Portál pre odborné publikovanie ISSN 1338-0087

Objavovanie znalostí v medicínskych dátach

EMEL a kol. · Informačné technológie, Medicína

05.11.2014

Článok sa zaoberá problematikou objavovania nových znalostí v neštruktúrovaných medicínskych dátach. Popisuje základné kroky všeobecného procesu objavovania znalostí – Knowledge discovery a taktiež sa zaoberá konkrétnejšie problematikou objavovania znalostí v textových dátach. Ďalej stručne popisuje problematiku openEHR, ktorý bol vybraný ako štandard pre uchovávanie a transformáciu medicínskych dát.

Úvod

V jednotlivých oblastiach ľudskej činnosti vznikajú každý deň kvantá dát a informácií. Poskytnúť exaktné číslo nie je možné, vieme však uviesť, že tempo prírastku dát každý deň sa počíta v stovkách petabajtoch. Jedným z príkladov kde denne vznikajú obrovské kvantá doslova životne dôležitých dát, je oblasť poskytovania zdravotnej starostlivosti. Je veľký predpoklad, že poskytovatelia zdravotnej starostlivosti disponujú množstvom analógových dokumentov nachádzajúcich sa v archívoch a taktiež elektronickými dátami uloženými v databázach. Tieto databázy sú však medzi rôznymi poskytovateľmi nekompatibilné, častokrát sú lekárske záznamy, aj keď v elektronickej, ale iba textovej neštruktúrovanej podobe. Tým pádom z veľkého množstva dát je veľmi obtiažne vyvodiť nejaké všeobecne platné vedomosti, či už čo sa týka pacienta ako jednotlivca, alebo väčšej skupiny pacientov.

Taktiež nie je zabezpečená interoperabilita dát o pacientovi medzi viacerými poskytovateľmi zdravotnej starostlivosti. Tým pádom dáta a informácie nie je možné elektronicky zdieľať a z pacienta sa tak stáva “poštár”, ktorý svoju zdravotnú dokumentáciu alebo jej časti prenáša medzi lekármi. Z uvedeného vyplýva, že pri prechode na systém elektornického zdravotníctva eHealth bude potrebné existujúce lekárske záznamy o pacientoch efektívnym spôsobom transfromovať do štruktúrovanej podoby a následne zabezpečiť ich správnu interpretáciu a interoperabilitu medzi poskytovateľmi zdravotnej stratoslivosti. V tomto článku sa chceme venovať práve problematike transformácie neštruktúrovaných elektronických lekárskych záznamov do štruktúrovanej podoby.

Knowledge Discovery a typy dát

Proces štruktúrovania neštruktúrovaných vysokodimenzionálnych dát je netriviálny proces. V súčasnosti existuje vedná disciplína zaoberajúca sa skúmaním vzťahov medzi dátami, ich štruktúrovaním a reprezentáciou pod názvom Knowledge Discovery. Práve

princípy tohto vedného odboru sme aplikovali v procese štruktúrovania lekárskych záznamov. Objavovanie znalostí je poloautomatický proces extrakcie a selekcie znalostí z rôznorodých dát (ako napr. extrakcia dát z textov, databáz, webu a pod.). Najdôležitejšie pri tomto procese je, aby znalosti, ktoré týmto procesom extrakcie získame, boli:

1. platné
2. efektívne využiteľné (pre daný softvér, ďalšie použitie, výskum)
3. nové (doteraz neznáme)

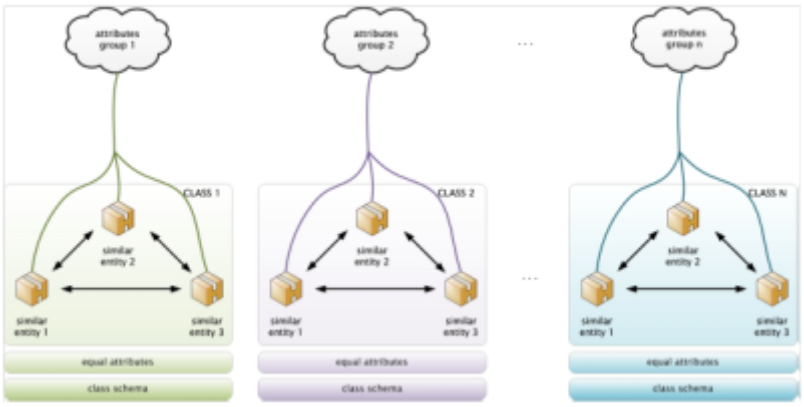
Oblasť Knowledge Discovery sa predovšetkým zaoberá vytváraním závislostí podobných informácií a následnou extrakciou vedomostí z týchto informácií a dát, teda na dáta sa budeme pozerať z pohľadu ich štruktúry:

1. homogénne – štruktúrované dáta,
2. heterogénne – semi-štruktúrované dáta,
3. primitívne a neprimitívne – neštruktúrované elektronické dáta.

