

PRIPOMOČKI ZA ISKANJE PO POLNEM BESEDILU V SLOVENŠČINI

spec. Anton Tomažič, dipl. pravnik

IUS SOFTWARE, d.o.o. Ljubljana

Brnčičeva 31

1231 Ljubljana, SLO

telefon, fax: 373 903

<http://www.ius-software.si/slo/podatki/tone.htm>

tone.tomazic@ius-software.si

POVZETEK

Sodobne informacijske rešitve upravljanja z znanjem (predvsem iskanja po polnem besedilu), uporabe naravnega jezika in prepoznavnja glasu zahtevajo tudi uporabo zahtevnih jezikovnih modulov (virov), katerih pa za slovenski jezik še ni. Ker gre tako za splošen, kot za komercialen interes, bi veljalo združiti sredstva in znanje za izdelavo splošnih in vsem dostopnih modulov (tezavri besed, slovarji, sezname izrazov, fraz, katalogi znanja, itd.). Kot najbolj primerna oblika sodelovanja med državo in gospodarskimi družbami se kaže poseben konzorcij, katerega bi sklicalo Ministrstvo za raziskovanje in razvoj.

ABSTRACT

The latest knowledge management information solutions (mostly full text search and retrieval), natural language and voice recognition require the use of sophisticated language moduls (resources), that do not exist for the Slovenian language yet. Because of the general as well as the commercial interest it would be worth joining the material and human resources to develop general and public domain moduls (thesaurus, glossaries, lists of expressions, phrases, knowledge cataloges, etc.). As the most propriate form of such a cooperation between the administration and commercial companies there seems to be a special consortium, inniciated by the Ministry for research and development.

1 UVOD

Hiter razvoj računalniških, komunikacijskih in (v zadnjih letih predvsem tudi) spletnih tehnologij ima za posledico, da je vse več in več dokumentov v elektronski obliki. To pa seveda ni končni cilj - važne so informacije, katere dokumenti vsebujejo. Obvladovanje teh informacij pa je tisto, kar loči dobre od slabih informacijskih sistemov. Tu se omejujemo na informacije v obliki (nestrukturiranega) polnega besedila, ki so bile sicer v informatiki dolgo zanemarjane, na račun - lažje obvladljivih - relacijskih

(predvsem številčnih) baz podatkov, danes pa že predstavljajo največjo količino informacij.

Obvladovanje nestrukturiranih podatkov se je začelo z zaporednim iskanjem [1], kot so ga omogočali predvsem urejevalniki besedil. Potem je prišlo na vrsto enostavno indeksiranje, pri katerem program vse besede razporedi po abecedi in si za vsako zapomni, kje se nahaja. Sledilo je zahtevno zbiranje z bogatimi iskalnimi orodji, strategijami, geslovniki, naravnim jezikom itd [2]. Hiter razvoj strojne in programske opreme je olajšal delo z velikimi količinami dokumentov v obliki sistemov za upravljanje z dokumenti, komunikacije pa skupinsko delo (groupware). Svetovni splet je prinesel iskalne stroje, agente in potisno tehnologijo. Ogromna podatkovna skladišča so obvladljiva le z interdisciplinarnim pristopom sistemov za upravljanje z znanjem (informatika, medčloveški odnosi, učenje in usposabljanje, motivacija, publiciranje, poslovne analize, lingvistika, itd.).

Nekateri pripomočki za iskanje po polnem besedilu oziroma (sodobneje) za obvladovanje znanja so splošni, nekateri pa izrazito jezikovno vezani. V tem referatu bi radi opozorili predvsem na slednje in na potrebo po tem, da se v Sloveniji reševanja tega problema lotimo koordinirano na nacionalnem nivoju.

2 POTREBE IN MOŽNOSTI

2.1 Iskalni operatorji

Iskalni (danes že klasični) operatorji omogočajo, da se v iskalni izraz lahko vnese dve ali več besed, v različnih medsebojih razmerjih. Osnovni so t.im. Boolovi operatorji.

