dohledatelná nejsou.

Tag *Manufacturer*, tedy výrobce, je vyplněný ve více než 99 % případů. Z tohoto tagu bychom mohli například zjistit, od jakého výrobce se přístroje, v kterých městech, krajích nebo celé ČR nejvíce využívají.

V hlavičkách souborů se v několika tazích nacházejí informace, i když ve velmi heterogenní podobě, o typu vyšetření. Pokud bychom vymysleli algoritmus, který by tato heterogenní data dokázal přečíst, dostali bychom informaci o typu vyšetření nebo informaci, která část těla byla zobrazována. Tím by se studie daly třídit podle typu vyšetření. Tuto informaci bychom pak mohli vztáhnout k dalším parametrům jako je typ modality nebo k některé z informací o pacientovi, jako je věk nebo pohlaví.

Zajímavé by mohlo být, spojit některé informace z hlavičky s diagnózou pacienta. PACS si většinou umí tato data vytáhnout z NIS. Poté se dají shlukovat podobné případy dle diagnóz, věku, lokality, určovat statistiku výskytů, zjišťovat, zda se u podobné diagnózy používaly stejné nebo jiné metody vyšetření apod.

10 Závěr

Diplomová práce se zabývá problematikou big dat ve zdravotnictví, konkrétně daty uloženými spolu s obrazy do DICOM souboru. Na 1215 studiích byla provedena analýza hlaviček DICOM souborů.

Ve 1215 studiích bylo alespoň jednou vyplněno 578 tagů. U 49 tagů byla relativní četnost vyplněnosti vyšší než 80 %. Velká část tagů se rozdílně vyplňuje u různých typů zobrazovacích technik, protože se jedná o informace spojené s technickými parametry přístroje.

Data, která jsou společná pro všechny typy zobrazovacích technik jsou převážně data o pacientovi a zdravotnickém zařízení, ve kterém byl snímek pořízen. Informace, ve kterém zdravotnickém zařízení byl snímek pořízen trpí velkou heterogenitou stejně tak jako informace o typu vyšetření. Obě tyto informace by mohli sloužit k dalším analýzám, popřípadě odkrýt další souvislosti v datech. Bude však potřeba vyřešit obtíže s jejich heterogenitou.

Práce mapuje vyplněnost tagů a byly navrženy další možnosti potupu, jak s těmito daty pracovat.

Zdroje

[1] WANG, Baoying, Ruowang LI a W PERRIZO. *Big data analytics in bioinformatics and healthcare.* ISBN 9781466666146

[2] PIANYKH, Oleg S. *Digital imaging and communications in medicine (DICOM): a practical introduction and survival guide*. Berlin: Springer, 2008. ISBN 9783540745709

[3] BELLE, Ashwin, Raghuram THIAGARAJAN, S. M. Reza SOROUSHMEHR, Fatemeh NAVIDI, Daniel A. BEARD a Kayvan NAJARIAN. Big Data Analytics in Healthcare. *BioMed Research International* [online]. 2015, **2015**, 1-16 [cit. 2017-05-16]. DOI: 10.1155/2015/370194. ISSN 2314-6133. Dostupné z: http://www.hindawi.com/journals/bmri/2015/370194/

[4] MUSTRA, M., K. DELAC a M. GRGIC, 2008. *Overview of the DICOM standard*. roč. 1. 50th International Symposium ELMAR, 2008. ISSN 1334-2630.

[5] CHEE, Adam. *Advances in Medical Imaging Informatics – Dealig with Big Data* [online]. Ambis 2012: Biomedical Symposium, Singapore, 25th May 2012. Dostupný z https://binaryhealthcare.files.wordpress.com/2009/03/advances-in-mii- big-data.pdf

[6] LANEY, Doug. *3D Data Management: Controlling Data Volume, Velocity, and Variety* [online]. roč. 949. META Delta. Application Delivery Strategies, 2001. ISSN 09505849. Dostupné z: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity- and-Variety.pdf

[7] National Electrical Manufacturers Association. *Digital Imaging and Communications in Medicine* [online]. Rosslyn. Virginia, 2011. Dostupný z: http://dicom.nema.org/

[8] K. Shvachko, H. Kuang, S. Radia and R. Chansler. *The Hadoop Distributed File System*. 2010. IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, 2010.

[9] Honc Josef. *Veritas Information Governance – získejte zpět kontrolu nad vašimi daty* [online]. Konference BigData Praha 2016, 2016. Dostupné z: https://eventworld.cz/akce/bigdata-2016-92/archiv-prezentaci-bigdata-2016

[10] Ústav zdravotnických informací a statistiky ČR. *Činnost zdravotnických zářízení ve vybraných oborech* [online]. 2015. ISSN: 1211-2585. Dostupný z: http://www.uzis.cz/katalog/zdravotnicka-statistika/cinnost-zdravotnickych-zarizeni-ve-vybranych-oborech

[11] MÍČ, Vladimír, Filip NÁLEPA. *DISA: Data Intensive Systém and Application: Prezentace laboratořepro den otevřených dveří na FI MU*. [online]. 2017. Dostupný z: http://disa.fi.muni.cz/

[12] BERKA, Petr. *Dobývání znalostí z databází*. Praha: Academia, 2003. ISBN 80-200-1062-9.

