

■ *Izvorni znanstveni članek*

Andrej Kastrin, Dimitar Hristovski

## Napovedovanje povezav v omrežju deskriptorjev MeSH: primer uporabe za odkrivanje zakonitosti iz literature

**Povzetek.** Odkrivanje potencialnega novega znanja predstavimo kot problem napovedovanja povezav med vozlišči, ki v omrežju niso neposredno povezana: večja kot je podobnost med vozliščema, večje je verjetje, da bo med vozliščema nastopila povezava. Za ilustracijo metode smo zgradili veliko realno omrežje, v katerem so vozlišča predstavljala deskriptorje iz geslovnika MeSH, povezave med vozlišči pa sopojavitev parov deskriptorjev MeSH v zbirki MEDLINE. Za računanje podobnosti med vozlišči smo uporabili število skupnih sosedov, Jaccardov in Adamic/Adarjev koeficient ter koeficient prednostnega povezovanja. Eksperimentalni rezultati kažejo dobro kakovost napovedi.

### Link Prediction in the MeSH Descriptors Network: Application to Literature-based Discovery

**Abstract.** Discovery of potential new knowledge is presented as the problem of link prediction of edges between nodes that are not inherently connected. The greater the similarity between a pair of nodes, the greater the likelihood that the link will be established. To illustrate the proposed approach we built a large-scale real network of co-occurring MeSH descriptors based on the MEDLINE bibliographical database. Link prediction was performed using node similarity computed using number of common neighbors, Jaccard and Adamic/Adar coefficients, and preferential attachment. The experimental results showed good predictive performance.

■ **Infor Med Slov** 2015; 20(1-2): 2-6

---

*Institucije avtorjev / Authors' institutions: Fakulteta za informacijske študije, Novo mesto, Slovenija (AK); Inštitut za biostatistiko in medicinsko informatiko, Medicinska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenia (DH).*

*Kontaktna oseba / Contact person: asist. dr. Andrej Kastrin, Fakulteta za informacijske študije, Ulica takev 3, 8000 Novo mesto. E-pošta / E-mail: andrej.kastrin@guest.arnes.si.*

*Prispelo / Received: 04.09.2015. Sprejeto / Accepted: 11.09.2015.*

## Uvod

Strojni priklic zelenih informacij in njihova umestitev v obstoječo zakladnico znanja predstavljata pomemben raziskovalni problem. Pri tem so nam v pomoč različne tehnologije rudarjenja besedil.<sup>1</sup> Na področju biomedicine se najpogosteje srečamo s štirimi problemskimi nalogami, ki zahtevajo rudarjenje po besedilih: prepoznavna različnih informacij v literaturi, priprava povzetkov dokumentov (angl. *document summarisation*), iskanje odgovorov na znanstvena in strokovna vprašanja (angl. *question-answering*) ter računalniško podprto odkrivanje zakonitosti iz literature (OZL; angl. *literature-based discovery*).

OZL je razmeroma mlado znanstveno področje, ki ponuja zbir različnih metodoloških orodij za samodejno konstrukcijo raziskovalnih hipotez.<sup>2</sup> Glavni cilj OZL je odkrivanje implicitnih, v literaturi še ne opisanih, povezav med znanstvenimi koncepti v obstoječi domeni znanja. Pionir na področju OZL je ameriški fizik Swanson,<sup>3</sup> ki je na osnovi ročne analize literature po naključju odkril povezavo med ribjim oljem in Raynaudjevimi sindromom. (Pri Raynaudjevem sindromu gre za občasna skrčenja manjših žilnih odvodnic, najpogosteje v prstih rok, lahko pa tudi na prstih nog, jeziku in nosu. Motnja v prekrvavitvi traja navadno nekaj minut do nekaj ur.) Kasnejši klinični eksperiment je pokazal, da ribje olje dejansko vpliva na zmanjšanje viskoznosti krvi, zmanjša strjevanje krvnih ploščic ter inhibira odziv žilne stene.<sup>4</sup>