Poleg osnovnih (IN, ALI, NE, IALI), so zelo uporabni predvsem še ISTO, Z, PRED in BLIZU, s katerim dosežete dokumente, v katerih se drugi pojem nahaja v istem stavku kot prvi. Zaporedje obeh pojmov v stavku ni pomembno. Ta iskalni izraz lahko spremenite (in s tem tudi število zadetkov), tako, da določite največje število

(iskalnih) besed, ki se lahko pojavijo med obema pojmomoma v stavku - na primer BLIZU4.

Informacijsko tehnično gledano iskalni operatorji za slovenske uporabnine (aplikacije) predstavljajo še najmanjši problem. Originalne (ponavadi angleške) izraze je le treba prevesti na dovolj nizkem nivoju (izvorna koda) in uporabniški vmesnik omogoča uporabo s povsem enako funkcionalnostjo, kot original.

Žal lahko ugotavljamo, da (predvsem zaradi komercialnih razlogov) večina programske opreme, delujoče v Sloveniji nima prevedenih niti iskalnih operatorjev. Lepo izjemo na tem področju predstavlja sistem IUS-INFO, ki temelji na programski opremi BRS/Search družbe Dataware Technologies.

2.2 Geslovník

Ob vnosu iskalnega izraza kvalitetni sistemi takoj preverijo tudi, če obstaja poseben geslovník (thesaurus), s

pomočjo katerega bi bilo treba obogatiti iskalni izraz. Geslovník prikazuje odnose med različnimi besedami. Iskanje je lahko bolj učinkovito, če so z iskanim pojmom povezani še drugi sorodni pojmi. Če se na primer išče pojem "zgradba", je dobro, da program najde tudi besedo "stavba". V slovenščini je možno geslovník izrabiti tudi za to, da se zmanjša potreba po uporabi posebnega znaka (zvezdice) za omejevanje besede na njen koren (skloni in druge različne oblike besed). Če so v tezaver n.pr. kot sinonimi vneseni vsi skloni oziroma oblike neke besede (n.pr. vožnja, vožnje, vožnji, vožnjo...), potem bo iskalni rezultat bogatejši (ob vnesenem pojmu "vožnja" bo našel tudi pojem "voženj").

Primer geslovníka, kot ga omogoča sistem BRS Search 6.3 [3]. Odnosi med besedami so definirani s posebnim znakom, sestavljenim iz dveh črk. Osnovni pojem je označen s črkama LT. Pomen ostalih razmerij pa je naslednji:

| Razmerje | Opis | Primer |
|----------------------------------------------|----------------------------------------------|---------------------------------|
| Širši pojmi (BT) | bolj splošni pojem | LT svet BT organ |
| Ožji pojmi (NT) <i>Narrower term</i> | bolj določen oziroma omejen pojem | LT organ NT svet |
| Podobni pojmi (RT) <i>Related Term</i> | dve besedi, ki sta podobni ali povezani | LT svet RT odbor |
| Drugi pojmi (OT) <i>Other Term</i> | polje, prosto za kakršnokoli razmerje | LT zadruga OT 3355XPZ |
| Sinonimi (SRT) <i>Synonym</i> | dve različni besedi, ki pomenita isto | LT arbitražna SRT razsodišče |
| Boljši izraz (USE) <i>Preferred Term</i> | pojem, ki je primernejši od osnovnega | LT šimel USE vzorec |
| Uporabi za (UF) <i>Non-preferred</i> | označba, da je osnovni pojem primernejši | LT vzorec UF šimel |
| Okrajšave (ABR) <i>Abbreviation</i> | okrajšava osnovnega pojma | LT delniška družba ABR d.d. |
| Okrajšava za (AF) <i>Abbreviation for</i> | označba, da obstaja okrajšava | LT d.d. AF delniška družba |
| Istopomensko (LE) <i>Language Equiv.</i> | sprejemljiva različna pisava v drugem jeziku | LT avto LE auto |

Razmerja med besedami iz geslovnika se lahko vidijo pri uporabi iskanja, v obliki drevesne strukture. Tam so prikazani vsi povezani pojmi, ne samo tisti, ki jih nastavite pri nastavitvi geslovnika. Izbira za naslednje iskanje je torej možna med vsemi oblikami besed.