[13] GIBAUD, Bernard. *The DICOM Standard: A Brief Overview* [online]. s. 229 [cit. 2017-05-11]. DOI: 10.1007/978-1-4020-8752-3_13. Dostupné z: http://link.springer.com/10.1007/978-1-4020-8752-3_13

[14] Ústav zdravotnických informací a statistiky ČR. *Činnost zdravotnických zařízení ve vybraných oborech* [online]. 2015. ISSN: 1803-3881. Dostupný z: http://www.uzis.cz/katalog/zdravotnicka-statistika/cinnost-spolecnych-vysetrovacich-lecebnych-slozek

[15] DINOV, Ivo D. *Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data*. GigaScience [online]. 2016, **5**(1), - [cit. 2017-05-15]. DOI: 10.1186/s13742-016-0117-6. Dostupné z: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-016-0117-6

[16] KHAN, Nawsher, Ibrar YAQOOB, Ibrahim Abaker Targio HASHEM, Zakira INAYAT, Waleed Kamaleldin MAHMOUD ALI, Muhammad ALAM, Muhammad SHIRAZ a Abdullah GANI. Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal* [online]. 2014, **2014**, 1-18 [cit. 2017-05-16]. DOI: 10.1155/2014/712826. ISSN 2356-6140. Dostupné z: http://www.hindawi.com/journals/tswj/2014/712826/

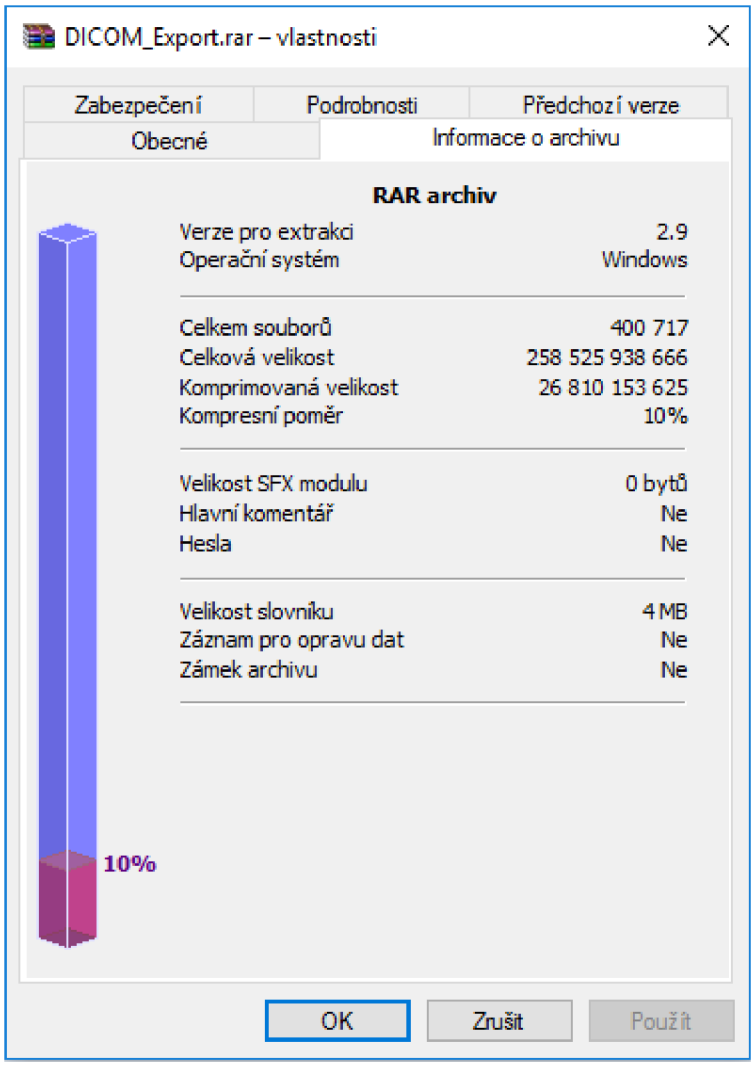
[17] KÄLLMAN, Hans-Erik, Erik HALSIUS, Magnus OLSSON a Mats STENSTRÖM. DICOM Metadata repository for technical information in digital medical images. *Acta Oncologica* [online]. 2009, **48**(2), 285-288 [cit. 2017-05-18]. DOI: 10.1080/02841860802258786. ISSN 0284-186x. Dostupné z: http://www.tandfonline.com/doi/full/10.1080/02841860802258786

[18]

