UDK 025.43:004.774.2

Bibliotekų žodynų pritaikymas semantiniam saitynui ir atvirkščiai: abipusiškai naudinga kelionė ten ir atgal

Bernard VATANT
Mondeca, Paryžius, Prancūzija, el. p. bernard.vatant@mondeca.com

Žodynų vaidmuo ypatingai svarbus labai vėluojančioje saityno ir bibliotekų paveldo sinergijoje. Semantiniame saityne turėtų būti tobulinami esami žodynai, o ne kuriami nauji, tačiau bibliotekų žodynų specifika leidžia daugiau ar mažiau pritaikyti juos semantiniam saitynui. Remdamiesi preliminariais TelPlus projekto rezultatais, siūlome būtinų permainų gaires, kurios padėtų padaryti žodynus tinkamus naudoti ir veiksmingus semantinio saityno srityje, įvertinti stambiųjų bibliotekų jau priimtus sprendimus dėl standartus ir sėkmingą praktiką atitinkančių žodynų publikavimo bei apžvelgti, kaip semantinio saityno priemonės galėtų palengvinti šių žodynų valdymą.

Reikšminiai žodžiai: semantinis saitynas; kontroliuojami žodynai.

Įvadas

Semantinis saitynas dažnai yra vadinamas didžiausia pasaulio biblioteka, tačiau į saityno paieškos technologijas ir bibliotekų klasifikacijos sistemas ilgą laiką buvo žiūrima kaip į konkuruojančias žinių organizavimo versle arba kaip orientuotas į atskiras vartotojų nišas. Saityno istorija, palyginti su šimtmečius siekiančia bibliotekų patirtimi, yra gana trumpa (mažiau negu dvidešimt metų), be to, vienas po kito sekė saityno sumanymų sėkmės ir žlugimai, neįtikėtinai sparčiai keičiasi informacinių technologijų gyvavimo ciklai todėl bibliotekininkų požiūryje į šį reiškinį susipina panieka ir susižavėjimas. Tačiau pamažu susiformavo nekintamas suvokimas, kad bibliotekose naudojamų „senų“ sistemų ir „naujų“ saityno technologijų sietis ir sinergija būtų sveikintina ir galiausiai galėtų būti labai naudinga abiem pusėms. Be įspūdingų techninių pasiekimų integruojant ir kataloguojant duomenis (pvz., *WorldCat* [1]), didžiulė semantinio saityno kalbų ir technologijų sankaupa [2] teikia vis daugiau naujų galimybių bibliotekų paveldėtų duomenų vertės atstatymui, naujų bibliotekininkystės mokslo ištobulintų klasifikavimo metodų perspektyvų plėtrai, naujoms paveldėtų duomenų apdorojimo priemonėms.

Be kitų semantinės konvergencijos aspektų, šiame straipsnyje daugiausia dėmesio bus skiriama ypatingam kontroliuojamų žodynų vaidmeniui. Prisiminsime, kaip žodynus paveikė bendra saityno ir informacijos sistemų plėtra; paaiškinsime, kodėl semantiniame saityne turi būti tobulinami esami žodynai, o ne kuriami nauji; pabandysime nustatyti, kokios šių žodynų savybės daro juos daugiau ar mažiau pritaikomus naudoti semantiniame saityne ir pasiūlysime gaires, kurios padėtų juos veiksmingiau taikyti šioje naujoje aplinkoje.

Pateikiamos problemos, praktiniai pavyzdžiai ir dalijamasi įgyta patirtimi, remiantis veikla dalyvaujant TelPlus [3] projekte, kurio metu Prancūzijos nacionalinė biblioteka surengė semantinio saityno kompanijų konsorciumą, kurio tikslas – lyginamoji semantinių anotacijų ir paieškos priemonių grandinės, apimančios išsamius bibliotekų žodynus (pvz., RAMEAU ir LCSH), bei šių žodynų sankirtų analizė.

Taip pat nurodome semantinio saityno priemonių bei programinės įrangos naudą tvarkant žodynus, remdamiesi patirtimi dirbant su EUROVOC [5] ir kitais žodynais, tvarkomais ir skelbiamais Europos Sąjungos leidinių biuro.