Dober primer geslovnika (za angleški jezik), nameščenega na svetovni splet je izdelek Princeton University – WordNet [4] z medseboj povezanimi samostalniki, glagoli, pridevniki in prislovi. Dosti manj ambiciozen je elektronski terminološki slovar (pojmov, akronimov in okrajšav) EURODICATOM [5].

V Evropski uniji že veliko let razvijajo večjezični (9) in multidisciplinarni geslovník EUROVOC, ki služi predvsem za gesliranje [6] predpisov in drugih dokumentov. Države, ki se pripravljajo za pridružitvev Evropski uniji, bolj ali manj intenzivno razvijajo tudi svoje variante EUROVOCA [7].

Za slovenski jezik splošen geslovník še ne obstaja, bil pa bi zelo koristen tudi za druge programe, ki omogočajo iskanje po polnem besedilu.

Poleg splošnega slovenskega geslovnika pa se bo brez dvoma pojavila tudi potreba po posebnih, tematsko orientiranih geslovnikih (kvalitetnih terminoloških slovarjih). Ti bodo sicer imeli veliko skupnega, po drugi strani pa bo del izrazov imel drugačen pomen ali vsaj drugačno težo (vrstni red pomembnosti). Beseda "organ" ima v pravi dokaj drugačen prvotni pomen, kot n.pr. v medicini. Ob izrazu "šimel" bo povprečen Slovenec najprej pomislil na vrsto konja, pravnik pa na koristen obrazec.

2.3 Podobnostno iskanje

Podobnostno iskanje (proximity search, fuzzy search) omogoča, da se najde tudi pojem, ki ne predstavlja enakega niza znakov, ampak ima enak ali podoben pomen ali pa se enako ali podobno izgovori. Jasno je, da to zahteva dobre spremljajoče geslovnike (tezavre), katerih je za angleški jezik že zelo veliko, za druge jezike pa dosti manj.

2.4 Iskanje v naravnem jeziku

Tu gre za naprednejšo programsko opremo na strani uporabnika. Temu je dovoljeno, da postavlja vprašanja v t.im. "naravnem jeziku" (natural language), n.pr. "Poišči mi besedo nagrada!" ali "Kje lahko dobim podatke o obrestnih merah?". S posebno jezikovno analizo se takšen izraz prevede v niz običajnih in strežniku razumljivih iskalnih izrazov. Tudi tu je prehod od angleških izvornikov na druge jezike dokaj zahteven.

Tudi za slovenščino bo potrebno sestaviti seznam ustreznih besednih zaporedij, izrazov oziroma fraz. Iz praktičnega razloga bi se bilo morda tega lotiti na ta način, da bi enostavno za osnovo vzeli kakšen dober seznam angleških izrazov, jih prevedli in tudi druge slovenske izraze prilagodili (porazdelili) angleškimi. V tem primeru bi modul (vir) gotovo lažje uporabili tudi ob

originalnih (na angleščini temelječih) programskih orodjih za iskanje v naravnem jeziku. Seveda pa bo treba veliko pozornosti in aktivnosti posvetiti tudi problemu skladja: brez vsaj osnovne sintaktične analize si je kvalitetno iskanje težko zamisliti.