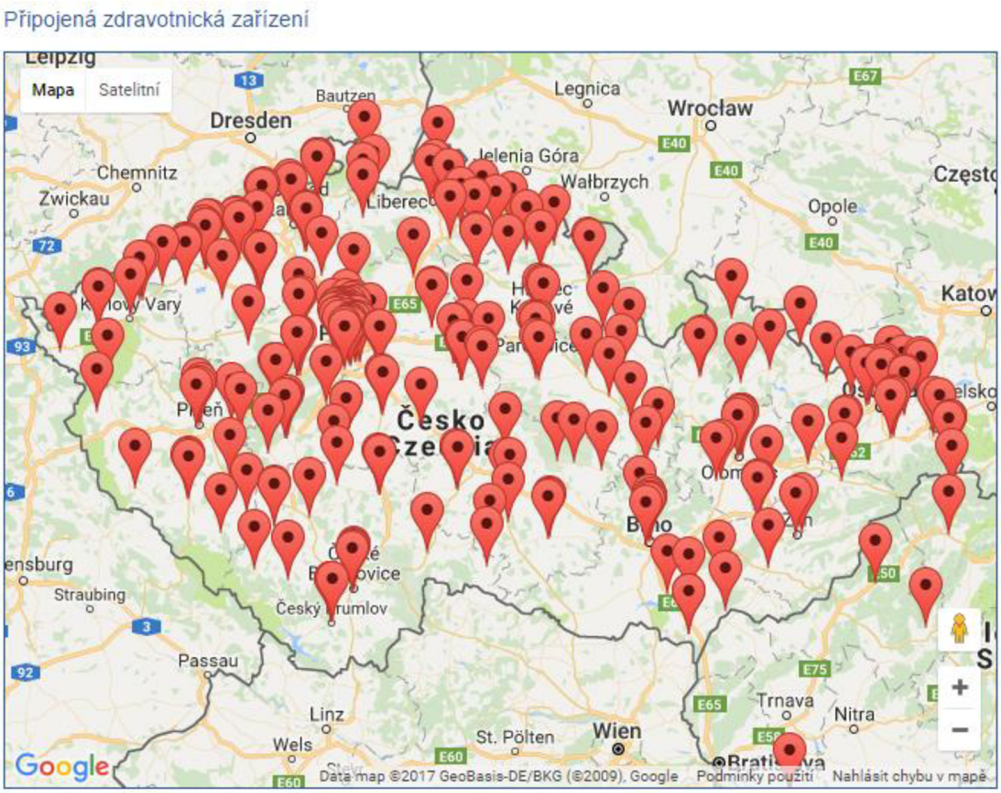
Příloha 1: Ukázka ze seznamu datových prvků v DICOM.

Tag	Name	Keyword	VR	VM
(0008,0001)	Length to End	LengthToEnd	UL	1
(0008,0005)	Specific Character Set	SpecificCharacterSet	CS	1-n
(0008,0008)	Image Type	ImageType	CS	2-n
(0008,0010)	Recognition Code	RecognitionCode	SH	1
(0008,0012)	Instance Creation Date	InstanceCreationDate	DA	1
(0008,0013)	Instance Creation Time	InstanceCreationTime	TM	1
(0008,0014)	Instance Creator UID	InstanceCreatorUID	UI	1
(0008,0016)	SOP Class UID	SOPClassUID	UI	1
(0008,0018)	SOP Instance UID	SOPInstanceUID	UI	1
(0008,0020)	Study Date	StudyDate	DA	1
(0008,0021)	Series Date	SeriesDate	DA	1
(0008,0022)	Acquisition Date	AcquisitionDate	DA	1
(0008,0023)	Content Date	ContentDate	DA	1
(0008,0024)	Overlay Date	OverlayDate	DA	1
(0008,0025)	Curve Date	CurveDate	DA	1
(0008,002A)	Acquisition DateTime	AcquisitionDateTime	DT	1

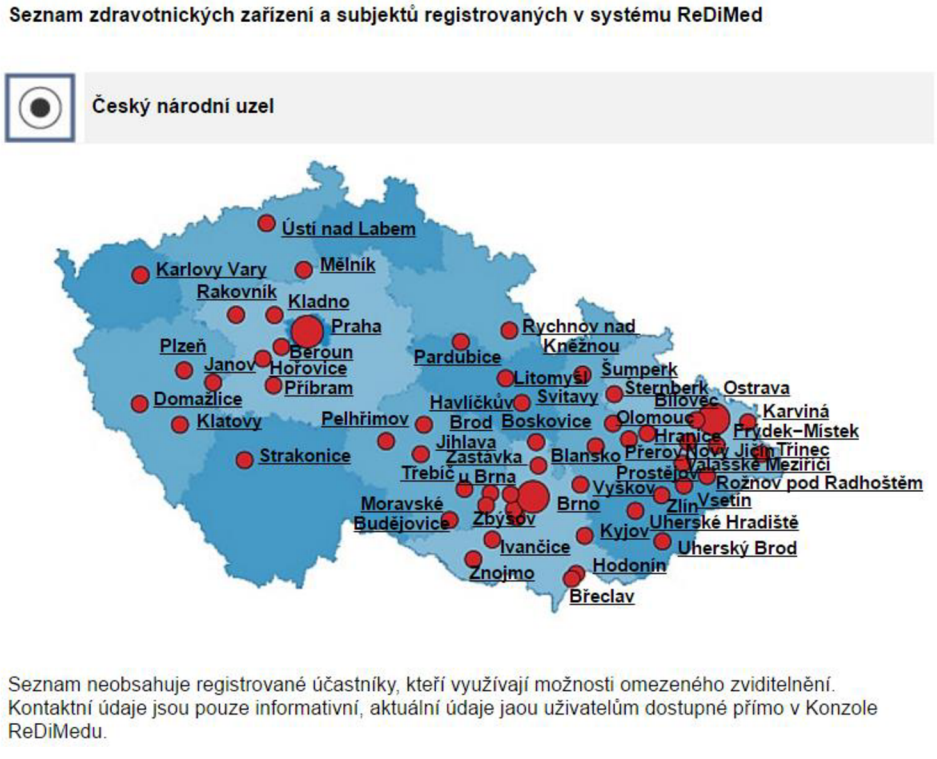
Příloha 2: Vlastnosti komprimovaných dat.



