DATOVÉ MODELY JAKO NÁSTROJ PRO INTEROPERABILITU

Michal Huptych, Lenka Lhotská, Miroslav Burša

Anotace

Největší problémy a zároveň největší prostor pro budoucí řešení jsou v oblasti správného mapování získávaných dat do datového modelu, který popisuje elektronický záznam pacienta. Zejména s ohledem na budoucí vývoj a možnost snímat a ukládat daleko větší objemy různorodých fyziologických parametrů je otázka interoperability stále důležitější. V této souvislosti před námi stojí celá řada úkolů. Tím nejdůležitějším je kvalitní datová analýza potřebných pacientských dat v jednotlivých lékařských odbornostech, na jejímž základě bude možné identifikovat množinu dat společnou pro všechny odbornosti (či alespoň část odborností) a definovat data specifická pro každou odbornost a následně vytvořit datové modely, které budou sloužit jako základ pro příslušné dokumenty v EHR.

Studie z posledních let ukazují, že otázka interoperability může významně ovlivnit efektivitu jak při návrhu a tvorbě integrovaného systému, tak i při vlastním provozu. Pokud interoperabilita mezi lékařskými přístroji a zdravotnickými IS opravdu funguje, je možné snížit náklady na integraci (udává se až o 30%), čas pro mapování datových typů až o 50%, a podstatně zvýšit přesnost dat v elektronickém zdravotním záznamu. Nemluvě o tom, že tato automatizace znamená usnadnění práce lékařského a zdravotnického personálu a možnost zaměřit se na vlastní práci s pacientem.

V příspěvku popíšeme proces tvorby datového modelu pro reprezentaci heterogenních dat, tvorbu ontologií nad datovým modelem a naznačíme možnosti využití v praktické aplikaci.

Klíčová slova

Informační systém, interoperabilita, standard, datový model, ontologie, heterogenní data

1. Úvod

Tématem problematiky využití ICT a hlavně standardů v medicíně jsme se na seminářích Medsoft již nejednou zabývali, např. v [1]. V tomto příspěvku se konkrétněji zaměřujeme na oblast, která je velmi důležitá, ale nemá zatím jasně dané propozice – heterogenní data a možnosti interoperability v oblasti signálů. Za heterogenní data považujeme zpravidla data, pocházející z několika různých zdrojů a uložená obvykle v různých formátech. V medicíně založené na důkazech (evidence-based medicine) se při rozhodování téměř vždy pracuje s heterogenními daty. Bohužel automatické zpracování a vyhodnocování takových dat je velmi omezené. Dosud nejsou k dispozici takové nástroje, které by umožnily plně automatické zpracování a vhodnou reprezentaci vzájemných vztahů mezi daty. V medicíně je definování vztahů mezi daty z různých zdrojů

integrální součástí kognitivního procesu. Jenže objemy a heterogenita získávaných dat představují velký problém jak pro manuální, tak pro automatické zpracování dat. Proto jsme se v rámci našeho výzkumu zaměřili na otázku nalezení vhodné metodologie pro zpracování heterogenních dat v obecné rovině. Zaměřili jsme se na jednotlivé fáze tohoto procesu: ukládání heterogenních dat, jejich fúze a zpracování. Hlavním důvodem je to, že tyto kroky představují důležitý základ pro vývoj mnohavrstvého datového modelu [2].

V oblasti znalostního a databázového inženýrství jsou heterogenní data často reprezentována datovým modelem, který je zaměnitelný a reprezentuje informace z každého zdroje a navíc umožňuje tyto informace kombinovat. Cílem je pak získat nové znalosti ze vzájemných vztahů v datech z několika zdrojů. Reprezentace použité pro integraci heterogenních dat musejí umožnit popis dat, schémat a kontextu. V literatuře můžeme najít celou řadu různých přístupů: objektově orientované modely [3], ontologie [4],[5],[6], statistické modely [7] nebo nepřímé (mediated) datové modely [8]. Jako velmi vhodné se ukazuje využití ontologií. Umožňují vhodně vyjádřit sémantickou heterogenitu [9]. Přehled hlavních typů ontologických architektur je velmi dobře popsán v [6].