2.5 Podatkovna skladišča

Podatkovna skladišča (Document Data Warehouses) so sistemi za obvladovanje velikih količin podatkov v najrazličnejših oblikah. Glavni cilj pri izgradnji takšnih sistemov je v inteligentnem obvladovanju velikih količin (ne)formatiranih podatkov - ob fizičnem povečevanju obsega podatkov (input) - obogatiti ga z dodano vrednostjo in zmanjševati potreben obseg informacij, ki iz njega izhajajo (output). Če so dostikrat ti sistemi nadaljevanje (nadgradnja) sistemov za upravljanje z dokumenti, potem so jim dodane predvsem naslednje funkcionalnosti: povzetki, ocenjevanje pomembnosti (relevancy ranking) poizvedbe v naravnem jeziku (natural language querying).

Uporabniki podatkovnih skladišč (predvsem vodilni v velikih organizacijah) morajo nujno biti opremljeni z učinkovitimi orodji (data mining tools), da lahko optimizirajo obstoječa in odkrijejo nova razmerja med osebjem, strankami in notranjimi postopki. Takšni programi po pravilu podpirajo tudi standardne protokole kot ODBC (Open Database Connectivity) in SQL (Structured Query Language), predvsem pa imajo pripravljene zelo dobre vzorce za zbirna poročila (reports). Za karšnakoli kvalitetna "rudarska" orodja pa so seveda nujni tudi kvalitetni jezikovni moduli (viri).

2.6 Upravljanje z znanjem

Upravljanje z znanjem (knowledge management) je danes najširši pojem za različne oblike iskanja in zbiranja informacij [8].

Predpogoj za učinkovitejšo delitev znanja je bil izveden že z razširitvijo Internet/intranet tehnologije. Toda to ni dovolj. Ne gre le za to, da se DOSEŽEJO informacije, ampak je ustvarjen veliko tesnejši odnos med ljudmi in informacijami. Gre za to, da se ustvari in obvladuje postopek uporabe posameznih (najsodobnejših) tehnoloških komponent v običajnem poslovnem procesu. Potrebno je zajeti, obvladovati in razdeljevati informacije ter jih ponuditi preko Interneta, intraneta, CD-ROMov, DVD-ROMov in omrežij (LAN), predvsem v večjih organizacijah.

Kot primer sistema za upravljanje z znanjem naj opozorimo na Dataware II Knowledge Management Suite. Ta vključuje tudi Web Crawler, ki izvaja iskanja po svetovnem spletu, Agent Server, ki naročnika opozori, kadar je najdena zanj pomembna informacija ter Query Server Middleware, ki z eno poizvedbo išče po različnih sistemih. Sistem je seveda tudi tesno povezan s klasičnimi relacijskimi bazami in kompatibilen s sodobnimi sistemi

za upravljanje z dokumenti (document management systems) [9].

Drugi primer je Oracle ConText Version 2, ki je vgrajen v datotečni strežnik, zato ni treba vzdrževati posebne baze z besedili. Poleg indeksiranja (keyword indexing) je glavna funkcija še v jezikovni analizi (linguistic analysis). API (Application Program Interface) za ConText je set PL/SQL (Program Language/Structured Query Language) paketov in razširitev jezika SQL, tako da se ni potrebno učiti dodatnega orodja pri izdelavi API-ja.

Prvi korak pri izdelavi datotečnega indeksa je naložitev baz v ConText in sicer v štirih možnih oblikah:

1. dokument se naloži direktno v stolpec tabele. Lastnosti stolpca so lahko VARCHAR2, LONG ali LONG RAW. Prvi je za krajša besedila - do 2.000 znakov (Oracle7) oziroma do 4.000 (Oracle8). Za vse obsežnejše dokumente je priporočljiva nastavitvev LONG. Možni formati vnosnih dokumentov so predvsem ASCII, Word, WordPerfect in HTML.
2. dokument se naloži v dve ločeni tabeli, posebej kot glavni (master) in podrobni (detail), kar je uporabno predvsem pri zelo velikih dokumentih (knjige, obsežne pogodbje).
3. naloži se le pot in ime datoteke, sam dokument pa ostane izven baze - v datoteki znotraj operacijskega sistema
4. naloži se le URL, dokument pa ostane v svetovnem spletu

Seveda so dokumenti, ki se nahajajo znotraj baze dosti lažje obvladljivi, kot tisti, ki se nahajajo izven nje, predvsem zato, ker se v prvem primeru samodejno ažurirajo indeksi ob vsaki spremembi. Velikokrat pa je dosti bolj praktično dokumente (posebno velike količine!) shranjevati izven same baze (omejene kapacitete strežnika).