Štruktúrované dáta

Štruktúrované dáta predstavujú typ dát majúcich jasne definovaný vlastný dátový model (obr.1), pričom medzi dátovými entitami existujú relácie. Pre takéto dáta existuje jasne popísaný dátový formát, dáta sú zhlukované v tzv. entitách, pričom podobné entity môžu byť zoskupované do ďalších usporiadaní. Štruktúrované dáta samé o sebe nenesú žiadne ďalšie špeciálne údaje o značkovaní (tzv. tagy) a taktiež nenesú nadbytočné informácie o svojej hierarchii alebo sémantike. Ich popisný model existuje nezávisle. O dátach ako o štruktúrovaných môžeme hovoriť až vtedy keď budú spĺňať nasledovné vlastnosti:

- sú organizované v sémantických entitách,
- podobné entity sú zoskupené do tried,
- medzi entitami v triedach existujú relácie,
- entity rovnakej triedy majú rovnaké atribúty,
- entity majú svoju schému:
  - zhodne definovaný formát,
  - preddefinovanú dĺžku.



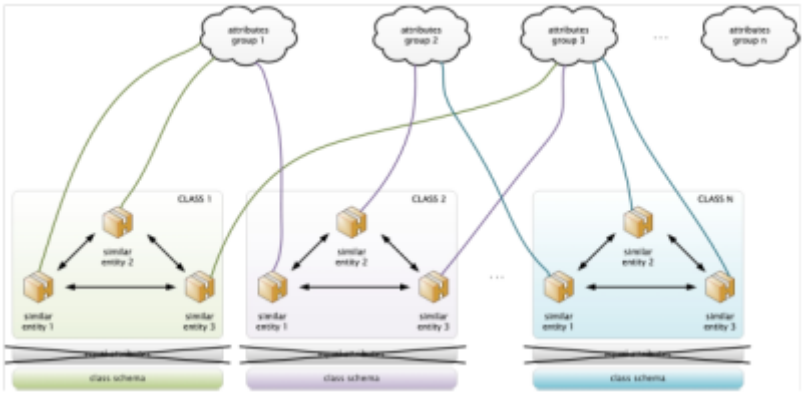
Obr. 1: Reprezentácia štruktúrovaných dát

Ako príklad štruktúrovaných dát môžeme uviesť dáta uložené v relačnej databáze. Pri

tvorbe modelu dát prostredníctvom databázy je v prvom kroku dôležité identifikovať aké entity a ich relácie sa náš model bude snažiť zachytiť. Samozrejme o každej entite musíme vedieť jednoznačne povedať akú množinu atribútov bude obsahovať. Ku každému atribútu následne presne špecifikovať jeho doménu, teda jeho typ a obmedzenia typu. Pri fyzickom návrhu entít sa ku každej entite vkladá množina atribútov, ktorá túto entitu vystihuje a typ týchto atribútov. Vytvárame teda popis štruktúry dát a popis typu dát, čo je vlastne logická schéma databázy. Úlohou tejto schémy je popisovať jednotlivé entity nachádzajúce sa v databáze a ich vzájomné relácie. Vytvorenie logickej schémy ako aj fyzický návrh entít v databázovom prostredí vytvára priestor pre ukladanie dát. Z tohto vyplýva, že štruktúrované dáta majú striktne oddelenú štruktúru od ich obsahu a pri transfere týchto dát, dáta nie sú sprevádzané žiadnymi tagmi, ktoré by určovali ich typ alebo formu. Ich štruktúra je určená priamo v ich cieľovom fyzickom úložisku.

Semi-štruktúrované dáta

- sú organizované v sémantický entitách,
- podobné entity sú zoskupené,
- entity rovnakej skupiny nemusia mať rovnaké atribúty,
- zoradenie atribútov nie je nutné,
- dimenzia atribútov zaradených v jednej skupine môže byť rôzna,
- typ atribútov zaradených v rovnakej skupiny môže byť rôzny.