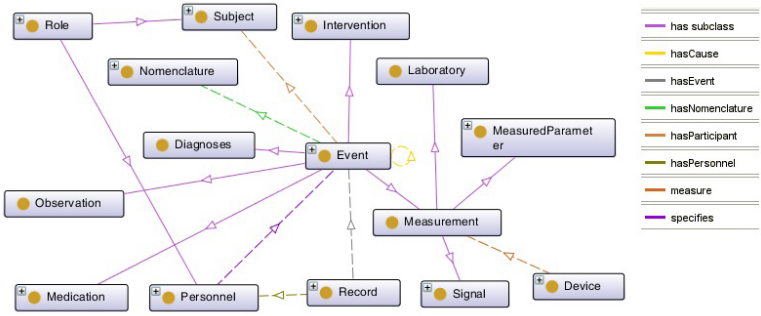
Jednou z nejdůležitějších částí zpracování heterogenních dat je fúze dat, jejímž hlavním cílem je vhodným způsobem kombinovat informace z různých zdrojů, např. senzorů, a přispět tak k hlubšímu pochopení dané situace. Příkladem může být práce [10], kde je popsána fúze EKG signálu, krevního tlaku, SpO2 a dýchání s cílem zlepšit klinickou diagnostiku na jednotce intenzivní péče. Řada hodnot je odvozena z původních záznamů fyziologických parametrů (např. tepová frekvence, frekvence dýchání, systolický, střední a diastolický tlak). Tyto odvozené hodnoty jsou pak použity jako příznaky pro fúzi dat. V této fázi lze použít řadu metod, vycházejících z oblasti strojového učení a dolování dat. Dalším příkladem fúze dat je práce [11], která popisuje spojení elektrofyziologických a hemodynamických signálů pro sledování ventrikulárního rytmu. Velmi důležitou roli zde mohou hrát funkce „vzdálenosti" parametrů. U heterogenních dat mohou být tyto funkce kombinované pro spojitá a kategoriální data. Některé z funkcí „vzdálenosti" pro heterogenní data lze najít např. v práci [12]. Další příklady využití strojového učení v systémech s heterogenními daty lze najít v pracích [13][14] a [15].

V dalších částech příspěvku se budeme podrobněji zabývat návrhem mnohavrstvého datového modelu.

2. Mnohavrstvý datový model

Základním stavebním kamenem modelu jsou události. Jsou definované jako pozorování, měření nebo nález určitých faktů o daném subjektu. Za událost považujeme např. stanovení diagnózy, medikaci, intervenci, měření nebo pozorování. V případě měření může být událostí měřený signál či parametr a/nebo výsledky laboratorního testu, který se provádí z odebraných vzorků.

Pozorování je naproti tomu založeno na subjektivním hodnocení osobou (expertem). Událost může být spojena s určitým slovníkem (číselníkem, nomenklaturou, jinou ontologií). Koncept ontologie události je pro ilustraci uveden na obr. 1. Pro vytvoření ontologií je použit systém Protégé [16].

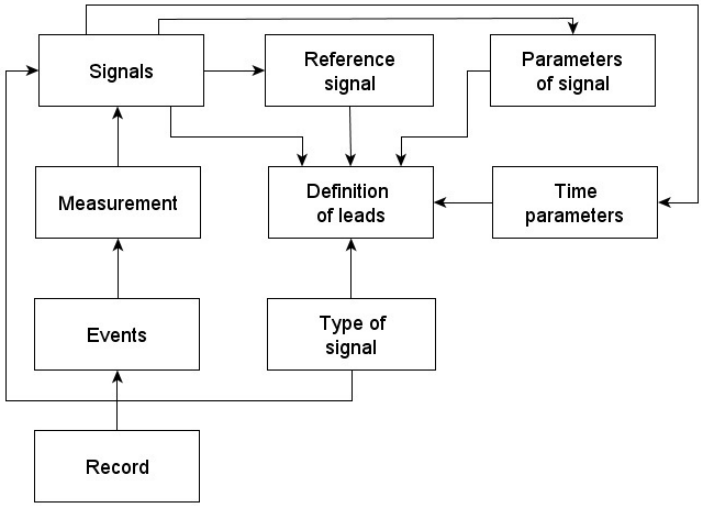


Obrázek 1 — Koncept ontologie události

Z hlediska popisu časových vlastností můžeme definovat dva typy událostí. První typ je definován svým začátkem a koncem (např. podávání léku infuzí kontinuálně), druhý typ události je jednorázový (např. injekční podání léku). Druhou možností pro specifikaci události je její definování jako parametrické (vyžaduje zadání hodnoty a jednotky) nebo neparametrické. Mezi událostmi může být definována vazba, která určuje kauzalitu událostí. Jinými slovy jedna událost může být definována jako důsledek jiné události. V moderní medicíně se čím dál častěji setkáváme s měřením biologických signálů, které jsou nedílnou součástí diagnostického procesu. Soubor invazivních a neinvazivních testů je velmi rozsáhlý a objem signálových dat, který je možné získat během vyšetření, může být značný. Existující standardy neobsahují žádné možnosti pro dekompozici signálů do sémantických popisů.