Straipsnio pabaigoje apžvelgiama: dabartinės pastangos visuotiniu mastu įtvirtinant šią konvergenciją, vadovaujantis Kongreso bibliotekos *Authorities and Vocabularies* [4] vaidmuo, įvairūs Europos bendrijos kontekste bendrai įgyvendinami projektai ir *W3C Library Linked Data Incubator Group Initiative*.

Žodynai saityne, arba meilės ir neapykantos istorija

Bibliotekininkai šimtmečius rengė kontroliuojamus ir struktūruotus žodynus, tokius kaip klasifikacijos ir tezau-

rai, teikiančius galimybę *žmogaus atliekamam indeksavimui* ir paieškai, kai tuo tarpu paiešką saityne vis labiau padėdavo atlikti automatinės priemonės, naudojančios *statistinius algoritmus*, grindžiamus kryptingąja visatekste paieška. Paieškos priemonės savo užduotis atlieka ir be struktūruotų žodynų, ir paieškos technologijų valdytojai jau seniai palaiko idėją, kad automatinis indeksavimas ir paieška gali pasiekti aukštą efektyvumo lygį, grindžiamą vis sumanesniais algoritmais, o tradicines bibliotekų klasifikavimo priemones ir metodus skelbia esančius daugiau ar mažiau pasenusiais, bent jau saityno kontekste. Antra vertus, bibliotekininkai ir žinių organizavimo specialistai atkakliai laikosi nuomonės, kad paieškos priemonės tinkamai nesusitvarko su terminų dviprasmiškumu ir kad išmaniojoje paieškoje turėtų derėti struktūruotos žinios ir į struktūruotus žodynus įterpiama žmogaus sukurta vertė.

Paieškos sistemų „aukso amžius"

Anksčiau aprašytos diskusijos buvo gana populiarios pirmaisiais saityno gyvavimo dešimtmečiais, kai paieškos sistemų efektyvumo lygis dar buvo žemas [6]. Tačiau įspūdingas *Google* iškilimas ir sėkmė per tolesnį dešimtmetį pasitarnavo kaip lemiamas argumentas automatinio indeksavimo ir paieškos naudai. Remdamiesi šia sėkme, kai kas netgi ėmė teigti, kad kontroliuojami žodynai – atgyvenęs dalykas, turintis užleisti vietą paieškos technologijoms. Žodynų sunykimas reikštų ir jų metaduomenų bei klasifikavimo apskritai, o galiausiai ir bibliotekininkystės mokslo sunykimą.

Žmonės tai atlieka nė kiek ne geriau

Kita vertus, bandymų sisteminti saityną naudojant didelės apimties pagal bibliotekų pavyzdį sudarytas ir pasaulinei bibliotekai pritaikytas klasifikacijos sistemas būta jau pačiais pirmaisiais saityno gyvavimo metais. *Yahoo!* katalogas, kuriuo pasekė bendrai su *Yahoo!* parengta alternatyva *Open Directory* (pastarasis buvo nukopijuotas šimtų kitų iniciatyvų, pvz., *Google Directory*) gana greitai susidorojo su rimtomis išplečiamumo problemomis, atsižvelgiant į beprecedentę saityno išteklių aprėptį, įvairovę ir visišką nestabilumą. Nepaisant drąsios frazės „Žmonės tai atlieka geriau" [7], tokie bandymai galiausiai patyrė pralaimėjimą, jeigu palygintume jų priežiūros kaštus su automatinio indeksavimo kaštais, apskritai diskredituodami visuotines saityno klasifikavimo priemones ir tarnybas paieškos sistemų algoritmų naudai. Tačiau iki šiol puikiai gyvuoja ir veikia daug specializuotų katalogų.