Osnovna zamisel Swansonovega pristopa narekuje obstoj dveh, med seboj nepovezanih, znanstvenih domen. Koncepti znanja v prvi domeni so sicer lahko povezani s koncepti znanja v drugi domeni, vendar so te relacije implicitne (tj. preko tretjih konceptov) in v literaturi še niso eksplicitno opisane. Idejo lahko ilustriramo s tremi teoretičnimi koncepti: X, Y in Z. Za primer vzemimo, da je skupina raziskovalcev ugotovila povezavo med boleznijo X in genom Y. V nadaljevanju privzemimo, da je druga raziskovalna skupina proučevala vpliv zdravila Z na gen Y ter med njima ugotovila vzročni odnos. Z uporabo metodologije OZL poskušamo odkriti implicitno relacijo med konceptoma X in Z preko koncepta Y, kar v našem primeru pomeni, da zdravilo Z lahko vpliva na bolezen X.

Obstoječa metodologija OZL temelji na načelu sopojavnosti (angl. *co-occurrence*) znanstvenih konceptov. V tem smislu obstoječe znanje konstruiramo kot nomološko mrežo (omrežje) konceptov, v kateri povezave med koncepti

predstavljajo njihovo sopojavnost v literaturi. Koncepta A in B sta povezana, če se skupaj pojavita v naslovu, povzetku ali med ključnimi besedami znanstvenega članka. Uporaba sopojavnosti konceptov v OZL temelji na naivni predpostavki, da sta taka koncepta med seboj tudi vsebinsko smiselno povezana.<sup>5</sup> Taka reprezentacija znanja je seveda dinamična, saj v omrežje dodajamo nove koncepte in povezave med njimi.

Raziskovalci so se dolgo ukvarjali z razumevanjem mehanizmov, ki so odgovorni za vzpostavljanje povezav v kompleksnih omrežjih. V zadnjem desetletju je znotraj analize omrežij vzniknilo novo raziskovalno področje, ki se ukvarja z napovedovanjem povezav (angl. *link prediction*). Gre za raziskovalno področje, ki meji tako na klasično analizo omrežij kot na računsko statistiko in strojno učenje. Tipična problemska naloga je konstrukcija seznama povezav, ki se bodo v omrežju pojavile v določeni časovni rezini.<sup>6</sup> Na problem napovedovanja povezav lahko prevedemo tudi proces OZL, tako da na podlagi vzorca obstoječih povezav med koncepti poskušamo napovedati formiranje novih eksplicitnih povezav med koncepti.

Namen prispevka je pokazati, da je tak omenjen pristop k napovedovanju novih povezav med koncepti komplementaren običajnemu procesu OZL. Omrežje povezav smo zgradili in preizkusili na omrežju, ki temelji na geslovníku deskriptorjev MeSH.<sup>7</sup>

## Metoda

### Priprava podatkov

MEDLINE je najbolj obsežna bibliografska zbirka za področje biomedicine. Trenutno obsega okoli 24 milijonov zapisov, ki segajo do konca 19. stoletja. Od sredine štiridesetih let prejšnjega stoletja so zapisi v MEDLINE označeni z deskriptorji MeSH. MeSH je kontroliran geslovník, ki vsebuje biomedicinske izraze (deskriptorje) na različnih nivojih specifičnosti. Distribucija MeSH 2015 vsebuje 27.455 različnih MeSH deskriptorjev. Zapis, ki poroča o mikromrežni analizi DNA v sivi možganski skorji pacientov z bipolarno motnjo, bo npr. vseboval deskriptorje "Bipolar Disorder", "Brains" in "Gene Expression Profiling". Vsak zapis v zbirki MEDLINE vsebuje v povprečju 12 deskriptorjev MeSH. Nekateri so v zapisu označeni kot glavni deskriptorji MeSH, kar pomeni, da označujejo glavno tematiko zapisa. V nadaljnji analizi smo uporabili le glavne deskriptorje MeSH.

V nadaljevanju raziskave smo prebrali celotno distribucijo MEDLINE do vključno leta 2014, ki je vsebovala 21.850.751 zapisov, označenih z deskriptorji MeSH. Distribucija je v XML zapisu, zato smo za nadaljnje potrebe za vsak zapis izločili identifikacijsko številko PMID, pripadajoče deskriptorje MeSH, indikator, ki označuje ali je dani deskriptor MeSH označen kot glavni deskriptor, ter letnico objave zapisa. Nato smo zgradili omrežje deskriptorjev MeSH, v katerem je posamezen deskriptor predstavljal posamezno vozlišče. Povezava med deskriptorjema je bila vzpostavljena, če sta se skupaj pojavila v istem MEDLINEovem zapisu. Omrežje smo sestavili kot neusmerjeno omrežje (relacija med deskriptorjema  $u$  in  $v$  je enakovredna relaciji  $v$  in  $u$ ). Omrežje smo shranili kot seznam povezav.