Příloha 3: Mapy zařazených nemocnic v projektech ePacs a ReDiMed.



Obrázek 8. Připojená zdravotnická zařízení v systému ePacs.



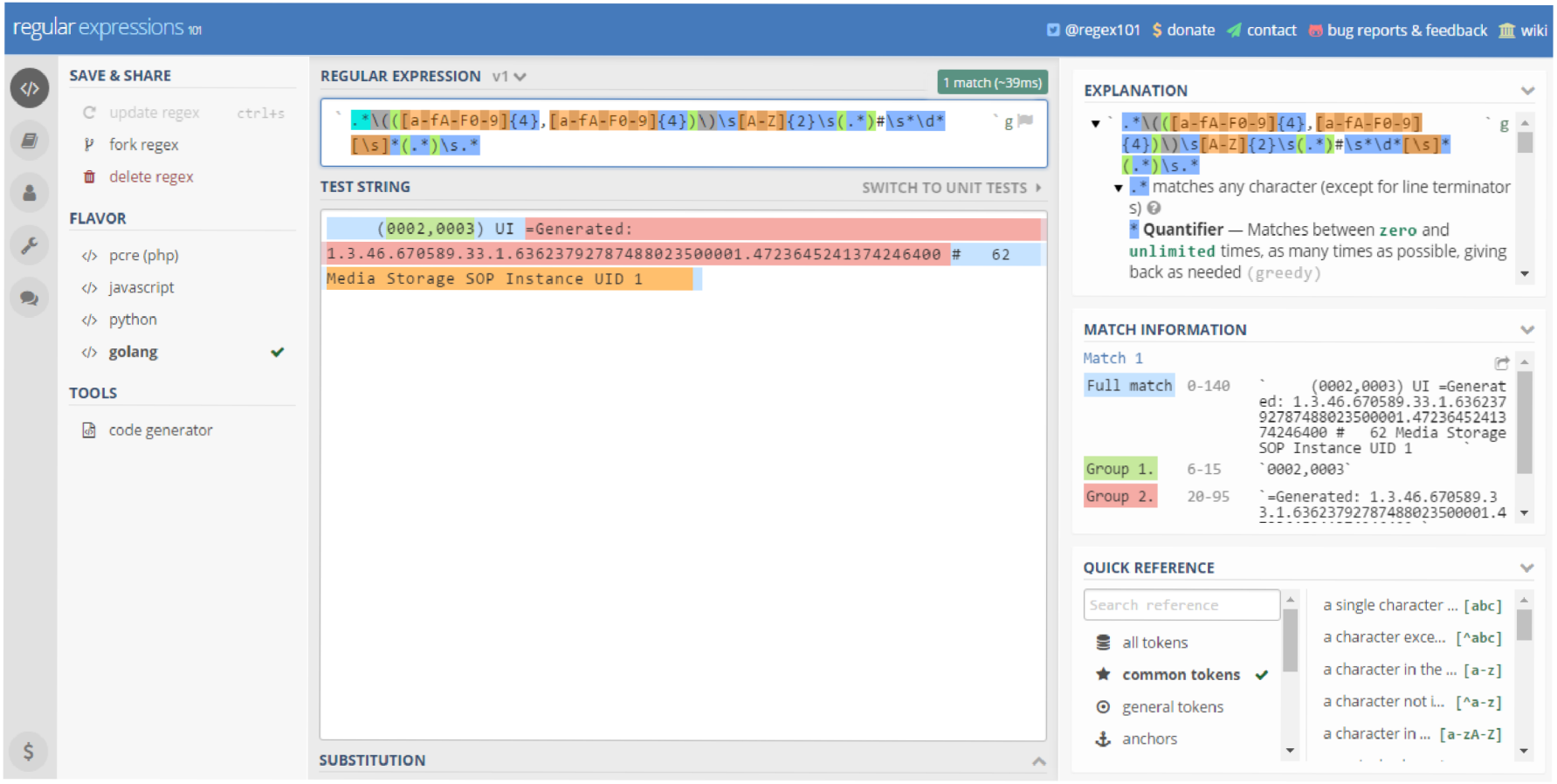
Obrázek 7. Připojená zdravotnická zařízení v systému ReDiMed v ČR

Příloha 4: Nejčastější výkony na odděleních nukleární medicíny.

