Dokumentační služba projektu Medigrid : dokumentování sémantiky lékařských dat

Adéla Jarolímková[1], Petr Lesný[2], Jan Vejvalka[2], Kryštof Slabý[2], Tomáš Holeček[3]

[1] Cesnet z.s.p.o.
[2] Fakultní nemocnice Motol
[3] Fakulta humanitních studií UK

INFORUM 2006: 12. konference o profesionálních informačních zdrojích
Praha, 23. - 25.5. 2006

Abstrakt: *Lékařské terminologie a klasifikační systémy obvykle sestávají z tezauru biomedicínských pojmů či konceptů na jedné straně a souboru vztahů mezi nimi na straně druhé. Základní vztahy mezi pojmy ve většině těchto systémů maji charakter taxonomický, dále bývá vyjádřena synonymie či meronymie (vztah termínů označujících celek a část). Medicína se však zabývá daleko složitějšími vztahy mezi koncepty a entitami. Příkladem takového vztahu jdoucího napříč tradičními hierarchickými systémy je BMI (Body Mass Index), vztah mezi tělesnou výškou a tělesnou hmotností. Aby bylo možno aplikovat potenciál obecných lékařských znalostí na obrovské množství dat souvisejících se zdravotní péčí (obsažených např. v elektronických záznamech), musí být znalosti a data popsány způsobem umožňujícím sémantické vyhledávání.*
Pro popis dat figurujících v lékařských výpočtech, které reprezentují vztahy složitější než tradiční klasifikační systémy a hierarchie, jsou existující klasifikační systémy použitelné pouze v omezené míře. Pro projekt Medigrid, který se zabývá právě takovými výpočty, byla proto navržena sémanticky orientovaná Dokumentační služba, která řeší problematiku jednoznačného popisu dat ve výpočtech a představuje solidní dokumentační bázi včetně databáze použitých citací.

Projekt Medigrid

Projekt MediGrid, jehož účastníky jsou CESNET z.s.p.o., Fakultní nemocnice v Motole a Masarykova nemocnice Ústí nad Labem, je financován z grantu č. 1ET202090537.

Jeho cílem je návrh, vývoj a pilotní implementace MediGridu - prostředí a modulárního systému aplikací pro distribuované zpracování datových a výpočetních úloh ve zdravotnictví (ve zdravotnickém výzkumu i praxi). Základní sjednocující technologií bude síť Grid, která bude umožňovat vkládání jednotlivých odborných modulů pro sdílení dat (včetně modulů s možností sdílení dat v reálném čase a telekonzultací), modulů pro sběr a analýzu dat a modulů pro výměnu odborných informací (1).

Jednotlivé součásti systému je nutno jednoznačně popsat tak, aby byl umožněn snadný přístup k aplikačním modulům a jejich spolupráce při složitějších operacích (např. automatické řetězení modulů, při čemž výstup jednoho modulu je vstupem dalšího). Protože popis pomocí tradičních terminologických systémů a klasifikací nepostihuje v dostatečné míře

složitost vztahů v rámci lékařských výpočtů, bylo třeba navrhnout nové řešení, které je součástí Dokumentační služby.

Uspořádání MediGridu

MediGrid pracuje s indikátory; indikátor je zde popisován v souladu s Husserlem jako záznam, který někdo pořídil sám (nebo pomocí automatu) pro své budoucí použití nebo pro použití někoho jiného (2). Podstatou indikátoru je , že jeho přečtení v někom vyvolává přesvědčení o něčem jiném. Například přečtení záznamu „Body Mass index je 20.4" v dokumentaci pacienta vyvolává v lékaři, který se do záznamu dívá, přesvědčení o tom, že pacientovi byl vypočítán Queteletův index tělesné hmotnosti (BMI) a jeho hodnota je 20.4. Cokoli takto slouží, je indikátor.

Z jednoho nebo více indikátorů lze někdy pořídit nový pro lékaře užitečný indikátor, aniž by se toho sám lékař účastnil. Lékař totiž může tuto práci nechat na člověku postupujícím podle algoritmu nebo na automatu. Například může nechat na automatu výpočet BMI ze záznamů o výšce a hmotnosti, čímž se transformují indikátory měření výšky a hmotnosti na indikátor BMI. Automatizovaný nástroj pro transformaci indikátorů v systému MediGrid nazýváme modul; tento modul implementuje relaci mezi indikátory.

Podle jejich úlohy v transformacích můžeme dělit indikátory do tříd (například indikátory dokumentující tělesnou výšku). Z hlediska dokumentace můžeme popisovat modul jako relaci mezi třídami indikátorů.