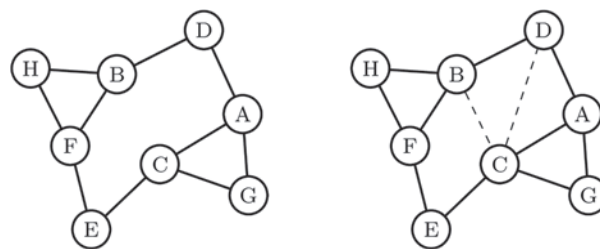
### Eksperimentalni načrt

Omrežje smo predstavili kot graf  $G(V, E)$ , ki ga sestavlja množica vozlišč  $V$ , ki označuje deskriptorje MeSH, ter množica  $E$  neusmerjenih povezav med njimi. Za podrobnejši vpogled v problematiko kompleksnih omrežij priporočamo pregledna članka Newmana<sup>8</sup> ter Boccalettija in sod.<sup>9</sup>

Za reševanje problema napovedovanja povezav moramo razumeti dinamiko pojavljanja povezav med posameznimi pari vozlišč. Nenadzorovani problem napovedovanja povezav lahko formalno predstavimo z naslednjimi koraki. Denimo, da imamo omrežje  $G[t_1, t_2]$ , ki ga sestavljajo vse povezave med vozlišči, vzpostavljene v časovni rezini  $[t_1, t_2]$ . Dalje predpostavimo, da je  $[t_3, t_4]$  časovna rezina, ki sledi intervalu  $[t_1, t_2]$ . Cilj postopka je sestaviti seznam povezav, ki bodo vzpostavljene v časovni rezini  $[t_3, t_4]$ , ne bodo pa prisotne v rezini  $[t_1, t_2]$ . Omrežje  $G[t_1, t_2]$  bomo v nadaljevanju imenovali učno omrežje, omrežje  $G[t_3, t_4]$  pa testno omrežje (slika 1). Za vsak par vozlišč v učnem omrežju lahko izračunamo različne statistike (mere podobnosti), ki odražajo verjetnost pojavitve povezave med vozliščema v testnem omrežju. Vsak tak par predstavlja pozitiven oz. negativen primer, odvisno od tega, ali je v testnem omrežju povezava med vozliščema vzpostavljena ali ne.

Bolj kot sta si vozlišči v paru podobni glede na vzorec povezanosti, večja je verjetnost, da bo med njima vzpostavljena povezava. Formalno za vsako vozlišče v paru  $(u, v)$  izračunamo mero podobnosti  $s(u, v)$ , ki predstavlja verjetje za povezanost obeh vozlišč. Pregled literature razkriva pisano paleto različnih mer podobnosti; uporabili smo metodo skupnih sosedov,

Jaccardov in Adamic/Adarjev koeficient ter koeficient prednostnega povezovanja.



**Slika 1** Učno (levo) in testno (desno) omrežje. Pojav novih povezav smo napovedovali na podlagi topologije učnega omrežja. Učinkovitost napovedovanja smo preverili na podlagi primerjave napovedanih povezav z dejanskimi novimi povezavami v testnem omrežju (črtkane povezave).

Metoda skupnih sosedov (CN) meri število skupnih vozlišč med dvema vozliščema. Za vozlišči  $u$  in  $v$  je CN definirana kot število vozlišč, ki so vozliščema  $u$  in  $v$  skupna. Formalno je mera CN definirana kot:

$$s_{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|.$$

Jaccardov koeficient (JC) je normalizirana različica metode skupnih sosedov. JC izračuna razmerje med številom skupnih sosedov in vsemi sosedi. Med vozliščema  $u$  in  $v$  bo vzpostavljena povezava, če bosta med vsemi svojimi sosedi imeli veliko število skupnih sosedov. Formalno je mera JC definirana kot:

$$s_{JC}(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}.$$

Adamic/Adarjev koeficient (AA) prav tako temelji na številu skupnih sosedov, le da močnejše obteži šibkejše povezave. Formalno je mera AA definirana kot:

$$s_{AA}(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(z)|}.$$

Koeficient prednostnega povezovanja (PA) je definiran kot produkt sosedov vozlišč  $u$  in  $v$ . Vozlišča z višjo stopnjo imajo večjo težo pri vzpostavljanju novih povezav. Formalno je mera PA definirana kot:

$$s_{PA}(u, v) = |\Gamma(u)| \times |\Gamma(v)|.$$

### Ovrednotenje kakovosti napovedovanja

Za oceno kakovosti napovedovanja novih povezav smo uporabili ploščino pod krivuljo ROC (AUC). AUC je posebej uporabna v primeru neuravnoteženih razredov. Algoritem napovedovanja povezav za vsak par vozlišč izračuna verjetje nastopa povezave  $s$ .

Vrednost AUC interpretiramo kot verjetnost, da slučajno izbrana manjkajoča (angl. *missing*) povezava dobi višji dosežek kot slučajno izbrana neobstoječa (angl. *nonexistent*) povezava.<sup>6</sup> Povedano drugače, slučajno izberemo manjkajočo in neobstoječo povezavo ter primerjamo njuna dosežka  $s$ . AUC je formalno definirana kot

$$AUC = \frac{n' + 0.5 \times n''}{n},$$

kjer je  $n$  število neodvisnih primerjav,  $n'$  število primerov, ko je bila vrednost dosežka  $s$  pri manjkajoči povezavi višja kot pri neobstoječi povezavi, ter  $n''$  število primerov, ko je bila vrednost dosežka  $s$  pri manjkajoči povezavi enaka kot pri neobstoječi povezavi.

## Rezultati

Omrežje, nad katerim smo izvajali eksperimente, je imelo  $|V| = 24.401$  vozlišč in  $|E| = 3.464.696$  neusmerjenih povezav. Povprečna stopnja vozlišča je znašala  $c = 284$  povezav, maksimalna stopnja pa  $k_{\max} = 7.761$  povezav. Premer omrežja je znašal  $D = 6$  povezav. Omrežje se je ponašalo z razmeroma kratko povprečno dolžino poti med vsemi pari vozlišč; v povprečju smo iz izbranega vozlišča do kateregakoli drugega vozlišča prišli v  $L = 2,51$  skokih. Koeficient zgoščanja omrežja je znašal  $C = 0,45$ . Zaradi majhnega premera omrežja in razmeroma visokega zgoščanja lahko govorimo o omrežju malega sveta. Največja komponenta (angl. *giant component*) omrežja vsebuje 98 % vseh vozlišč. Več podrobnosti o omrežju je navedenih v predhodnem članku.<sup>10</sup>

Sledijo rezultati eksperimentalnega preverjanja natančnosti napovedovanja povezav. Povzetek mer točnosti napovedovanja je za vse štiri mere prikazav v tabeli 1. Vsaka vrstica v tabeli se nanaša na petletni izsek omrežja, ki je bil uporabljen kot testno omrežje. Ustrezno učno omrežje je bilo sestavljeno na podlagi sopojavnosti deskriptorjev MeSH pred testnim obdobjem (npr. za testno omrežje 1996 – 2000 smo učno omrežje zgradili na podlagi citatov v MEDLINE od začetka obstoja zbirke do konca leta 1995).

Najboljši rezultat AUC dosega mera AA, ki ji sledijo mere CN, JC in PA. Razlike v kakovosti delovanja mer so statistično značilno različne (enosmerna ANOVA:  $F(3, 44) = 5,19$ ;  $p = 0,004$ ). Test naknadnih primerjav (Tukey HSD) je pokazal, da je srednji dosežek za PA ( $M = 0,66$ ;  $SD = 0,07$ ) statistično značilno nižji ( $D = 0,08$ ;  $p = 0,020$ ) kot srednji dosežek za mero CN ( $M = 0,74$ ;  $SD = 0,06$ ). Srednji dosežek za PA je bil prav tako statistično značilno

nižji ( $D = 0,10$ ;  $p = 0,003$ ) kot srednji dosežek za AA ( $M = 0,76$ ;  $SD = 0,06$ ).