6.5 Nejčastější výkony (in vivo) na odděleních nukleární medicíny*)

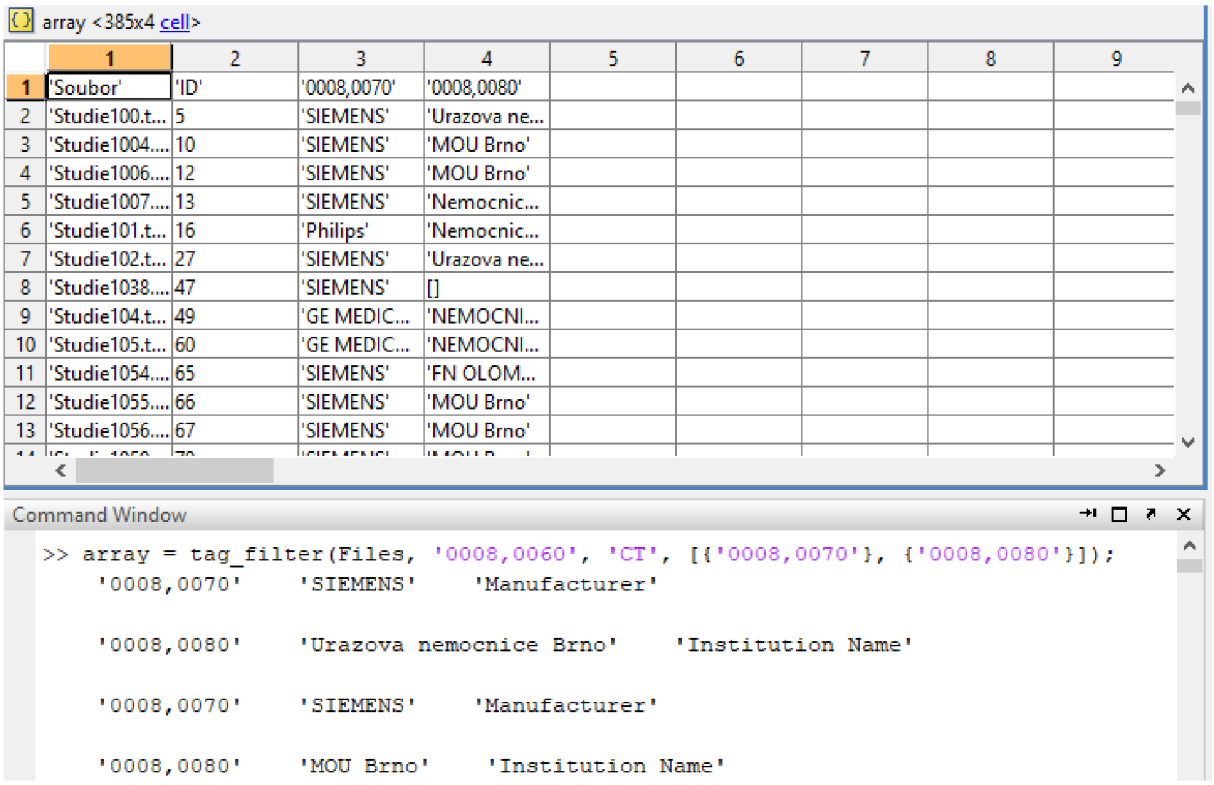
Číslo výkonu	Název výkonu	Počet výkonů (in vivo)			
		celkem		z toho pro (v %)	
		absolutně	podíl (v %)	ústavní péči	ostatní zařízení
47269	Tomografická scintigrafie - SPECT	85 010	20,1	20,8	79,2
47273	Kvantifikace dynamických a tomografických scintigrafických vyšetření	61 528	14,6	20,6	79,4
47241	Scintigrafie skeletu	56 033	13,3	16,2	83,8
47271	Kvantifikace výsledku statického scintigrafického vyšetření	43 804	10,4	29,3	70,7
47257	Scintigrafie plic perfúzní	28 064	6,6	42,8	57,2
47302	Hybridní výpočetní a pozitronová emisní tomografie	15 847	3,7	5,6	94,4
47259	Scintigrafie plic ventilační statická	10 508	2,5	46,0	54,0
47245	Scintigrafie skeletu cílená třífázová	10 105	2,4	15,5	84,5
47351	PET trupu	7 106	1,7	5,8	94,2
47219	Scintigrafie ledvin dynamická včetně stanovení GF resp. ERPF	5 729	1,4	20,7	79,3
47217	Scintigrafie ledvin dynamická	5 226	1,2	26,6	73,4
47147	Scintigrafie štítné žlázy prostá	5 146	1,2	48,2	51,8
47275	Scintigrafie sentinelové uzliny	4 383	1,0	75,9	24,1
47263	Radionuklidová lymfografie	3 919	0,9	19,7	80,3
47139	Radionuklidová flebografie	3 096	0,7	20,2	79,8
47022	Cílené vyšetření odborníkem v nukleární medicíně	2 845	0,7	63,1	36,9
47215	Scintigrafie ledvin s výpočtem relativní funkce	2 581	0,6	10,7	89,3
47255	Tomografická scintigrafie perfúse mozku po podání difúsibilních RAF	2 358	0,6	26,0	74,0
47267	Scintigrafie nádoru	2 186	0,5	23,8	76,2
47151	Celotělová scintigrafie u karcinomu štítné žlázy	1 931	0,5	38,6	61,4
	Ostatní	65 353	15,5	22,2	77,8
Výkony celkem		**422 758**	**100,0**	**23,7**	**76,3**

*) Pouze výkony odbornosti 407 Nukleární medicína



Obrázek 9. Ukázka testování regulérního výrazu pro separaci dat.