Netiesioginė galima saityno katalogų paradigmos nesėkmės žala gali būti ir tai, kad vidutiniam galutiniam vartotojui gali susiformuoti nuojauta, jog struktūruoti žodynai – tai priemonės, paveldėtos iš tamsių bibliotekinio klasifikavimo amžių prieš atsirandant *Google*. Kita vertus, galutinis vartotojas nesuvokia, kuo skiriasi tezauras, dalykinių rubrikų schema, klasifikacija, rodyklė, turinio sąrašas ir taksonomija... Jam atrodo, kad bet koks žodynas, padeda jam rūšiuoti ir rasti tai, ko jam reikia. Tarkime, Vikipedijos vartotojui pateikiami tiesiog natūraliai besidauginantys įvairūs skirtingi „sąrašai" ir „kategorijos", kurių nevienija joks bendras sumanymas, neapibrėžia jokios aiškios gairės ir kurios parengtos žmonių, dažnai neturinčių pakankamų klasifikavimo įgūdžių [8].

Apskritai vidutinio vartotojo patirtis dirbant su saityno žodynais rodo, kad jie netvarkingi kaip kartais ir pats saitynas, su bendrame chaose vietomis pasitaikančiomis sutvarkytomis sritimis. Užuot padėdami vartotojui aiškiai ir greitai atrinkti ir rasti informaciją, saityno žodynai dažnai tiesiog įneša papildomos medžiagos į visuotinę žinių mišrainę [9], geriausiu atveju paskatindami sėkmingą vartotojo įžvalgą.

Taksonomijų „aukso amžius"

Dabar, esant uždaroms organizacijų informacijos sistemoms, vis labiau pripažįstama struktūruotų ir kontroliuojamų žodynų svarba, o naudojimasis tokiais žodynais labai dažnai sustiprina daugelio kompanijų verslo modelį. Verslo srityje kontroliuojami žodynai paprastai vadinami „taksonomijomis". *The Accidental Taxonomist* [10] įžangoje Heather Hedden rašo: „Šiuolaikiniame informacijos valdyme „taksonomijos" terminas vartojamas siaurąja prasme, t. y. hierarchinei klasifikacijai ar skirstymui kategorijomis reikšti, ir plačiąja prasme, kuri susijusi su bet kokiais žinijos sąvokų organizavimo būdais. Kai kurie specialistai šio termino nemėgsta, tvirtindami, kad labai dažnai jis yra dviprasmiškas arba neteisingai vartojamas. Tačiau terminas tapo pakankamai populiarus, ir panašu, kad alternatyvos jam nėra."

Todėl, patinka tai jums ar ne, žodynas, jeigu tik jis publikuotas, ir kad ir kokia būtų jo originali struktūra bei specifinė paskirtis, populiariai bus žinomas kaip „taksonomija" [11].

Semantinės taikos link

Duomenų atvėrimas ir semantinio suderinamumo poreikis

Pastaraisiais metais atsirado nauja tendencija: informacijos sistemų atvėrimas bendriausia prasme. Plačiai pripažįstama, kad visiškai nerentabilu apdoroti duomenis ir juos dubliuoti uždarose saugyklose bei duomenų bazėse, kai daugelis duomenų galėtų būti viešai prieinami „debesų" serveriuose ir naudojami pagal poreikį. Vis labiau

nyksta skirtumai tarp organizacijų informacijos sistemų ir atvirojo saityno. SaaS (*Software as a Service*) principas – *programinė įranga kaip paslauga* – jau atgyveno, jo vietą užėmė DaaS (*Data as a Service*) – *duomenys kaip paslauga*. Sistemoms keičiantis duomenimis, neabejotinai svarbu ne tik taikyti vienodus duomenų formatus (pvz., XML), bet ir susitarti dėl šių duomenų semantikos. Į pirmąjį planą iškilo *semantinio suderinamumo* problema, todėl tapo aišku, kad būtini bendri informaciniai žodynai, kuriems būdinga formaliai išreikšta semantika.