Obr. 2: Reprezentácia semi-štruktúrovaných dát

Reprezentatmi tejto skupiny sú napr. XML dáta, skupina značkovacích jazykov (HTML, SVG, CFML, atď.), dáta typu EDI, vedecké dáta, atď.

Neštruktúrované dáta

Neštruktúrované dáta sú typom dát, ktoré nemajú svoj vlastný dátový model. Prvou skupinou neštruktúrovaných dát sú dáta, ktoré nemajú formálne definovanú štruktúru. Na základe analýzy dát môže byť táto štruktúra odvodená. Popisné dáta však nedokážu pomôcť pri procese spracovávania týchto vstupných dát a teda výstupom procesu spracovania sú nezmysluplné informácie. Za neštruktúrované sú vyhlásene vtedy, ak ich nie je možné na základe známych postupov spracovávať informačným systémom. Tieto dáta označíme ako neštruktúrovane. Všeobecne môžeme o neštruktúrovaných dátach vyvodiť nasledujúce vlastnosti:

- dáta ľubovoľného formátu,

- nesledujúce striktne daný formát alebo sekvenciu,
- neriadiace sa žiadnymi pravidlami,
- nepredikovateľné dáta.

Typickými reprezentantmi skupiny neštrúkturovaných sú dáta, ktoré sú vytvorené na základe lingvistických alebo vizuálnych štruktúr, ktoré sú predmetom tvorby dát konkrétnym programom. Skupinu neštrúkturovaných dát tvoria:

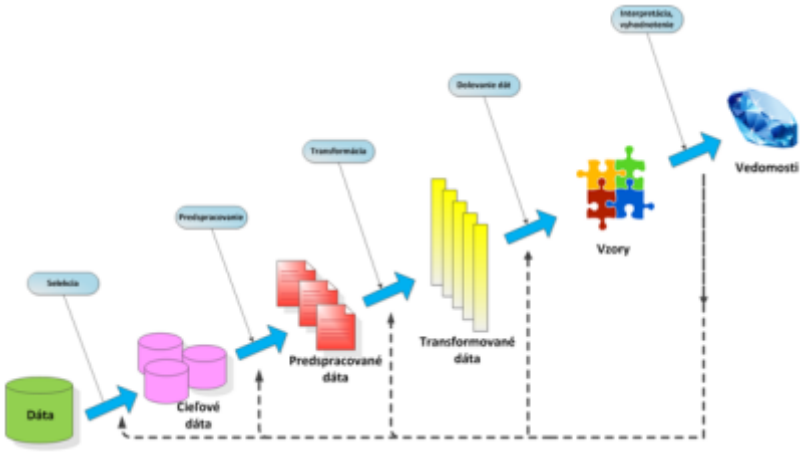
- audio záznam,
- video záznam,
- obrazové dáta (jpg, gif, atď),
- text,
- hlavičky emailových správ,
- …

Neštruktúrované dáta sú práve najproblematickejším zdrojom dát, keďže v každom kontexte môžu mať iný význam. Informácie z týchto dát nevieme jednoznačne popísať, nastáva tu problém s vytvorením štruktúrovaných dát z neštruktúrovaného vstupného dátového zdroja. Najjednoduchším spôsobom ako popísať štruktúru je priradiť neštruktúrovaným dátam metainformácie (dáta o dátach). Tento spôsob však nedokáže efektívne popísať celý kontext obsiahnutých dát.

Všeobecný proces Knowledge Discovery

Knowledge Discovery je nedeterministický proces. Preto je tento proces modifikovaný podľa vstupných dát a požadovaných výstupných informácii a vedomostí. Všeobecný proces objavovania znalostí obsahuje tieto základné kroky:

- Pochopenie aplikačnej domény
- Selekcia relevantnej (cieľovej) množiny dát
- Predspracovanie dát
- Transformácia dát
- Dolovanie dát
- Interpretácia, vyhodnotenie získaných poznatkov



Obr. 3: Všeobecný proces Knowledge Discovery

Ako už bolo spomenuté, proces objavovania znalostí v neštruktúrovaných dátach je veľmi špecifický a líši sa podľa charakteru vstupných dát. Jedným z

najkomplikovanejších procesov je dolovanie nad textovými dátami, ináč povedané nad prirodzeným jazykom písaným človekom. Tieto dáta sú poväčšine absolútne neštruktúrované a obsahujú tzv. kontextové vedomosti, teda vedomosti skryté v celých významoch slovných spojení a viet, ktoré nie je jednoduché strojovo identifikovať. Navyše slovenčina je flektívny jazyk, čo znamená, že slová sa ohýbajú, menia svoj tvar. Po zmene tvaru častokrát nie je zachovaný slovotvorný základ slova – koreň, teda slovo zmení svoj tvar oproti základnému tvaru. Následné je veľmi obtiažne strojovo identifikovať základ slova, teda jeho pôvodný význam a ďalej s týmto slovom pracovať.