**Tabela 1** Dosežki AUC za nenadzorovano učenje.

Testna množica	CN	JC	AA	PA
1951 – 1955	0,77	0,64	0,79	0,74
1956 – 1960	0,84	0,77	0,85	0,79
1961 – 1965	0,74	0,67	0,75	0,71
1966 – 1970	0,69	0,69	0,70	0,64
1971 – 1975	0,69	0,70	0,71	0,62
1976 – 1980	0,63	0,64	0,65	0,54
1981 – 1985	0,68	0,69	0,70	0,57
1986 – 1990	0,75	0,76	0,77	0,64
1991 – 1995	0,75	0,76	0,77	0,64
1996 – 2000	0,76	0,77	0,78	0,65
2001 – 2005	0,80	0,80	0,82	0,70
2006 – 2010	0,82	0,80	0,83	0,73
Aritm. sredina	0,74	0,72	0,76	0,66

Pojasnilo: Vsaka vrstica tabele se nanaša na petletno obdobje, nad katerim smo zgradili testno omrežje. Pripadajoče učno omrežje smo sestavili na podlagi sopojavnosti vozlišč pred testnim obdobjem. Za podrobnosti glej besedilo.

Za ilustracijo predstavljenega pristopa si oglejmo povezavo med shizofrenijo in histaminom. Frekvenca citatov v zbirki MEDLINE z deskriptorjem "Schizophrenia" do leta 1950 znaša 264 citatov, medtem ko znaša frekvenca citatov z deskriptorjem "Histamine" 505 citatov. Če za učno množico vzamemo obdobje od konca 19. stoletja do leta 1950, v tem časovnem oknu ne najdemo nobenega citata, ki bi vseboval oba deskriptorja hkrati. Nenadzorovano učenje, natančneje indeks AA, povezavo med deskriptorjema "Schizophrenia" in "Histamine" v testnem obdobju 1951 – 1955 napove z dosežkom AA = 16,98.

V začetku šestdesetih let prejšnjega stoletja sta Carlson in Lindqvist<sup>11</sup> pokazala, da ima pri psihozi dopamin ključno vlogo. Danes praktično vsa zdravila za zdravljenje psihotičnih motenj temeljijo na dopaminski hipotezi. Raziskovalci pa so nedavno pokazali, da so histaminski mehanizmi dejansko pomembni pri shizofreniji. Histamin namreč služi kot regulator nekaterih drugih nevrottransmiterjev.<sup>12</sup> Pacienti s shizofrenijo imajo običajno nižji nivo receptorjev za histamin H1. Nedavna raziskava je tudi pokazala pozitiven učinek antagonizma histamina H2 pri shizofreniji.<sup>13</sup>

## Razprava

V prispevku smo predstavili in empirično ovrednotili uporabo metodologije napovedovanja povezav v omrežju sopojavnosti deskriptorjev MeSH. Rezultati

kažejo, da je metodologija primerna za proces OZL. Med štirimi preizkušenimi merami se za napovedovanje povezav najbolj obnese Adamic/Adarjev koeficient.

Potrebno je poudariti, da so razlike med preizkušenimi merami podobnosti zelo majhne in niso vse statistično značilne. Razmeroma nizka učinkovitost mere PA lahko nakazuje slabo prileganje omrežja potenčni porazdelitvi, kot smo to že opisali drugje. Naši rezultati so zelo podobni izsledkom, o katerih poročajo Liben-Nowell<sup>14</sup> in Zhou<sup>15</sup> s sodelavci, ki so sistematično analizirali razlike med posameznimi merami podobnosti. Empirične ugotovitve kažejo, da se meri CN in AA praviloma obneseta bolje kot mera PA.