Semantikos pagrindai: dviprasmiškumas ir bendroji sietis

Semantinio saityno duomenų aibė dažnai suvokiama kaip dirbtinio intelekto technologija, kurios pagrindinis aspektas yra formali ko nors pateiktis aprašo logika paremtomis kalbomis, pvz., OWL. Tai neabejotinai yra labiausiai pastebima aibės dalis, kurioje taikomi sudėtingi ribojimai ir taisyklės, padedančios pašalinti nesuderinamumą ir gauti naujų žinių. Tačiau tam (ir kas iš tikrųjų yra daug svarbiau) būtinas semantinio saityno kalbų ir priemonių funkcionalumas, susijęs su dviprasmiškumo šalinimu ir bendrosios sieties nustatymu. Trumpiau tariant, šios kalbos ir priemonės leidžia aiškiai išreikšti, ar kas nors su tais pačiais arba skirtingais vardais yra tas pats, ar ne. Pavyzdžiui, ar dviejų kokių nors dokumentų tema yra ta pati, ar jų autorius yra tas pats, ar kokie nors du ištekliai randasi ten pat ir pan. Šiuo atžvilgiu semantinio saityno principai visada buvo artimesni bibliotekininkystės mokslui negu vadinamoji *Web 1.0* infrastruktūra. *Web 1.0* daugiausia dėmesio skiriama prieigai prie informacijos išteklių per URL ir netipinėms hipertekstinėms nuorodoms, o semantiniame saityne pirmiausia siekiama *identifikuoti, su kuo susijęs vienas ar kitas išteklius*.

Ko nors tapatumo nustatymas, leidžiantis pašalinti dviprasmiškumą arba nustatyti bendrąją sietį, semantiniame saityne užtikrinamas universaliųjų identifikatorių (URI), kurie skiriasi nuo bibliotekų žodynuose vartojamų natūralios kalbos terminų. Vis dėlto semantinio saityno principai artimesni bibliotekininkystės mokslui negu paieškos priemonių logikai. Abiejuose universumuose informacija kaupiama į sąvokas orientuotu principu, net jeigu sąvokos semantiniame saityne pateikiamos vartojant URI, o bibliotekų žodynuose – natūralios kalbos terminus. Tam, kad abu universumai priartėtų vienas prie kito ir funkcionuotų kaip vienuma, reikalinga kalba, bibliotekų žodynus susiejanti su saityno identifikatoriais (URI).

SKOS: žodynas, pralenkiantis visus kitus?

SKOS (Paprasta žinių organizavimo sistema) [12] kalbos plėtra tęsiasi nuo 2005 m. vykstant vaisingam bibliotekų ir semantinio saityno bendruomenių dialogui, kuriuo siekiama suteikti galimybę nesudėtingai perkelti paveldėtus žodynus. Laikantis W3C rekomendacijos, nuo 2009 m. SKOS naudojama perkeliant tokius paveldėtus žodynus, kaip RAMEAU [13], LCSH [14], AGROVOC [15], EUROVOC ir daugelį kitų. Panašu, kad artimiausiais mėnesiais ir metais šiuo pavyzdžiu paseks ir daugelis žinomų žodynų rengėjų.

SKOS nėra sukurta specialiai bibliotekų žodynams, o skirta taikyti įvairiausioms žinių organizavimo sistemoms ir apima „taksonomijas“ plačiausia anksčiau apibūdinta prasme. Į SKOS labai lengvai išverčiami standartiniai struktūruoti žodynai, pvz., tezaurai, tačiau kai kuriems sudėtingesnių savybių žodynams SKOS raiškos galimybes būtina praplėsti.

Be to, jeigu žodynas išverstas į SKOS formatą, tai dar nereiškia, kad jį bus galima veiksmingai naudoti semantiniame saityne. Kitame skirsnyje panagrinėsime kai kuriuos praktinius aspektus ir pamėginsime nustatyti, kas daro žodyną tinkamą naudoti bei kaip jį tobulinti, siekiant padaryti veiksmingesnį.