Příloha 6. Ukázka funkce `tag_filter` a jejího výstupu.



array <385x4 cell>

	1	2	3	4	5	6	7	8	9
1	'Soubor'	'ID'	'0008,0070'	'0008,0080'					
2	'Studie100.t...	5	'SIEMENS'	'Urazova ne...					
3	'Studie1004....	10	'SIEMENS'	'MOU Brno'					
4	'Studie1006....	12	'SIEMENS'	'MOU Brno'					
5	'Studie1007....	13	'SIEMENS'	'Nemocnic...					
6	'Studie101.t...	16	'Philips'	'Nemocnic...					
7	'Studie102.t...	27	'SIEMENS'	'Urazova ne...					
8	'Studie1038....	47	'SIEMENS'	[]					
9	'Studie104.t...	49	'GE MEDIC...	'NEMOCNI...					
10	'Studie105.t...	60	'GE MEDIC...	'NEMOCNI...					
11	'Studie1054....	65	'SIEMENS'	'FN OLOM...					
12	'Studie1055....	66	'SIEMENS'	'MOU Brno'					
13	'Studie1056....	67	'SIEMENS'	'MOU Brno'					

Command Window

```
>> array = tag_filter(Files, '0008,0060', 'CT', [{'0008,0070'}, {'0008,0080'}]);
    '0008,0070'    'SIEMENS'    'Manufacturer'

    '0008,0080'    'Urazova nemocnice Brno'    'Institution Name'

    '0008,0070'    'SIEMENS'    'Manufacturer'

    '0008,0080'    'MOU Brno'    'Institution Name'
```

Příloha 7. Četnosti vyplnění tagů, kde je relativní četnost více než 80 %.

Tag	Tag Name	Absolutní četnost vyplněnosti tagu	Relativní četnost vyplněnosti tagu
'0002,0001'	File Meta Information Version	1215	100
'0002,0002'	Media Storage SOP Class UID	1215	100
'0002,0003'	Media Storage SOP Instance UID	1215	100
'0002,0010'	Transfer Syntax UID	1215	100
'0002,0012'	Implementation Class UID	1215	100
'0002,0013'	Implementation Version Name	1215	100
'0008,0008'	Image Type	1177	96,87242798
'0008,0016'	SOP Class UID	1215	100
'0008,0018'	SOP Instance UID	1215	100
'0008,0020'	Study Date	1215	100
'0008,0021'	Series Date	1123	92,42798354
'0008,0022'	Acquisition Date	1103	90,781893
'0008,0023'	Content Date	1108	91,19341564
'0008,0030'	Study Time	1214	99,91769547
'0008,0031'	Series Time	1121	92,26337449
'0008,0032'	Acquisition Time	1101	90,61728395
'0008,0033'	Content Time	1089	89,62962963
'0008,0050'	Accession Number	1138	93,66255144
'0008,0060'	Modality	1215	100
'0008,0070'	Manufacturer	1212	99,75308642
'0008,0080'	Institution Name	1197	98,51851852
'0008,1010'	Station Name	1178	96,95473251
'0008,1030'	Study Description	1082	89,05349794
'0008,103e'	SeriesDescription	997	82,05761317
'0008,1090'	Manufacturer"s Model Name	1143	94,07407407
'0010,0010'	Patient"s Name	1215	100
'0010,0020'	Patient ID	1215	100
'0010,0030'	Patient"s Birth Date	1198	98,60082305
'0010,0040'	Patient"s Sex	1209	99,50617284
'0010,1000'	Other Patient IDs	1215	100
'0010,1001'	Other Patient Names	1215	100
'0010,1010'	Patient"s Age	1058	87,0781893
'0018,1020'	Software Version(s)	1131	93,08641975
'0020,000d'	StudyInstanceUID	1215	100
'0020,000e'	SeriesInstanceUID	1215	100
'0020,0010'	Study ID	1215	100
'0020,0011'	Series Number	1214	99,91769547
'0020,0013'	Instance Number	1214	99,91769547
'0028,0002'	Samples per Pixel	1196	98,43621399
'0028,0004'	Photometric Interpretation	1196	98,43621399
'0028,0010'	Rows	1196	98,43621399
'0028,0011'	Columns	1196	98,43621399
'0028,0030'	Pixel Spacing	1030	84,77366255
'0028,0100'	Bits Allocated	1196	98,43621399
'0028,0101'	Bits Stored	1196	98,43621399
'0028,0102'	High Bit	1196	98,43621399
'0028,0103'	Pixel Representation	1196	98,43621399
'0028,1050'	Window Center	1144	94,1563786
'0028,1051'	Window Width	1144	94,1563786

Četnosti všech tagů naleznete v CD příloze v souboru tabulky_statistika.xls.

Příloha 8. Výpis všech alespoň jednou vyplněných tagů v našich datech.