Zdroj dát, v ktorých sme objavovali nové vedomosti boli lekárske záznamy, teda lekársky záznam o pacientovi, jeho anamnéze a vyšetrení. Tieto dáta boli síce v digitálnej podobe, ale všetko neštruktúrované obyčajné textové dokumenty, písané v prirodzenom jazyku, t.j. v nečitateľnej podobe pre počítač. Našim cieľom bolo:

- Analyzovať vstupné dáta
- Odstrániť dáta od šumov a rôznych redundantných dát
- Rozdeliť záznamy do kategórii podľa zamerania lekára (všeobecný lekár, gynekológ, pediater a pod.)
- Zoskupiť údaje v jednotlivých záznamoch do logických celkov
- Mapovať údaje v záznamoch do štruktúrovanej podoby

Ako finálna štruktúra, do ktorej sa budú mapovať dáta z lekárskych záznamov bol zvolený štandard openEHR, ktorý spĺňa normu ISO 13606-2 pre zdravotnícku informatiku. OpenEHR platforma je otvorený štandard pre zdravotné údaje, ktorý popisuje správu, ukladanie, vyhľadávanie a upravovanie zdravotných údajov v elektronických zdravotných záznamoch. Štandard nepopisuje výmenu dát medzi systémami EHR, pretože to je obsahom iných noriem. V openEHR sú všetky zdravotné údaje jednej osoby uložené v jednom „životnom“ zázname.

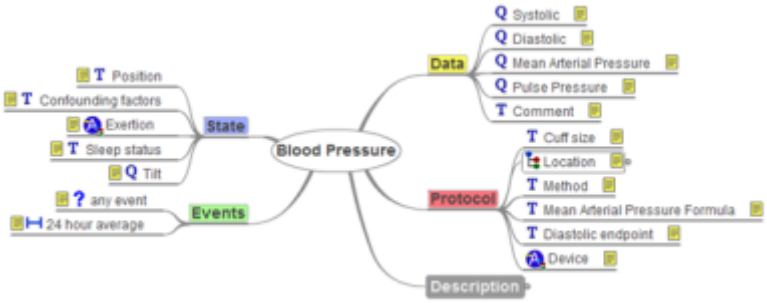
OpenEHR platforma je open-source softvérová infraštruktúra pre implementovanie komplexného EHR v klinickom prostredí. Platforma je založená na kombinácii pätnásťročného európskeho a austrálskeho výskumu, vývoja EHR a nových paradigiem, vrátane toho čo sa dnes nazýva archetypová metodika pre špecifikáciu obsahu. Platforma definuje zdravotný informačný referenčný model, jazyk pre vytváranie „klinických modelov“ respektíve archetypov, ktoré sú oddelené od softvéru a dotazovacieho jazyka. Architektúra je navrhnutá, aby mohla využívať externé zdravotné terminológie ako SNOMED CT, LOINC a iCDX.

Dvojúrovňový model

Kľúčovou inováciou openEHR platformy je oddelenie špecifikácií zdravotných údajov od referenčného modelu popisujúceho štruktúru. Najdôležitejšou úlohou je, aby platforma bola prostriedkom pre ukladanie potrebných vyjadrení lekára a pacienta, takže informácie sú k dispozícii kdekoľvek sú potrebné.