Naša raziskava ima seveda tudi pomanjkljivosti. Analiza je temeljila le na konceptu sopojavnosti deskriptorjev MeSH. Čeprav se v sopojavnost biomedicinskih raziskavah pogosto uporablja, še ne implicira vzročnega odnosa med dvema konceptoma. Prav tako so nekatere sopojavnosti presplošne, da bi bile uporabne (npr. "Humans" – "Disease"). Temu problemu se lahko izognemo z uporabo semantičnih relacij, kot jih npr. ponuja sistem SemRep<sup>16</sup> (npr. "Cognitive therapy TREATS Depressive Symptoms"). Z uporabo sistema SemRep lahko relacije med koncepti bolj natančno operacionaliziramo. Poleg tega je naša analiza zanemarila uteži na povezavah in je obravnavala vse povezave kot enako pomembne. Ustrezna uporaba uteži v problemu napovedovanja povezav še ni rešeno vprašanje. Pričakujemo, da bi njihova ustrezna implementacija pomembno izboljšala rezultate napovedovanja.

Možnosti za nadaljnje delo so številne. Najprej je smiselno implementirati večje število mer podobnosti in preveriti njihovo uspešnost pri napovedovanju povezav. Smiselno bi bilo tudi upoštevati časovno komponento razvoja omrežja. Omrežje je potrebno tudi filtrirati, kar pomeni, da se znebimo redundantnih povezav (npr. povezav med deskriptorji MeSH, ki so preveč splošni). Slednjega se bomo lotili tako, da bomo uporabili UMLS orodje Semantic Network.<sup>17</sup> Vsak biomedicinski koncept lahko namreč mapiramo v metatezaver UMLS Metathesaurus,<sup>18</sup> UMLS Semantic Network pa nam omogoča preveriti smiselnost povezave med dvema izbranimi konceptoma.

## Reference

1. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R: Text-mining solutions for biomedical research: Enabling integrative biology. *Nat Rev Genet* 2012; 13: 829-39.
2. Hristovski D, Rindflesch T, Peterlin B: Using literature-based discovery to identify novel therapeutic approaches. *Cardiovasc Hematol Agents Med Chem* 2013; 11: 14-24.
3. Swanson DR: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986, 30: 7-18.
4. DiGiacomo RA, Kremer JM, Shah DM: Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *Am J Med* 1989; 86: 158-64.
5. Cohen KB, Hunter L: Getting started in text mining. *PLoS Comput Biol* 2008; 4: e20.
6. Lü L, Zhou T: Link prediction in complex networks: a survey. *Phys A Stat Mech its Appl* 2011; 390: 1150-70.
7. Coletti MH, Bleich HL: Medical subject headings used to search the biomedical literature. *J Am Med Informatics Assoc* 2001; 8: 317-23.
8. Newman MEJ: The structure and function of complex networks. *SIAM Rev Soc Ind Appl Math* 2003; 45: 167-256.
9. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U: Complex networks: structure and dynamics. *Phys Rep* 2006; 424: 175-308.
10. Kastrin A, Rindflesch TC, Hristovski D: Large-scale structure of a network of co-occurring MeSH terms: statistical analysis of macroscopic properties. *PLoS One* 2014; 9: e102188.
11. Carlsson A, Lindqvist M: Effect of chlorpromazine or haloperidol on formation of 3methoxytyramine and normetanephrine in mouse brain. *Acta Pharmacol Toxicol (Copenh)* 1963; 20: 140-4.
12. Arrang J-M: Histamine and schizophrenia. *Int Rev Neurobiol* 2007; 78: 247-87.
13. Meskanen K, Ekelund H, Laitinen J, Neuvonen PJ, Haukka J, Panula P, Ekelund J: A randomized clinical trial of histamine 2 receptor antagonism in treatment-resistant schizophrenia. *J Clin Psychopharmacol* 2013; 33: 472-8.
14. Liben-Nowell D, Kleinberg J: The link-prediction problem for social networks. *J Am Soc Inf Sci Technol* 2007; 58: 1019-31.
15. Zhou T, Lü L, Zhang Y-C: Predicting missing links via local information. *Eur Phys J B* 2009; 71: 623-30.
16. Rindflesch TC, Fiszman M: The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 2003; 36: 462-77.
17. National Institutes of Health, Department of Health & Human Services, Lister Hill National Center for Biomedical Communications: The UMLS Semantic Network. <http://semanticnetwork.nlm.nih.gov/> (30.9.2015)
18. National Institutes of Health, Department of Health & Human Services, U.S. National Library of Medicine: UMLS – Metathesaurus. [https://www.nlm.nih.gov/research/umls/knowledge\\_sources/metathesaurus/](https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/) (30.9.2015)