Nova funkcija ConTexta je predvsem indeksiranje po tematiki, glede na vgrajeni katalog znanja: velika hierarhična klasifikacija, razdeljena na šest področij: posel in ekonomija, vlada in vojska, zemljepis, znanost in tehnologija, družba, abstraktne ideje (v tem je gotovo že pomanjljivost... (kaj imata n.pr. skupnega znanstvena fantastika in ljubosumje, ki spadata v isti kategorijo?...)) Običajen izraz kot n.pr. "predstava" ima kot dodatne indekse tudi vse pojme, ki so kategorizirani hierarhično nad tem izrazom (n.pr. "drama", "gledališče", "umetnost", "družba"). Indeksi ConTexta so shranjeni v podatkovne tabele, tako da se lahko po njih izvajajo običajne SQL poizvedbe. Dodatna funkcija je CONTAINS, ki je namenjena določanju relativne teže (tehtnosti) iskalnega izraza. Obstajajo še dodatni logični operatorji, ki na matematičen način kombinirajo indekse z različno težo (najmanjšo, največjo, zbirno, razliko) in relativnim položajem v besedilu (bližje ali bolj oddaljeno v celotnem besedilu). Obstajajo še operatorji za isti koren besede, za podobno napisane in podobno izgovorjene besede.

2.7 Lingvistična naliza

Dobro pripravljene in kvalitetno indeksirane dokumenti (kot n.pr. pri ConTextu [10]) omogočajo tudi dobre jezikovne razdelave (lingvistične analize). Najprej je besedilo razdeljeno tako, da se ugotovijo slovnične značilnosti (gramatične funkcije) besed. Nato je treba ugotoviti, katere besede nosijo ključne tematske enote (theme-bearing elements). Vse takšne enote znotraj dokumenta se nato med seboj primerjajo, tako da se ugotovijo njihove relativno večje ali manjše vrednosti. Končno se izvede klasifikacijski postopek, ki semantične informacije, pridobljene v prejšnjih fazah povzame v nekaj ključnih tem dokumenta. Tu je seveda bistvena zakladnica znanja - ogromen leksikon v obliki številčnih tabel.

Nekateri sistemi ob tem celo opremijo dokument s posebnim povzetkom (summary), ki pa doslej še ni imel takšne uporabne vrednosti, da bi ga uporabniki tudi brali. Zaenkrat služi bolj samemu programu v iskalnem postopku. Pripomniti je treba, da so jezikovne razdelave računalniško zelo zahtevne in skoraj zahtevajo poseben procesor (če tega ni, jih je dostikrat bolje odklopiti).

3 POMANJKLJIVA FUNKCIONALNOST PROGRAMSKE OPREME

Vse več naj sodobnejše programske opreme temelji med drugim tudi na kvalitetnem iskanju po polnem besedilu, za kar obstajajo posebni moduli. Opozoriti je treba na potrebno distanco do programske opreme, ki (ne) predpostavlja, da se iskanje izvaja nad besedilom, ki je v kakšnem drugem jeziku, kot je angleški. Če smo se n.pr. Slovenci pri prvih enostavnih iskanjih še srečevali z drobnimi (a pomembnimi) banalnimi problemi (ko n.pr. programi niso našli nekaterih naših črk), in če smo jih v nedavnem obdobju že več kot rešili (prevedeni so celo notranji iskalni operatorji, kot n.pr. pri IUS-INFU), pa si lahko v bodoče obetamo še veliko večje probleme. Ti so seveda načeloma vsi rešljivi, vendar bodo takšne rešitve večkrat (pre)drage.