Systém MediGrid je tedy složen z kolekce modulů (každý implementuje jednu relaci), řadiče a dokumentační služby. Řadič je rozšířením mechanismů sítě GRID, sloužící k automatickému řetězení modulů a dokumentační služba slouží k jednoznačnému popisu sémantiky lékařských dat, zejména modulů (relací) a tříd indikátorů.

Problém jednoznačného popisu dat a jejich vztahů

V oblasti medicíny existuje velké množství terminologických systémů a klasifikací, které jsou využívány k různým účelům a v rozdílných oblastech, například k indexování záznamů v bibliografických databázích slouží heslář MeSH, pro klasifikaci chorob se užívá International Classification of Diseases a další. Společnou vlastností těchto systémů je jejich hierarchické uspořádání, při čemž vztahy mezi jednotlivými pojmy mají převážně charakter taxonomie (is_a), synonymie či meronymie (part_of). Pro účely znalostně orientovaných aplikací nejen v medicíně se toto ukázalo jako nepostačující, proto se od 90. let v této souvislosti objevuje termín ontologie ve smyslu explicitní specifikace konceptualizace (tedy

nikoliv v původním filozofickém významu jako nauka o bytí či univerzální soustava znalostí popisující objekty, jevy a zákonitosti světa „tak jak je") (3).

Mezi medicínské ontologie jsou řazeny především některé tradiční systémy, jako je SNOMED či UMLS (Unified Medical Language System), které si nárokují pokrytí celé oblasti medicíny, avšak z přísně formálního hlediska obsahují řadu inkonzistencí a nepřesností (4) a dále systémy nově vyvíjené právě s ohledem na udržení konzistence a jednoznačnosti používaných termínů a vztahů, k nimž patří např. OpenGALEN nebo On9.

Kromě časté vágnosti a nejednoznačnosti používaných termínů je dalším problémem při popisu dat v lékařských algoritmech pomocí existujících klasifikačních systémů nebo ontologií také binární charakter vztahů mezi pojmy. Obtížnost popisu lze demonstrovat na příkladě výpočtu indexu tělesné hmoty (body mass index, BMI). Vstupem je tělesná váha a tělesná výška pacienta, výstupem BMI. Podívejme se, jak by bylo možné tento algoritmus popsat s pomocí UMLS.

UMLS je soubor databází (UMLS Knowledge Sources) a softwarových nástrojů určených pro budování elektronických informačních systémů, které vytvářejí, zpracovávají, vyhledávájí, integrují či shromažďují medicínské informace (5). Skládá se z Metathesauru, multilinguální databáze biomedicínských koncpetů, jejich názvů a synonym a vzájemných vztahů, z více než 100 biomedicínských a zdravotnických slovníků, tezaurů, klasifikací a kódovníků (např. MeSH – indexace biomedicínských databází a katalogů, SNOMED, ICD10, LOINC, RxNorm, GO aj.), „sémantické sítě" (Semantic Network), která obsahuje informace o typech/kategoriích, do nichž jsou přiřazovány koncepty Metatezauru), a všech vztazích mezi těmtito typy, a s jejíž pomocí lze odvodit vztahy mezi koncepty. Třetí součástí je SPECIALIST Lexicon obsahující syntaktické informace pro termíny a jejich komponenty vyskytující se v Metatezauru, společně s řadou obecných anglických výrazů. Pro potřeby popisu dat jsou podstatné koncepty Metathesauru a sémantická síť.

Pro všechny třídy indikátorů v relaci BMI lze v Metathesauru nalézt odpovídající koncepty: body weight (cui C0005910), body height (cui C0005890), body mass index (cui C1305855). Při bližším zkoumání zjistíme, že koncepty „body weight" a „body height" náležejí k sémantickému typu Organism Attribute, „body height" ještě k typu Quantitative Concept, a „body mass index" k typu „Clinical Attribute". Vztahy mezi koncepty Metathesauru nejsou explicitně specifikovány, s vyjímkou is_a, je možné je pouze odvodit na základě vztahů mezi odpovídajícími sémantickými typy v sémantické síti. Mezi typy „Organism Attribute" a „Clinical Attribute" jsou tyto vztahy: Clinical Attribute is_a Organism Attribute, Clinical Attribute associated with Organism Attribute, Clinical Attribute degree of Organism

Attribute, Organism Attribute associated with Clinical Attribute a Organism Attribute degree of Clinical Attribute, z nichž žádný nevypovídá o vztahu váhy, výšky a BMI.