'0002,0001'	'0018,0090'	'0018,1700'	'0018,9058'	'0028,0010'	'0070,0023'
'0002,0002'	'0018,0091'	'0018,1702'	'0018,9059'	'0028,0011'	'0070,0024'
'0002,0003'	'0018,0093'	'0018,1704'	'0018,9060'	'0028,0014'	'0070,0041'
'0002,0010'	'0018,0094'	'0018,1706'	'0018,9062'	'0028,0030'	'0070,0042'
'0002,0012'	'0018,0095'	'0018,1708'	'0018,9064'	'0028,0034'	'0070,0052'
'0002,0013'	'0018,1000'	'0018,1720'	'0018,9069'	'0028,0100'	'0070,0053'
'0008,0005'	'0018,1004'	'0018,5010'	'0018,9073'	'0028,0101'	'0070,0062'
'0008,0008'	'0018,1008'	'0018,5012'	'0018,9074'	'0028,0102'	'0070,0066'
'0008,0012'	'0018,1010'	'0018,5020'	'0018,9075'	'0028,0103'	'0070,0067'
'0008,0013'	'0018,1012'	'0018,5021'	'0018,9077'	'0028,0300'	'0070,0080'
'0008,0014'	'0018,1014'	'0018,5022'	'0018,9078'	'0028,0301'	'0070,0081'
'0008,0016'	'0018,1016'	'0018,5050'	'0018,9079'	'0028,1040'	'0070,0082'
'0008,0018'	'0018,1018'	'0018,5100'	'0018,9080'	'0028,1041'	'0070,0083'
'0008,0020'	'0018,1019'	'0018,5101'	'0018,9081'	'0028,1050'	'0070,0084'
'0008,0021'	'0018,1020'	'0018,6000'	'0018,9082'	'0028,1051'	'0070,0100'
'0008,0022'	'0018,1030'	'0018,6012'	'0018,9087'	'0028,1052'	'0070,0101'
'0008,0023'	'0018,1040'	'0018,6014'	'0018,9089'	'0028,1053'	'0070,0102'
'0008,0030'	'0018,1041'	'0018,6016'	'0018,9090'	'0028,1054'	'0070,0401'
'0008,0031'	'0018,1042'	'0018,6018'	'0018,9091'	'0028,1055'	'0070,0403'
'0008,0032'	'0018,1043'	'0018,6020'	'0018,9093'	'0028,1056'	'0088,0140'
'0008,0033'	'0018,1044'	'0018,6022'	'0018,9094'	'0028,1201'	'0400,0005'
'0008,0050'	'0018,1046'	'0018,6024'	'0018,9098'	'0028,1202'	'0400,0010'
'0008,0060'	'0018,1047'	'0018,6026'	'0018,9100'	'0028,1203'	'0400,0015'
'0008,0061'	'0018,1048'	'0018,6028'	'0018,9101'	'0028,1300'	'0400,0020'
'0008,0064'	'0018,1049'	'0018,6030'	'0018,9147'	'0028,1350'	'0400,0100'
'0008,0068'	'0018,1050'	'0018,6031'	'0018,9151'	'0028,2110'	'0400,0105'
'0008,0070'	'0018,1063'	'0018,6032'	'0018,9155'	'0028,3002'	'0400,0110'
'0008,0080'	'0018,1065'	'0018,6036'	'0018,9168'	'0028,3003'	'0400,0115'
'0008,0081'	'0018,1081'	'0018,7000'	'0018,9170'	'0028,6010'	'0400,0120'
'0008,0090'	'0018,1082'	'0018,7001'	'0018,9171'	'0028,9001'	'2010,0010'
'0008,0100'	'0018,1083'	'0018,7004'	'0018,9172'	'0028,9002'	'2010,0030'
'0008,0102'	'0018,1084'	'0018,7005'	'0018,9174'	'0032,0012'	'2010,0040'
'0008,0103'	'0018,1088'	'0018,7006'	'0018,9177'	'0032,1000'	'2010,0100'
'0008,0104'	'0018,1090'	'0018,7008'	'0018,9178'	'0032,1001'	'2010,0140'
'0008,0105'	'0018,1094'	'0018,7010'	'0018,9179'	'0032,1021'	'2020,0010'
'0008,0106'	'0018,1100'	'0018,7011'	'0018,9180'	'0032,1030'	'2050,0020'
'0008,1010'	'0018,1110'	'0018,7012'	'0018,9181'	'0032,1032'	
'0008,1030'	'0018,1111'	'0018,7014'	'0018,9182'	'0032,1033'	
'0008,1040'	'0018,1114'	'0018,7016'	'0018,9183'	'0032,1060'	
'0008,1048'	'0018,1120'	'0018,7020'	'0018,9199'	'0032,4000'	
'0008,1050'	'0018,1130'	'0018,7022'	'0018,9218'	'0038,0010'	
'0008,1060'	'0018,1134'	'0018,7024'	'0018,9220'	'0038,0300'	
'0008,1070'	'0018,1138'	'0018,7026'	'0018,9231'	'0038,0500'	
'0008,1080'	'0018,1140'	'0018,7028'	'0018,9232'	'0040,0001'	
'0008,1090'	'0018,1141'	'0018,7030'	'0018,9240'	'0040,0002'	
'0008,1150'	'0018,1143'	'0018,7032'	'0018,9241'	'0040,0003'	
'0008,1155'	'0018,1147'	'0018,7034'	'0018,9302'	'0040,0004'	
'0008,1160'	'0018,1149'	'0018,7040'	'0018,9303'	'0040,0005'	
'0008,2111'	'0018,1150'	'0018,7041'	'0018,9305'	'0040,0006'	
'0008,2142'	'0018,1151'	'0018,7042'	'0018,9306'	'0040,0007'	
'0008,2143'	'0018,1152'	'0018,7044'	'0018,9307'	'0040,0009'	
'0008,2144'	'0018,1153'	'0018,7046'	'0018,9309'	'0040,0010'	
'0008,3010'	'0018,1154'	'0018,7048'	'0018,9310'	'0040,0241'	
'0008,9007'	'0018,1155'	'0018,7050'	'0018,9311'	'0040,0242'	
'0008,9123'	'0018,1156'	'0018,7052'	'0018,9313'	'0040,0243'	
'0008,9205'	'0018,1160'	'0018,7054'	'0018,9318'	'0040,0244'	
'0008,9206'	'0018,1162'	'0018,7060'	'0018,9323'	'0040,0245'	