Vsi tisti sistemi, ki so zadnji tehnološki dosežki in navidezno najprivlačnejši, delujejo le nad besedili, ki so v angleškem jeziku (ali morda še v kakšnem od večjih svetovnih jezikov). To velja tudi za Oraclov ConText: da bi polno funkcioniral, morajo biti vanj vgrajeni ogromni in kompleksni tezavri besed, slovničnih pravil, leksikoni znanja, itd. - v jeziku PODATKOV. Če tega ni, so funkcionalnosti tako okleščene, da ne ostane kaj dosti več kot dobro indeksirano iskanje po polnem besedilu. Nekaj se bo na tako majhnem jezikovnem "trgu" kot je Slovenija, gotovo dalo narediti, vendar se bo zaostajanje verjetno še povečevalo. Odslej se bo vrhunska programska oprema gotovo skoncentrirala v smeri naravnega jezika, umetne inteligence in - prepoznavanja glasu. Ti moduli pa bodo tako kompleksni in dragi, da jim bo težko slediti za slovenski jezik. Brez teh nujnih

sestavin pa si ni možno predstavljati kvalitetnega upravljanja z znanjem.

Priporočljivo bi bilo, da bi se reševanja tega problema lotili na nacionalni ravni, saj se bodo le tako lahko zbrala zadostna sredstva za pripravo in izdelavo modulov (virov), ki bodo omogočala uporabo najsodobnejših sistemov z elementi naravnega jezika, umetne inteligence in tudi prepoznavanja glasu.

4 MOŽNE REŠITVE

4.1 Na državnem nivoju

Država (in še posebno njena uprava) bi gotovo morala biti zainteresirana, da se tudi za slovenski jezik vzpostavijo potrebni moduli (viri), ki bi omogočali polno izrabo sodobne programske opreme za iskanje po polnem besedilu (obvladovanje znanja), prepoznavanje glasu, naravni jezik in uvajanje elementov umetne inteligence. Takšen interes se bo brez dvoma artikuliral na vladnem nivoju, gotovo pa še posebej s strani posameznih ministrstev, predvsem n.pr. tudi Ministrstva za znanost in tehnologijo. Našla se bodo lahko tudi dokaj velika sredstva, od raziskovalnih, do informacijskih in (mednarodno) projektnih. Že samo sodelovanje dovolj kompetentnih organov in oseb pri tovrstnih aktivnostih bo zagotavljalo stabilnost, resnost in tudi koristno medijsko podporo.

Slabe strani državnega (administrativnega) pristopa pa so seveda v tem, da država sama po eni strani ne čuti dovolj dejanskih potreb, po drugi strani pa sta ogroženi učinkovitost in tempo izdelovanja konkretnih modulov (virov), še bolj pa seveda njihovega vzdrževanja.

4.2 Na komercialnem nivoju

Kot zainteresirani komercialni subjekti se kažejo predvsem velike družbe, ki bodisi prodajajo programsko opremo, katera potrebuje (specifično slovenske) module (vire), ali pa so ponudniki informacijskih baz, ki bi bile z ustreznimi moduli bolje izkoriščene. Primeri: Microsoft, IBM, Oracle, Atlantis, IUS-INFO...

Obstaja seveda možnost, da bi posamezna takšna družba (ali več skupaj) kar sama izdelala ustrezne module (vire), vendar je to težko pričakovati. Po eni strani bi bil namreč takšen projekt (pre)drag, po drugi strani, pa se vsaka družba zaveda, da bo svoje delo težko ustrezno avtorskopravno zavarovala in da se bodo lahko kasnejši interese enostavno okoristili z njenimi vlaganji. Tisti, ki bo n.pr. naredil dober geslovník sinonimov, lahko pričakuje, da ga bodo tisti, ki se bodo tega lotili za lastne potrebe, gotovo uporabili za primerjavo, ko bodo izdelovali svojega (s tem pa jim bo prihranjeno zelo veliko dela in stroškov).