Bibliotekų žodynų pritaikymas semantiniam saitynui

Žodynų integravimas semantinio saityno priemonėms, tokioms kaip anotavimo, paieškos ir naršymo, semantinės paieškos išplėtos priemonės, neabejotinai sukuria pridėtinę vertę. Tačiau panašu, kad šia prasme kai kurie žodynai veiksmingesni už kitus, o tie, kurie gerai pritaikyti naudotis žmogui, nebūtinai labiausiai tinkami mašinoms. Mašinoms nebūdingas subtilus žmogiškas supratingumas, joms būtinos aiškios apibrėžtys ir dviprasmiškumo šalinimo taisyklės, ypač jei jos veikia atviroje aplinkoje, kur būtinas platus funkcinis suderinamumas. Mašinų reikalavimai paprasti. Joms būtina aiškiai išreikšta semantika, o sintaksę jos visur ir visada interpretuoja vienodai. Žodyno pritaikymas semantiniam saitynui turės ir šalutinį poveikį, kuris pasireikš pirmiausia ir bus palankus – paaiškės, kad semantiniame saityne žodynas ne visada reikš tai, ką jis reiškia jo autorių ir vartotojų galvose.

Mūsų kompanija dabar apžvelgia kai kurias problemas ir patirtį, įgytą lyginamosios analizės, naudojant Prancūzijos nacionalinės bibliotekos dalykines rubrikas (RAMEAU) ir Kongreso bibliotekos dalykines rubrikas (LCSH), metu TelPlus projekto ribose. Išryškinami tie jų aspektai, kuriuos neabejotinai reikia tobulinti siekiant, kad žodynai būtų iš tikrųjų tinkami veiksmingai naudoti. Jie pateikiami sėkmingos patirties taisyklių forma; kai kurios iš jų tinka bet kokių duomenų pritaikymui semantinėms priemonėms, kitos – tik bibliotekų žodynams.

1. Apibrėžkite ir aprašykite nedviprasmiškas sąvokas

Pagrindinės prielaidos, kuriomis vadovaujamasi semantiniame saityne apibrėžiant sąvokas, lygiai kaip ir bet kurį URI identifikuojamą „daiktą“, yra paprastos, tačiau griežtos.

1. Semantika (URI reikšmė) nepriklauso nuo konteksto. Kad ir kur URI būtų, jis reiškia (žymi) tą patį dalyką. URI yra universalieji unikalūs vardai.

2. Šią semantiką perteikia ir sąvokos dviprasmiškumą pašalina formalizuotas sąvokos aprašas, gaunamas pasitelkiant URI ir saityno protokolą (dažniausiai HTTP).

Kitais žodžiais tariant, formalizuotas sąvokos aprašas turi būti pakankamai tikslus, kad būtų galima aiškiai atskirti sąvoką nuo bet kurios kitos sąvokos (arba sutapatinti su bet kuria kita sąvoka), apibrėžiama bet kurio kito saityno URI. Tai gana griežtas reikalavimas, tačiau kiekvieno potencialaus saityno vartotojo uždavinys yra tokiu būdu patikrinti savo skelbiamą žodyną.

Būtina sąlyga yra unikalių identifikatorių politikos buvimas, tačiau įprasti bibliotekų žodynai dažnai grindžiami vardų identifikavimu pagal kontekstą. Lemiamą reikšmę turi sąvokų ir jas reiškiančių terminų vartojimas, todėl identifikatorius turi būti tokia notacija, kuri nesusijusi su jokiomis etiketėmis. Jeigu žodyno valdymo sistema sąvokoms naudoja vidinius unikalius raktus, pasiteisina praktika kurti URI tų raktų pagrindu, kaip daroma su toliau pateikiamais LCSH URI.

2. Formalizuokite savo žodyno sintaksės semantiką

Žodynus naudojant bibliotekų kontekste, daugeliu atvejų dviprasmiškumą pašalinti gali ir žmogus vartotojas. Natūralia kalba išreikštas terminas gali būti dviprasmiškas, o žodyno kontekste – ne. Anglų (taip pat ir prancūzų) kalbos terminas „table“ dviprasmiškas, tačiau jo dviprasmiškumas sumažėja, jį apibrėžus kaip „home furniture“ (baldas) arba „data base“ (duomenų bazė). Semantinio saityno priemonėms šie du terminai „table (home furniture)“ ir „table (data base)“ turi būti identifikuojami kaip skirtingos sąvokos, skirtingų URI ir formaliai susieti su atitinkamais kontekstais. Šie kontekstai savo ruožtu turi būti formaliai identifikuojami ir skelbiami kaip platesnė arba siauresnė sąvoka arba sąvokų schema (priklausomai nuo smulkumo lygmens). Nepakanka paprasčiausiai nurodyti kontekstą vartojant patikslinimą, kaip tezaurų praktikoje.