Problémem ontologií, jako je On9, které jsou po formální stránce propracovanější a disponují daleko širší škálou vztahů mezi pojmy, je malý počet doposud zpracovaných konceptů, který prozatím neumožňuje jejich širší využití.

Řešením, které bude použito v Dokumentační službě MediGridu, je tzv. ad-hoc či ex-post ontologie, generovaná automaticky na základě strojem zpracovatelných součástí Dokumentační služby. Funkčnost a výtěžnost této ontologie bude otestována ve zkušebním provozu MediGridu.

Dokumentační služba

Hlavním úkolem Dokumentační služby (dále jen DS) je, jak již bylo řečeno, řešit problematiku jednoznačného popisu dat ve výpočtech, zároveň však také sloužit jako prostředek pro komunikaci odborníků z jednotlivých oblastí, sdílení často unikátních znalostí a podklad pro hodnocení systému v rámci EBM (evidence based medicine).

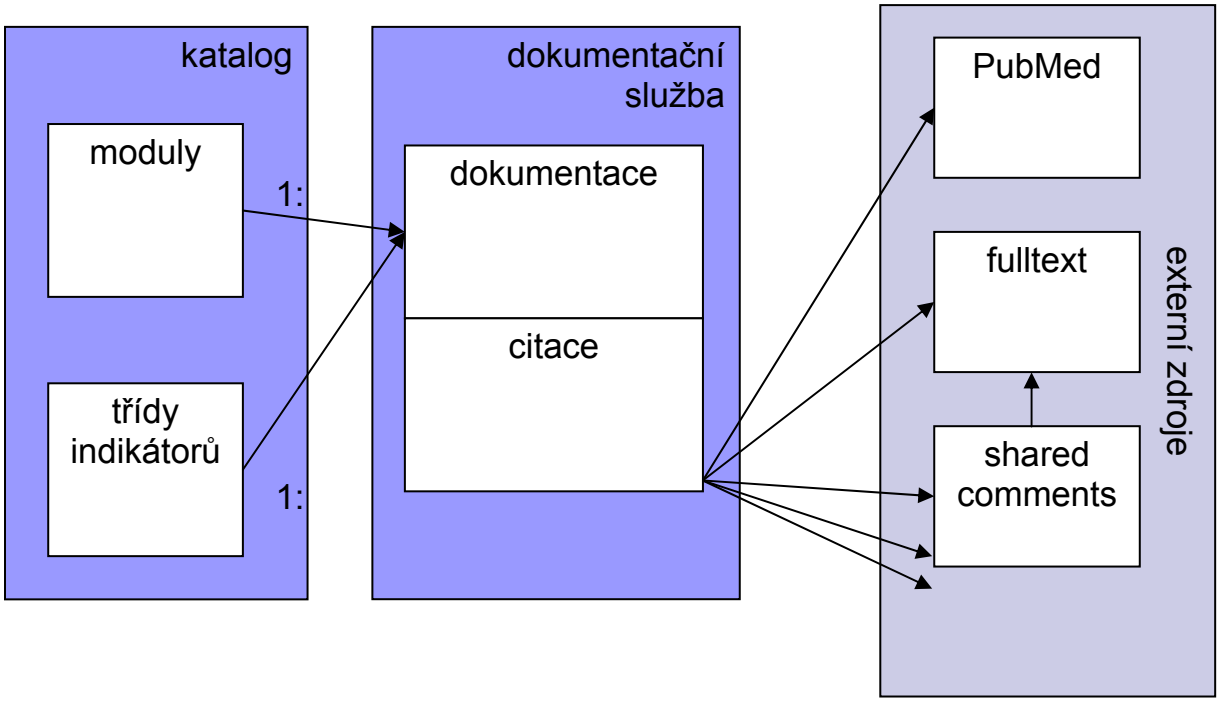
V rámci DS jsou uchovávána metadata tří základních kategorií entit: odborných modulů, tříd indikátorů a citací, a to jak ve strojem zpracovatelné podobě, která je podkladem pro vytváření ad-hoc ontologie, tak v podobě lidsky čitelného popisu doplněného citacemi relevantních dokumentů.

Dokumentace tříd indikátorů a modulů

Třída indikátorů je určena pojmem, který je převzat z některého z kontrolovaných slovníků. Původním záměrem bylo využít pro popis indikátorů pouze konceptů UMLS Metathesauru, avšak v řadě případů jsou tyto koncepty příliš vágní, případně se v Metathesauru nevyskytují vůbec. Je proto možné k označení třídy indikátorů použít i jiné kontrolované slovníky a v případě, že se pojem v žádném dostupném slovníku nevyskytuje, také slovník uživatelský. Součásti dokumentace je i kód autora a popis v rozsahu nezbytném pro odlišení od jiných tříd indikátorů, který může citovat relevantní dokumenty z dané oblasti.