Z pohledu standardů pro komunikaci a výměnu informací je nejčastějším řešením připojit signál pomocí reference. Pro uložení signálů existuje celá řada formátů, z nichž některé používají složité záhlaví (obsahuje úplnou informaci o typu signálu, vzorkovací frekvenci, senzitivitě a velikosti datového bloku). Tyto informace jsou zásadní pro následné zpracování, ale neumožňují sémantický přenos. Problémem je totiž variabilita signálů a fakt, že složitější popis v záhlaví znamená složitější syntaktický analyzátor pro získání informací. Signál principiálně obsahuje velký počet charakteristických bodů a intervalů, které mohou být významné pro následnou analýzu signálu. Identifikace těchto bodů a intervalů je cílem většiny analýz. Musíme si však uvědomit, že tyto intervaly nemusejí být nutně dány pouze povahou signálu. Často jsou spojeny s fyziologickými, resp. patofyziologickými procesy v těle. Události, které ovlivňují subjekt, jsou přídavkem k charakteristickému intervalu.

Může tedy nastat překryv těchto dvou definic intervalů v signálu. Například začátek a konec ventrikulární fibrilace lze určit přímo ze signálu. Ale může být také definován uloženými událostmi začátku a konce ventrikulární fibrilace v nezávislém systému s přesnou časovou synchronizací. Dochází k shodě a je zřejmé, že diagnóza v tomto případě koreluje s vlastnostmi signálu a znalost lékaře a systému je shodná (i když její reprezentace a proces vyvozování shodné být nemusí). Definované body a komplexy lze pak reprezentovat pomocí nástrojů sémantické interoperability. V dalším textu se budeme zabývat popisem návrhu, který definuje uložení signálu se všemi informacemi, a popisem transformace, která převádí signál do strukturované formy. Základní koncept svodu (signálu) je znázorněn na obr. 2.



Obrázek 2 — Základní koncept svodu (signálu) v závislosti na události a parametrů signálu

Nejprve definujeme jednotlivé pojmy. Předpokládáme, že svod (signál) může být odkazován nebo že jeden signál či kombinace signálů lze použít jako referenci a datový model pracuje s více typy signálů.

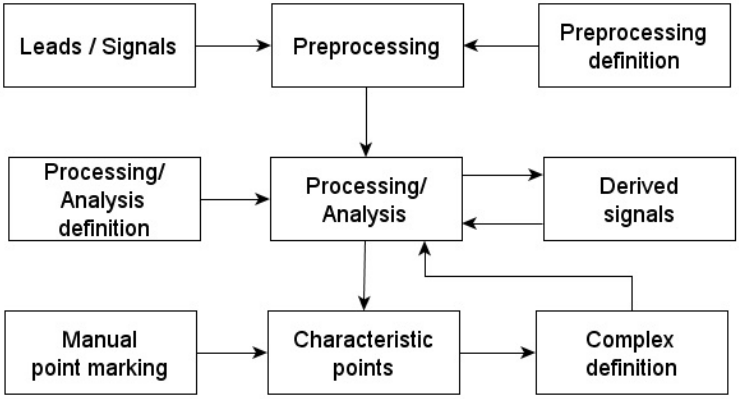
Typ signálu je velmi důležitou položkou v našem konceptu, protože další podmínky zpracování jsou závislé na typu signálu. Klíčová je tato informace pro výběr metod pro extrakci příznaků a evaluaci.

Parametry signálu obsahují základní informaci o signálu (např. vzorkovací frekvenci, senzitivitu, maximální hodnotu, minimální hodnotu, formát souboru). Tyto informace slouží pro jednoduchý přenos původních dat a jejich reprodukovatelnost.