Panašiai kaip ir patikslinimų atveju, numanomą semantiką, kurią būtina formalizuoti, slepia ir specifinės sintaksinės struktūros. LCSH ir RAMEAU taikomas MARC formatas apima tokias sintaksines struktūras, kaip „Actors- -Psychology“ (veikėjai- -psichologija) (http://id.loc.gov/authorities/sh85000748). Nors žmogui vartotojui toks santykis ir atrodo savaime suprantamas, formalus šios sąvokos aprašas nenukreipia nei į „veikėjus“ (http://id.loc.gov/authorities/sh85000744), nei į „psichologiją“ (http://id.loc.gov/authorities/sh85108459). Kad šį santykį būtų galima pritaikyti semantinio saityno programinėms priemonėms, jį būtina formalizuoti.

Žodynų rengėjui tokie pavyzdžiai naudingi tuo, kad jie verčia susimąstyti apie tokios sintaksės pagrindą sudarančią semantiką, dėl kurios jam niekada nekildavo klausimų. Siekiant tapačias sintaksines konstrukcijas automatiškai perkelti į semantinio saityno formatą, jos turi būti vienodai interpretuojamos ir apdorojamos. Tačiau pasirodo, kad nėra taip paprasta anksčiau minėtų struktūrų skirtukui „- -“ rasti bendrą semantiką, kuri peržengtų bendros ir neapibrėžtos „skos:related“ savybės ribas ir galiotų visame išsamiame žodyne, tokiame kaip RAMEAU ar LCSH. Ne visada naudingos ir žodynų tvarkytojams skirtos natūralia kalba parašytos rekomendacijos, paaiškinančios, kaip teisingai vartoti kiekvieną sintaksinę konstrukciją, nes žodynų tvarkymo sistemos paprastai nesugeba vadovautis tokiomis rekomendacijomis, tuo labiau įvertinti neteisingo vartojimo pasekmes.

3. Tikrinkite faktinį hierarchijų tranzityvumą

Savybėms „skos:broader“ arba „skos:narrower“ pritaikomos žodynų struktūros semantinių priemonių paprastai bus naudojamos praplėsti paiešką iki bendresnių arba susiaurinti iki konkretesnių sąvokų, siekiant sumažinti paieškos triukšmą arba, atvirkščiai, jį padidinti. Tokia išplėta greičiausiai būtų atliekama pasitelkiant hierarchijos tranzityvumą. Nors apie tai SKOS bendruomenėje būta daug diskusijų, kurių metu galiausiai buvo prieita prie bendros nuomonės, kad šios savybės paprastai nėra tranzityvios, daugelio programinių priemonių veikimas įrodo, kad vis dėlto jos tokios yra.

Atidžiau pažvelgus į tokias sudėtingas hierarchijas, kaip LCSH ar RAMEAU, darosi aišku, kad semantinė išplėta naudojant tranzityvumą dažnai galioja „iki tam tikro taško“, t. y. „vietiniu lygmeniu“. Kadangi dėl žodyno išsamumo globali jo vizija žodynų tvarkytojams ne visada atvira, kuriame nors hierarchijos lygmenyje gali įvykti semantinę išplėtą apsunkinanti „semantinė paklaida“. Toliau pateikiamas LCSH pavyzdys iš tikrųjų nėra toks kraštutinis, kaip gali atrodyti – panašių pavyzdžių rasta ir daugiau. RAMEAU aptikome cikłų, aprėpiančių iki trylikos sąvokų ir gluminančių paieškos sistemas.

