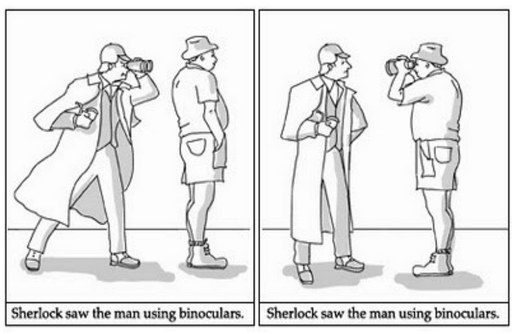
Ontologie a získávání znalostí v biomedicíně

Vendula Papíková
Oddělení biomedicínské informatiky, Ústav informatiky Akademie věd ČR, v. v. i.
INFORUM 2012: 18. konference o profesionálních informačních zdrojích
Praha, 22. – 24. 5. 2012

Úvod

Textový dokument je převažujícím prostředkem sdělování informací mezi vědci. Objem vědecké literatury roste tak rychle, že je obtížné najít a zpracovat obrovské množství publikovaných informací. Nástroje pro efektivní a inteligentní **dobývání informací z elektronických textových databází** se proto vyvíjejí v poslední době ve všech oblastech výzkumu včetně biomedicíny.

Klíčovou úlohu v této souvislosti sehrávají **biomedicínské ontologie**, které jednak pomáhají zvyšovat efektivitu tradičního vyhledávání informací (**text search**, **information retrieval**), jednak jsou klíčem k automatizované analýze textu, dobývání informací z nestrukturovaného textu (**information extraction**, **text-mining**), vytváření nových hypotéz (**hypotheses generation**) a získávání znalostí (**knowledge discovery**) z rozrůstajících se biomedicínských databází.



Obecně lze říci, že **ontologie vyjadřují asociaci slov a jejich významů. Zpřístupňují text počítačům i člověku** a usnadňují tak vývoj počítačových systémů, které se chovají jakoby „rozuměly" problematice dané znalostní domény.

Faktory určující potřebu sémantické standardizace oborových terminologií:

- **Různorodost používaných jazyků:** tzn. víceznačnost termínů používaných v různých kontextech, více termínů užívaných pro tentýž prvek nebo jev, texty psané v cizím jazyce nebo překládané ap.
- **Různorodost prostředí**, v němž texty vznikají: tj. oborově specifické komunikační a publikační zvyklosti, slangové výrazy, zaužívané nestandardní zkratky atp.
- **Potřeba dát širší kontext** informacím formulovaným v rámci úzce specifikovaného zaměření jednotlivých publikací.
- **Potřeba propojit různé informační zdroje** s cílem kombinovat a společně analyzovat různé typy dat (např. genetické informace a textové informace ap.).
- **Potřeba integrovat textové informační zdroje** do jiných typů informačních systémů ve zdravotnictví (např. nemocniční informační systém, laboratorní informační systém, elektronický zdravotní záznam ap.).

Biomedicínské ontologie

Biomedicínské **ontologie lze rozdělit** do tří základních kategorií:

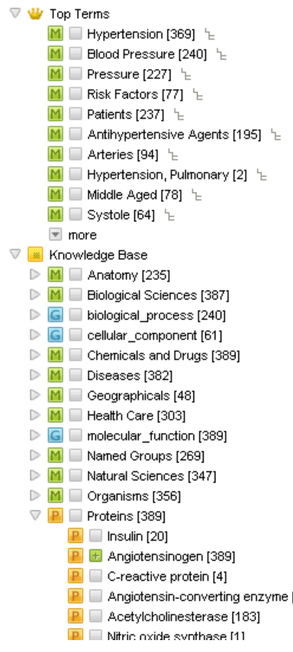
1. **Terminologie (kontrolované slovníky):** poskytují seznam pojmů (konceptů) a textových popisů jejich významu (často organizovaný hierarchicky) a seznam pojmů odpovídajících každému konceptu. Typickými příklady jsou:
   Genová ontologie (GO): termíny vyjadřující molekulární funkce, biologické procesy a buněčné součásti genových produktů (tj. ribonukleových kyselin a bílkovin kódovaných těmito geny). *
   Medical Subject Headings (MeSH): slovník tvořený Národní lékařskou knihovnou USA; poskytuje soubor termínů používaných k popisu témat publikací za účelem indexace biomedicínské literatury.
   NCI Thesaurus: integruje molekulární a klinické informace související s nádorovou tématikou.
   RadLex: kontrolovaný slovník pro obor radiologie; poskytuje terminologii pro techniky, nálezy a onemocnění související se zobrazovacími metodami.

2. **Informační (nebo datové) modely:** poskytují strukturu pro informace příslušné domény (např. data z mikročipů) a popisují vzájemné souvislosti jednotlivých částí modelu. Do této kategorie patří například
   Microarray Gene Expression Object Model (MAGE-OM): spolu s kontrolovanou terminologií používanou k obsazení informačního modelu je nazýván „MGED ontologie". Slouží k popisu minimálních informací o mikročipovém experimentu nezbytných proto, aby čísla tvořící mikročipová data byla smysluplná.