Časové parametry určují základní časové vlastnosti signálu. Začátek

a konec signálu nejsou jedinými důležitými parametry. Často jsou signály zaznamenávány po dlouhou dobu spolu s mnoha významnými akcemi. Definice os času a hodnot jsou velmi důležité pro integraci dat (může dojít k přerušení záznamu signálu a architektura systému s tím musí počítat). Je zřejmé, že musíme nalézt reprezentaci signálu ve vztahu k času a ostatním signálům, parametrům a událostem. Ke všem záznamům signálů, segmentaci a analýze je vhodné přistupovat přes jednotné rozhraní, které zároveň umožňuje přidávat nové formáty, nové metody segmentace a analýzy.

V dalším kroku popíšeme základní součásti modelu – charakteristické body, komplexy, analýzu a příznaky (viz obr. 3).



Obrázek 3 — Schéma části modelu definující vztah svodu (signálu) , zpracování, charakteristických bodů, komplexů a analýz

Jak jsme se již zmínili, zpracování signálu je nedílná součást diagnostického procesu. Proto je nutné nějakým způsobem vhodně popsat potřebné atributy.

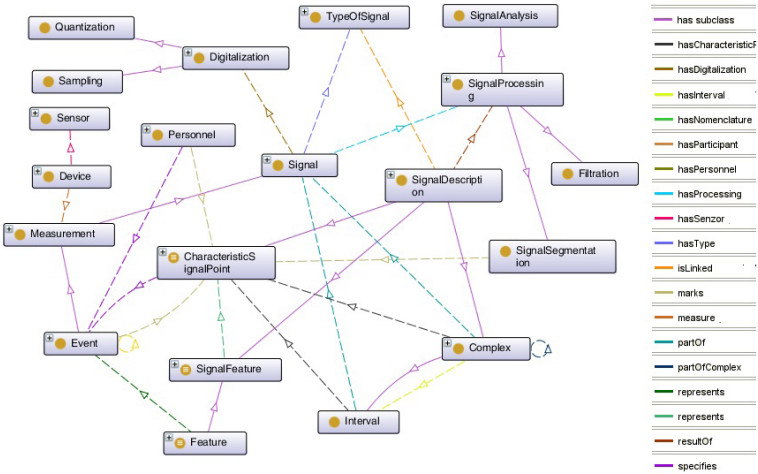
Základním elementem jsou charakteristické body. Můžeme je definovat přímo v signálu. Jedná se pak o body, které přímo souvisejí s vlastnostmi signálu. Typickým příkladem jsou charakteristické body na signálu EKG (P, Q, R, S, T). Druhou možností, jak lze určit charakteristické body, je využití událostí. Události jsou zaznamenávány odděleně. Jejich projekce na signál je definována příslušnými charakteristickými body s daným časem výskytu. Zde je jasné, proč požadujeme časovou synchronizaci. Události mohou být představovány jediným okamžikem nebo intervalem (komplexem). Události jsou obecně vztažené ke třídě charakteristických bodů a také tím definují své třídy komplexů. Může tedy dojít k jistému uzavření kruhu, kdy charakteristické body události definují v signálu komplex a tento komplex svými vlastnostmi definuje událost, která je shodná s původní událostí (jak je zmíněno už výše dochází ke shodě výsledku – událost je reprezentována signálem). Podstatné je

zde říci, že každá událost, pokud v jejím trvání existuje měřený signál, definuje interval (komplex) v tomto signálu. Avšak ne každý komplex je automaticky událostí. K tomu, aby se komplex stal událostí, je zapotřebí znalost, která ho jako událost umožní reprezentovat. Zde vzniká vazba na příznaky, které vyjadřují nějakou specifickou vlastnost signálu a/nebo komplexu.

Komplexy jsou tedy definovány pomocí charakteristických bodů. Základní formou komplexu je interval, který je definován pouze dvěma body (začátkem a koncem). V obecné definici komplexu můžeme použít libovolný počet bodů a komplex je pak určen přesným pořadím bodů nebo v jednodušším případě pouze jejich přítomností v komplexu. Body komplex mohou být definovány jako nutné a možné. Lze pak definovat komplex velmi podrobně a s vyšší mírou možností.

Zpracování/analýza reprezentuje transformace a analýzy, jejichž výstupem jsou nové signály (spektrogramy, wavelety) a automaticky určené charakteristické body pro definici komplexů. Blok zpracování může být použit opakovaně na již odvozené signály a signálové komplexy.