Seveda pa bi izdelava jezikovnih modulov (virov) samo s strani posamezne družbe tudi po vsej verjetnosti ne predstavljala optimalne rešitve s splošnega slovenskega vidika. Struktura takšnega modula bi bila gotovo izrazito

prilagojena samo (programskim) pogojem te družbe in zato kakšne splošne uporabnosti ni pričakovati.

4.3 Konzorcij

Iz zgoraj navedenih razlogov se kot optimalna rešitev ponuja le konzorcijalni pristop, pri katerem bi sodelovali vsi (najbolj) zainteresirani, nekateri (državni organi) zastopajoč bolj splošne interese, drugi (komercialne družbe) pa bolj svoje specifične. Prisotnost prvih bi imela za posledico to, da bi bili moduli planirani, zasnovani in izdelani dovolj splošno in objektivno ter avtorskopravno prosti, prisotnost drugih pa, da bi bila izvedba hitrejša in cenejša.

Predstavljamo si, da bi konzorcij združeval kakšnih 20 subjektov, približno polovico iz javnega (vključno z raziskovalnimi, izobraževalnimi oziroma univerzitetnimi organizacijami) in polovico iz komercialnega sektorja. Koristno bi bilo, da bi na prvi sestanek zainteresirane povabil dovolj kompetenten državni organ (n.pr. Ministrstvo za znanost in tehnologijo), kasneje pa bi se udeleženci konzorcija sami dogovorili o načinu vodenja in dela. Nekaj možnih osnovnih izhodišč bi bilo lahko naslednjih:

- prostovoljno sodelovanje; vsak udeleženec nosi svoje stroške sodelovanja
- odločitve se sprejemajo s konsenzom
- sklicevanje in vodenje sestankov prehaja izmenično za določeno časovno obdobje od udeleženca do udeleženca (čemur pa se lahko vsakdo odpove)
- po možnosti se čimveč aktivnosti izvede tako, da se na podlagi skupaj sprejetih planov delo in stroški porazdele med udeležence, tako da vsakdo nosi svoj del (v skupaj dogovorjenem odstotku)
- za morebitno skupno nastopanje v pravnem prometu (n.pr. zaradi prijav za financiranje z javnimi sredstvi) in za morebitne medsebojne poslovne odnose se sklepajo ustrezni dvostranski ali večstranski pravni akti
- rezultati skupnih aktivnosti se dajo v "public domain", hkrati pa jih lahko udeleženci uporabljajo (in nadalje razvijajo) v okviru svojih komercialnih aktivnosti

5 REFERENCE

[1] Razlika med iskanjem in zbiranjem je v tem, da pri iskanju pred izvedbo ukaza ni znano, kje se iskani pojem nahaja, pri zbiranju pa je to že znano, le da je treba dobiti pravo informacijo o tem.

[2] Primeri kvalitetnega iskanja: <http://ius-info.ius-software.si/iusinfo/tour/primeri.htm>

[3] Dataware Technologies, Inc.:

<http://www.dataware.com>

[4] Spletni naslov za WordNet:

<http://www.cogsci.princeton.edu/~wn/>

- [5] EURODICATOM: <http://www2.echo.lu/echo/databases/eurodicautom/en/eu92home.html>
- [6] Primer: CELEX: <http://ius-info.ius-software.si/euroius/60.htm>
- [7] Primer Češke: <http://www.psp.cz/kps/knih/eurovoc/>
- [8] Upravljanje z znanjem: <http://ius-info.ius-software.si/znanje/>
- [9] Getting the most from data; By Christy Walker and Jim Kerstetter, PC Week Online, 02.23.98
- [10] Unlock the Hidden Value in Documents; by Dan Sullivan, Databased Web Advisor, February 1998