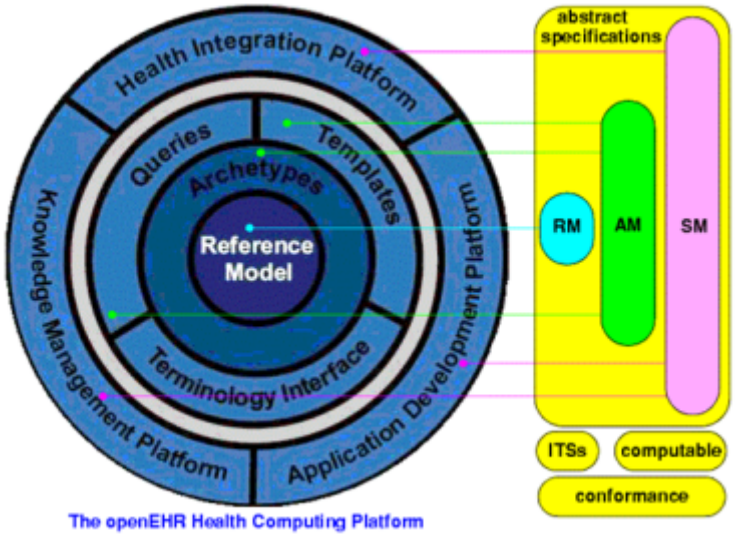
Zdravotné údaje sú špecifikované v dvoch typoch artefaktov, ktoré sú nezávislé na referenčnom modeli. Prvý typ archetyp (viď obr. 4), poskytuje miesto pre formálne definovanie znovu použiteľných údajových položiek a údajových skupín, t.j. obsah položiek, ktoré budú opätovne použité. Typickým príkladom je meranie krvného tlaku.

Každé meranie krvného tlaku pozostáva zo systolického a diastolického tlaku a iných parametrov, ktoré sa pri každom meraní tlaku opakujú. Druhým typom artefaktu je šablóna (template), ktorá je použitá na reprezentáciu súboru dát v konkrétnych prípadoch, ako napríklad sumárne vyšetrenie pacienta alebo rádiologický záznam. Šablóna je skonštruovaná z odkazov na príslušné položky z viacerých archetypov. Šablóny sú takmer vždy vyvinuté pre miestne použitie softvérovými vývojármi spolu so zdravotníckymi analytikmi.



Obr. 4: Príklad archetypu Tlak krvi

Celý rozsah špecifikácie štandardu openEHR je zobrazený na nasledovnom obrázku:



Obr. 5: Open EHR špecifikácia

Štandard nesie zodpovednosť za vytvorenie abstraktných špecifikácií, na ktorých je postavená openEHR zdravotnícka platforma. Tieto abstraktné špecifikácie sú tri:

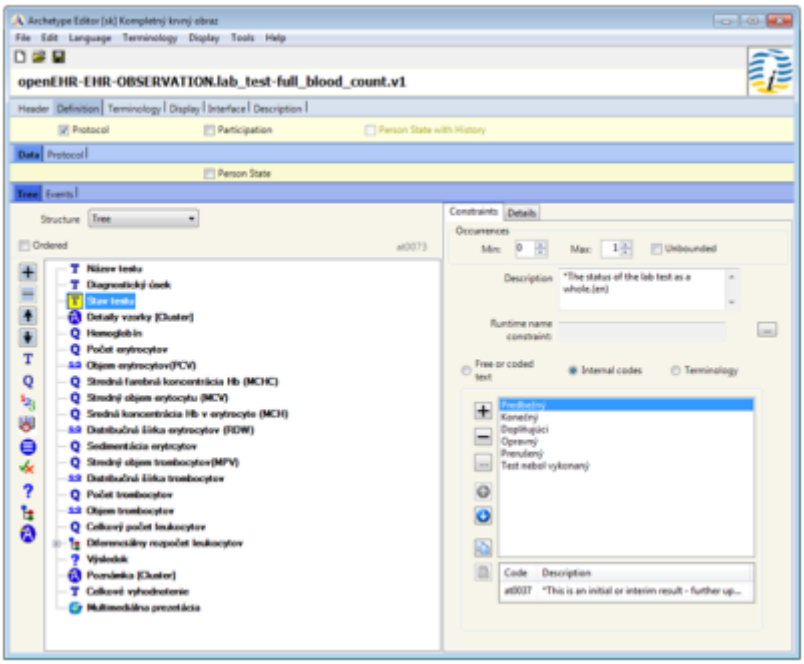
- Referenčný model (RM)
- Archetypový model (AM)
- Model služieb (SM)

Uvedené špecifikácie predstavujú formálny podklad pre vytvorenie osobitných vrstiev openEHR platformy, pozostávajúcej z referenčného modelu, archetypov a šablón, dotazov a terminologických rozhraní. V openEHR špecifikácií sú zahrnuté tri hlavné balíčky, a to: RM (referenčný model), AM (archetypový model) a SM (model služieb).