Obdobně dokumentace modulu obsahuje kromě kódu autora, názvu modulu a jeho URI také popis ve formě strukturovaného textu a relevantní citace.

Jednoduché schéma DS



Ukládání citací

Část DS, která slouží pro správu citací, funguje jako jednoduchý „reference manager" (typickými představiteli tohoto typu aplikací jsou např. EndNote či ProCite), umožňující zejména uložit záznam jakékoliv publikované i nepublikované informace, a to manuálně či stažením záznamu z externí databáze, vložit citaci do textu dokumentace, propojit citaci na externí zdroj, např. fulltext či bibliografický záznam ze zdrojové databáze. Doplňkovými funkcemi čistě pro účely MediGridu je hodnocení kvality citace, které probíhá semiautomaticky na základě uživatelem zadaných údajů, a rovněž specifikace vztahu mezi citujícím dokumentem, tj. samotným textem dokumentace, a citovaným dokumentem.

Ve verzi 0 jsou (byl) citace provizorně ukládány jako plain-text v podobě podle tzv. vancouverské konvence (uniform requirements), která se používá ve většině zahraničních lékařských časopisů. Pro další verze byl hledán formát záznamu, který splňuje následující požadavky:

1. založený na XML s XML Schematem použitelným na dané aplikační úrovni. Veškerá metadata obsažená v DS jsou ukládána v XLM a jejich podobu určují XML schémata. Formát nesmí být příliš komplikovaný, neboť nejde o budování plnohodnotné bibliografické databáze pro knihovnické účely a aplikace bude používána odborníky z lékařské, nikoliv knihovnicko-informační, oblasti.
2. Umožňuje citovat jakoukoliv publikovanou i nepublikovanou informaci v jakékoliv podobě.
3. Obsahově splňuje minimálně požadavky vancouverské konvence.

4. Je libovolně rozšiřitelný, neboť jak bylo již řečeno, pro potřeby MediGridu budou přidávána některá pole.
5. Kompatibilní zejména s PubMed.

Původním záměrem bylo využít některé z XML schémat stávajících široce rozšířených formátů pro ukládání metadat, jakými jsou např. MARC 21 či Dublin Core. Za tímto účelem byla analyzována dostupná XML schémata, případně DTD (pokud není schéma k dispozici) jak zmíněných formátů MARC 21 a Dublin Core, tak dalších, především NLM Medline/PubMed, odkud budou nejčastěji přebírány záznamy, dále např. formát používaný pro projekt DiVA, ShaRef, aj. Jako nejschůdnější se jeví použití formátu MODS (Metadata Object Description Schema), vyvíjeném Kongresovou knihovnou jako XML obdoba MARC 21, z něhož po přizpůsobení potřebám DS vznikne vlastní formát.

Závěr

Funkčnost DS služby v navrhované podobě, především její „ontologické" části, bude muset být teprve prokázána v testovacím provozu, protože se jedná o unikátní, dosud neověřené řešení popisu dat v lékařských algoritmech. Pokud se tento model osvědčí, vznikne nejen prostředí pro zpracování lékařských algoritmů, ale i jedinečný odborný zdroj využitelný v dalších projektech týkajících se zpřístupňování informací v medicíně.

Literatura

1. MediGrid : o projektu [online]. [c2005/2006]. [cit. 2006-04-18]. Dostupný z WWW: < http://www.medigrid.cz/cs/oprojektu/index.html>
2. HUSSERL E. Logical Investigations. Volume 1. Investigation I. London: Routledge, 2001. s. 184, § 2.
3. SVÁTEK, Vojtěch. Ontologie a WWW. In: DATAKON 2002. Brno : Masarykova univerzita, 2002. S. 1-35.
4. CEUSTERS, Werner, SMITH, Barry, KUMAR, Anand, DHAEN, Christeoffel. Mistakes in medical ontologie : where do they come from and how they can be detected?. In: PISANELLI, Domenico (ed.) Ontologies in medicine : proceedings of the Workshop on Medical Ontologies. Amsterdam : IOS Press, 2003.
5. Fact Sheet UMLS [online]. Bethesda> National Library of Medicine, 2006. [cit. 2006-04-18]. Dostupný z WWW: <http://www.nlm.nih.gov/pubs/factsheets/umls.html>