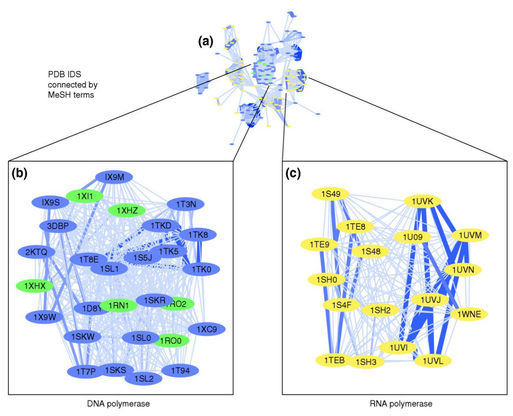
3. **Ontologie ve smyslu vyjádření (reprezentace) znalostí** s definicemi pojmů (konceptů), jejich vlastností (atributů) a vztahů mezi nimi, například:
   Foundational Model of Anatomy (FMA): znalostí zdroj pro anatomii symbolizující třídy a vztahy potřebné pro počítačové modelování struktury lidského těla.

* Využití jednotlivých ontologií může být různé, některé ontologie proto mohou spadat současně do více kategorií (např. GO slouží jako kontrolovaná terminologie, ale také jako informační model).

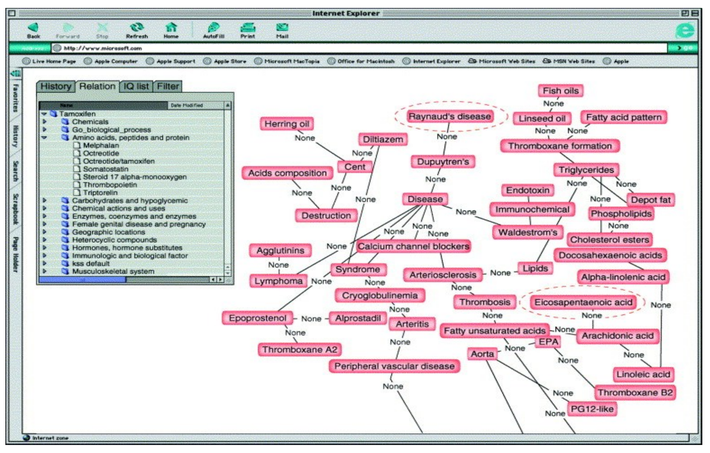
Příklady



GoPubMed umožňuje efektivnější vyhledávání informací v databázi MEDLINE/PubMed prostřednictvím základních biomedicínských znalostí zahrnutých v Genové ontologii (**GO**), v Medical Subject Headings (**MeSH**) a v Universal Protein Resource (**UniProt**). Lineární seznam výsledků **vyhledávání třídí do kategorií na základě biomedicínských konceptů,** jež vyhledané články obsahují, a seskupuje je do kategorií definovaných ontologiemi GO, MeSH a UniProt. Obrázek ilustruje část konceptů nalezených v publikacích, které byly vyhledány na téma hypertenze. Kliknutím na uvedené termíny lze **původní dotaz snadno rozšířit nebo zúžit**.



Příklad grafu vytvořeného pomocí nástroje **PubNet,** který z publikací vyhledaných v databázi MEDLINE/PubMed **extrahuje identifikační kódy 3D modelů bílkovinných struktur** pocházející z databáze Protein Data Bank (PDB), a **propojí je prostřednictvím termínů MeSH.** V takto vzniklé síti představují identifikační kódy bílkovin uzly, kategorie MeSH jsou vyjádřeny jako spojnice mezi těmito uzly. Uvedený obrázek zachycuje (a) graf dotazu „DNA polymerase 2004[dp]" (modré uzly) a „RNA polymerase 2004[dp]" (žluté uzly). **Shluk uzlů téže barvy v těsné blízkosti vyjadřuje strukturální podobnost daných proteinů (b), (c).** Modré uzly spojují termíny MeSH, které souvisí s DNA polymerázou (b), žluté uzly jsou propojeny prostřednictvím termínů MeSH souvisejících s RNA polymerázou (c). Zelené uzly zastupují proteiny vztahující se k DNA i RNA polymeráze současně. Uvedená **síť zpřístupňuje nové objevy a poznatky. Umožňuje studium vztahů mezi vyhledanými bílkovinami** reprezentovanými příslušnými identifikátory. Ty jsou graficky zobrazeny **na základě současných biomedicínských znalostí, které jsou zde reprezentovány pomocí terminologie MeSH.**



Příklad **slovní sítě** ilustrující tzv. **skryté (tj. doposud nepopsané) vazby** (**„hidden links"**) mezi Raynaudovou nemocí (uzel v horní části grafu) a mastnými kyselinami v rybím oleji (zastoupenými např. termíny „fish oils" v horní části grafu nebo „eicosapentaenoic acid" vpravo dole). Spojnice na obrázku vyjadřují vztahy mezi termíny. Spojení označená jako **„none"** ukazují, že mezi danými termíny doposud žádný vztah nebyl popsán. **Stromová struktura** vlevo je tvořena **kategoriemi MeSH** a podle nich roztříděných termínů, které byly extrahovány z vyhledaných dokumentů. Tyto kategorie mohou být použity k **filtrování a další analýze termínů** zobrazených v rámci sítě.

Souhrn

Ontologie odrážejí skutečnost, že propojením jednotlivých faktů jsou tvořeny větší entity, jež **vyjadřují stav poznání v dané oblasti**. Ontologie jsou důležité pro zachycení, vyjádření a organizaci znalostí, jež se stále rychleji objevují v heterogenních a nestrukturovaných textových zdrojích. Jsou proto základem pro **popis funkcí genových produktů, předpověď molekulárních funkcí a buněčné lokalizace doposud nepopsaných genů, určování vztahů mezi genotypem a fenotypem, tvorbu a interpretaci genových sítí a sítí cílových molekul léků atd.**

Poděkování:
zčásti podpořeno projektem 1M06014 MŠMT ČR

Kontakt:
papikova@cs.cas.cz, www.euromise.cz