Pre nás je najzaujímavejší práve Archetypový model, lebo ten definuje, do akej štruktúry majú byť neštruktúrované dáta lekárskych záznamov ukladané. Archetypy sú

základným kameňom openEHR architektúry. Kým v minulosti navrhovali dátové štruktúry používané v zdravotníctve najmä informatici, poprípade informatici v spolupráci s lekármi, tak pri archetypoch je to úplne inak. Jednotlivé archetypy sú navrhované a tvorené doménovými odborníkmi a lekármi. Každý archetyp predstavuje špecifikáciu pre jednoduchý, diskrétny lekársky obsah (napr. krvný tlak, pulz, plán liečenia, nález, záznam o ošetrení, rôzne lekárske pozorovania). Tento obsah je zahrnutý v dátových elementoch, ktoré dávajú obsahu lekársky význam pre všetky predstaviteľné lekárske situácie. Definícia archetypov je široká a obmedzenia sú minimálne, tak aby bolo možné čo najviac maximalizovať interoperabilitu v smere zdieľania a znovupoužitia archetypov v rôznych smeroch zdravotnej starostlivosti.

Špecifikácia archetypov je vyjadrená pomocou jazyka ADL (Archetype Definition Language), ktorý predstavuje ISO štandard. Hoci je jazyk ADL primárne určený pre čítanie a zápis pomocou nástrojov, je čitateľný aj pre človeka. Archetypy možno upravovať manuálne pomocou bežného textového editora. Definícia archetypov obsahuje tri časti: popisné dáta, pravidlá pre vyjadrenie obmedzení a ontologické definície.



Obr. 6: Príklad archetypu Kompletného krvného obrazu v Archetype editore, preložený do slovenčiny

Popisné dáta obsahujú jedinečný identifikátor archetypu, ďalej informácie ako meno autora archetypu, verziu archetypu, popis jeho využitia a informáciu o jazykovej verzii archetypu. Pravidlá pre obmedzenia špecifikujú obmedzenia pre dosiahnutie validnej štruktúry, definujú počet výskytov jednotlivých prvkov v archetype a formu ich obsahu v súlade s archetypom. Ontologická časť definuje súvisiaci terminologický slovník (t.j. strojovo kontrolované kódy), ktoré môžu byť použité v špecifických uzloch inštancií archetypov. Archetypy sú jazykovo neutrálne, čo znamená, že môžu byť prepísané a preložené do ľubovoľného iného jazyka. Taktiež sú plne adresovateľné štýlom podobným ako pri XML dátach, s využitím tzv. path výrazov, ktoré je možné konvertovať priamo na XPath výrazy.

Archetypy sme zvolili ako cieľovú štruktúru, do ktorej sme vkladali údaje z lekárskych

záznamov. Bolo teda potrebné lekárske záznamy spracovať a pripraviť na štruktúrovanie a vkladanie do archetypov. Preto bol navrhnutý nasledovný postup, ktorý vychádzka zo všeobecného procesu objavovania znalostí v textoch:

- Odstránenie šumov v lekárskych záznamoch
- Kategorizácia záznamov podľa odboru lekára
- Rozdelenie lekárskych správ na jednotlivé odstavce
- Lematizácia
- Skúmanie zoznamu najčastejších slov v správach
- Doplnenie lematizátora
- Kategorizácia odstavcov lekárskych správ
- Zoskupovanie údajov do logických celkov
- Mapovanie údajov na archetypy
- Ukladanie informácii do EHR banky

Poďakovanie

Podporujeme výskumné aktivity na Slovensku / Projekt je spolufinancovaný zo zdrojov EÚ

Literatúra

1. LESLIE, H. Introduction to Archetypes and Archetype classes. [online]. 2012. [cit. 2013-03-06]. Dostupné na: http://www.openehr.org/wiki/display/healthmod/Introduction+to+Archetypes+and+Archetype+classes
2. Eichelberg, M. et al. A survey and analysis of electronic healthcare record standards. In ACM Computing Surveys. [online]. 2005, vol. 37, no. 4 [cit. 2013-05-05]. Dostupné k stiahnutiu na: http://dl.acm.org/citation.cfm?id=1118891&bnc=1
3. http://trac.openehr.jp/wiki/Archtectural%20Overview%20Overview
4. http://en.wikipedia.org/wiki/SNOMED_CT
5. Paralič, Ján, a iní. 2010. Dolovanie znalostí z textov. Košice : Technická univerzita v Košiciach, 2010. ISBN 978-80-89284-62-7