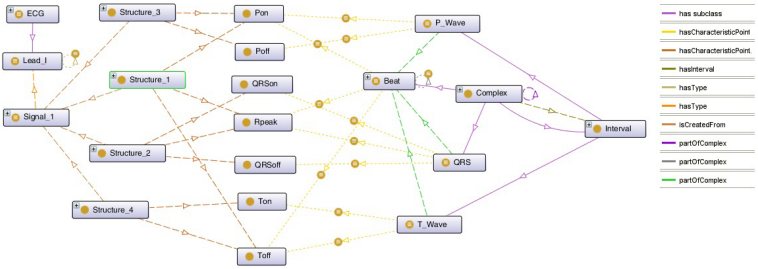
Obr. 4 ukazuje část ontologie reprezentující signály, body a komplexy a shrnuje definice uvedené výše.



Obrázek 4 — Část ontologie reprezentující signály, body a komplexy

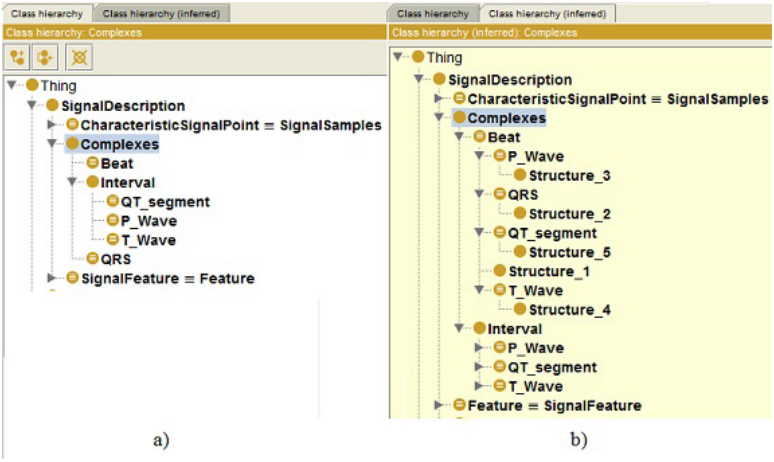
Příklad struktury části ontologie signálů a komplexů pro EKG je zobrazen na Obr. 5. Na levé straně je definována struktura příslušnosti signálu, kde je patrná vazba mezi typem signálu (ECG), typem svodu (Lead I) a samotným signálem. Ten může být dále segmentován do struktur. Na pravé straně jsou zobrazeny třídy interval a komplex, v rámci kterých jsou definovány intervaly Pwave a Twave a komplexy beat a QRS. K intervalům a komplexům jsou přiřazeny

charakteristické body, které obsahují (jsou jimi definovány). Propojení mezi oběma stranami je provedeno právě přes vytvoření instancí tříd příslušných bodů segmentací. V této chvíli je každé struktuře přiřazena třída intervalu nebo komplexu, do které náleží.



Obrázek 5 — Příklad struktury části ontologie signálů a komplexů pro EKG

Na Obr. 6 je zobrazena struktura tříd komplexu a intervalu před spuštěním inference, kdy ještě nejsou k jednotlivým intervalům a komplexům přiřazeny struktury signálu (Obr. 6a). Po provedení inference jsou segmentované části signálu přiřazeny k příslušným intervalům a komplexům a to na základě definice charakteristických bodů, resp. přiřazení jejich instancí k strukturám signálu (Obr. 6b).



Obrázek 6 — Příklad inference (přiřazení) segemntů signálu k příslušným komplexů pro EKG: a) struktura intervalů a komplexů, b) přiřazení segmentů ke komplexům (intervalům)

3. Závěr

V příspěvku jsme prezentovali navrženou obecnou ontologii a datový model, který umožňuje vytvořit sémantický popis signálu. Je nutné zdůraznit, že ontologie a model lze použít na jakýkoliv typ měřeného signálu nezávisle na jeho morfologii, časových a frekvenčních charakteristikách. Navíc signál (nebo jeho části) může být spojen s událostmi, které se objevily během měření. Signál lze pomocí segmentace rozdělit na segmenty a individuální body, které definují reprezentaci v symbolické podobě s přímými vazbami na události, které nastaly při měření. Symbolická reprezentace také umožňuje vytvoření vazby mezi strojovým zpracováním a znalostmi experta, protože informace jsou definovány v kontextu signálu a připojených událostí a navíc jsou čitelné a srozumitelné jak pro počítač, tak pro člověka.

Poděkování

Práce byla podporována projektem MAS Nanoelectronics for Mobile Ambient Assisted Living (AAL) Systems (grant no. 120228) (projekt je částečně financován ENIAC Joint undertaking´s Funding (no. 120228) a MŠMT ČR (č. projektu 7H10019).