'0008,9207'	'0018,1164'	'0018,7062'	'0018,9324'	'0040,0250'
'0008,9208'	'0018,1166'	'0018,7064'	'0018,9327'	'0040,0251'
'0008,9209'	'0018,1170'	'0018,7065'	'0018,9328'	'0040,0252'
'0010,0010'	'0018,1180'	'0018,8150'	'0018,9330'	'0040,0253'
'0010,0020'	'0018,1190'	'0018,8151'	'0018,9332'	'0040,0254'
'0010,0021'	'0018,1191'	'0018,9004'	'0018,9334'	'0040,0280'
'0010,0030'	'0018,1200'	'0018,9005'	'0018,9345'	'0040,0301'
'0010,0032'	'0018,1201'	'0018,9008'	'0020,0010'	'0040,0302'
'0010,0040'	'0018,1210'	'0018,9009'	'0020,0011'	'0040,0303'
'0010,1000'	'0018,1250'	'0018,9010'	'0020,0012'	'0040,0306'
'0010,1001'	'0018,1251'	'0018,9011'	'0020,0013'	'0040,0310'
'0010,1010'	'0018,1260'	'0018,9012'	'0020,0020'	'0040,0314'
'0010,1020'	'0018,1261'	'0018,9014'	'0020,0032'	'0040,0316'
'0010,1030'	'0018,1310'	'0018,9015'	'0020,0037'	'0040,0318'
'0010,1040'	'0018,1312'	'0018,9016'	'0020,0052'	'0040,1001'
'0010,2000'	'0018,1314'	'0018,9017'	'0020,0060'	'0040,1003'
'0010,2110'	'0018,1315'	'0018,9018'	'0020,0062'	'0040,1007'
'0010,2160'	'0018,1316'	'0018,9019'	'0020,0100'	'0040,1010'
'0010,2203'	'0018,1318'	'0018,9020'	'0020,0105'	'0040,1400'
'0010,4000'	'0018,1400'	'0018,9021'	'0020,0110'	'0040,2004'
'0012,0062'	'0018,1401'	'0018,9022'	'0020,1002'	'0040,2005'
'0018,0010'	'0018,1402'	'0018,9024'	'0020,1040'	'0040,2016'
'0018,0015'	'0018,1403'	'0018,9025'	'0020,1041'	'0040,2017'
'0018,0020'	'0018,1404'	'0018,9026'	'0020,1208'	'0040,2400'
'0018,0021'	'0018,1405'	'0018,9027'	'0020,4000'	'0040,8302'
'0018,0022'	'0018,1450'	'0018,9028'	'0020,9056'	'0040,9210'
'0018,0023'	'0018,1460'	'0018,9029'	'0020,9057'	'0040,9224'
'0018,0024'	'0018,1500'	'0018,9030'	'0020,9072'	'0040,9225'
'0018,0025'	'0018,1508'	'0018,9032'	'0020,9128'	'0050,0004'
'0018,0040'	'0018,1510'	'0018,9033'	'0020,9157'	'0054,0400'
'0018,0050'	'0018,1511'	'0018,9034'	'0020,9164'	'0054,1001'
'0018,0060'	'0018,1530'	'0018,9035'	'0020,9165'	'0070,0002'
'0018,0080'	'0018,1531'	'0018,9036'	'0020,9167'	'0070,0004'
'0018,0081'	'0018,1600'	'0018,9037'	'0020,9254'	'0070,0005'
'0018,0082'	'0018,1602'	'0018,9041'	'0020,9255'	'0070,0006'
'0018,0083'	'0018,1604'	'0018,9043'	'0020,9256'	'0070,0010'
'0018,0084'	'0018,1606'	'0018,9044'	'0020,9421'	'0070,0011'
'0018,0085'	'0018,1608'	'0018,9047'	'0028,0002'	'0070,0014'
'0018,0086'	'0018,1610'	'0018,9048'	'0028,0004'	'0070,0015'
'0018,0087'	'0018,1612'	'0018,9050'	'0028,0006'	'0070,0020'
'0018,0088'	'0018,1620'	'0018,9051'	'0028,0008'	'0070,0021'
'0018,0089'	'0018,1622'	'0018,9053'	'0028,0009'	'0070,0022'

Příloha 9. Rozdíly v relativních četnostech vyplnění tagů mezi rentgeny s přímou a nepřímou digitalizací.