Literatura

[1.] Lhotská, L., Burša, M., Huptych, M.: ICT v medicíně a problematika standardů. In Sborník příspěvků MEDSOFT 2011. Praha: Agentura Action M, 2011, s. 172-181. ISSN 1803-8115

[2.] Huptych, M. - Lhotská, L. (supervisor), Multi-layer Data Model. [PhD Thesis]. Prague: CTU FEE, Department of Cybernetics, BIO Laboratory, 2013. 110 p.

[3.] A. Levy, A. Rajaraman and J. Ordille, „Querying Heterogeneous Information Sources Using ource Descriptions," in Proceedings of 22th International Conference on Very Large Data Bases, Mumbai, 1996.

[4.] M. Gagnon, „Ontology-Based Integration of Data Sources," in 10th International Conference on Information Fusion, Quebec, 2007.

[5.] Rubin, D.L., Shah, N.H., & Noy, N.F. (2007). Biomedical ontologies: a functional perspective. Briefings in bioinformatics. 9(1), 75-90

[6.] H. Wache, T. Vögele, U. Visser, H. Struckenschmidt, G. Schuster, H. Neumann and S. Hübner, „Ontology-based Integration of Information – A Survey of Existing Approaches 108-117," Proccedings of IJCAI-01 Workshop: Ontologies and Ingormation Sharing, pp. 108-117, 2001.

[7.] M. Dactu, F. Melgani, A. Piardi and Serpico S., „Multisource Data Classification With Dependence Trees," IEEE Transaction on Geoscience and Remote Sensing, vol. 40, no. 3, 2002.

[8.] C. Chirathamjaree, „A Data Model for Heterogeneous Data Sources," in IEEE International Conference on e-Business Engineering, Xi'an, 2008.

[9.] V. Kashyap and A. Sheth, „Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies," ACM SIGMOD Record, vol. 28, no. 1, pp. 5-12, 1996.

[10.] N. Kannathal N., U. Acharya, E. Ng, S. Krishnan, L. Min and S. Laxminarayan, „Cardiac health diagnosis using data fusion of cardiovascular and haemodynamic signals.

Computer methods and programs in biomedicine, Elsevier," Computer methods and programs in biomedicine, Elsevier, vol. 88, pp. 87-96, 2006.

[11.] L. Thoraval, G. Carrault, J. Schleich, R. Summers, M. van de Velde and J. Diaz, „Data Fusion of Electrophysiological and Haemodynamic Signals for Ventricular Rhythm Tracking," IEEE Enginnering in Medicine and Biology, vol. 16, pp. 48-55, 1997.

[12.] D. Wilson and T. Martinez, „Improved Heterogeneous Distance Functions," Journal of Artificial Intelligence Research, vol. 6, pp. 1-34, 1997.

[13.] E. Fromont, R. Quiniou and M. Cordier, Learning Rules from Multisource Data for Cardiac Monitoring, Berlin Heidelberg: Springer-Verlag, 2005.

[14.] V. Podgorelec, P. Kokol, M. Stiglic, M. Heročko and I. Rozman, „Knowledge Discovery with Classification Rules in a Cardiovascular Dataset," Computer Methods and Programs in Biomedicine, Elsevier, vol. 80, pp. 39-49, 2005.

[15.] L. Busse, P. Orbanz and J. Buhmann, „Cluster Analysis of Heterogeneous Rank Data," in 24th International Conference on Machine Learning, Corvallis, 2007.

[16.] Stanford Center for Biomedical Informatics Research, „The Protégé Project," 2000. [Online]. Dostupné na: http://protege.stanford.edu

Kontakt:

Michal Huptych
ČVUT FEL Prah, katedra kybernetiky
Technická 2, 166 27 Praha 6
tel.: 224357325
fax: 224311081
e-mail: *michalhuptych@seznam.cz*
web: *http: cyber.felk.cvut.cz*

Lenka Lhotská
ČVUT FEL Praha, katedra kybernetiky
Technická 2, 166 27 Praha 6
tel.: 224353933
fax: 224311081
e-mail: *lhotska@fel.cvut.cz*
web: *http: cyber.felk.cvut.cz*

Miroslav Burša
ČVUT FEL Praha, katedra kybernetiky
Technická 2, 166 27 Praha 6
tel.: 224357325
fax: 224311081
e-mail: *bursam@fel.cvut.cz*
web: *http: cyber.felk.cvut.cz*