Auxiliary sciences of history
.Civilization
..Learning and Scholarship

...Humanities
....Philosophy
.....Attention
.....Listening
.....Eavesdropping
.....Wiretapping

Jeigu žvelgtume į šią hierarchiją vietiniame kontekste, neaptiktume jokio semantinių ryšių išdėstymo netikslumo: prasmingas kiekvienas santykis. Tačiau tai naudojant semantinei išplėtai visuotiniu mastu, rezultatai gali būti daugiau negu keisti.

Vėlgi siekiant tikslios semantikos, žodyno tvarkytojui kiltų abejonių dėl tokių plačių hierarchijų tinkamumo ir dėl jų atsirandančios semantinės paklaidos netgi bibliotekų sistemose, jau nekalbant apie atvirąjį saityną.

4. Išlaikykite žodyną kiek įmanoma siauresnį ir paprastesnį

Iš pateiktų pavyzdžių pakankamai aiškiai matyti, kad didelės apimties žodynai nėra tinkamai pritaikyti semantinėms technologijoms ir kad būtų geriau, jei jie būtų organizuojami vertikaliu principu, palengvinančiu dviprasmiškumo šalinimą. Veiksmingas žodynas yra *siauras savo reikšme ir išsamus savo apimtimi* arba *siauras savo apimtimi ir išsamus savo reikšme*. Bandymų parengti žodynus, kurie būtų plačios apimties ir išsamūs savo reikšme, rezultatas būna milžiniški, apimantys tūkstančius sąvokų žodynai, kuriuos sunku tvarkyti ir jais naudotis. Tai suteikė karčios patirties, nes šie milžiniški žodynai plačiai naudojami indeksavimui. Dabar prioritetiniu uždaviniu turėtų būti nustatyti šių didelių bendrų ir siauresnių vertikaliu principu paremtų žodynų sutaptį. To neįmanoma atlikti per vieną dieną, tačiau čia galėtų padėti semantinių saityno kalbų pagrindu sukurtos tvarkymo priemonės. Apie tai dar bus kalbama paskutiniame skirsnyje.

5. Parenkite sankirtas su kitais žodynais

Svarbią pridėtinę vertę žodynui suteikia sankirtos su tapačiomis kitų žodynų sąvokomis, ypač daugiakalbėje aplinkoje. Netgi vienakalbėje aplinkoje bendri žodynai gali būti praplečiami, suteikiant jiems labiau specializuotų žodynų, parengtų nepriklausomų kūrėjų, savybių. Tokioms sankirtoms išraiškos suteikia SKOS, o semantinės paieškos priemonės gali pritaikyti tokias sankirtas semantinei išplėtai. Šios srities veikloje pirmauja MACS [16] projektas.

Tokios bendrų išsamių žodynų sankirtos su siauresniais palengvintų jų naudojimą žvalgymo taksonomijoms, pateikiančioms supaprastintą ir individualizuotą bibliotekų rodyklių ir katalogų vaizdą.

6. Sąvokoms būdingas gyvavimo ciklas, tačiau šaunieji URI nesikeičia

Kadangi semantinio saityno programinės priemonės reikalauja, kad jų naudojami URI turi būti kiek įmanoma „šaunesni“ [17], žodynų sąvokoms priskirti URI neturi kisti laikui bėgant. Keičiantis žodynų sąvokoms, turi būti griežtai užtikrinamas URI ir pridėtų aprašų pastovumas, net pasenusių sąvokų atveju. Vartojant semantinio saityno kalbas ir protokolus, gali būti naudojamos filtravimo ir peradresavimo priemonės.

7. Pateikite ir viešinkite savo žodyną kaip paslaugą

Semantinis saitynas – tai ne tik kalbų rinkinys, tai ir paslaugų struktūra. Tam, kad žodynas būtų visiškai tinkamas naudoti, jis turi būti prieinamas visų tipų saityno programinėms priemonėms, suderinamoms su saityno struktūra. Išsamios rekomendacijos išdėstytos puikiai parašytuose vadovuose [18], čia tik apibendrintai pateiksime pagrindines iš jų.

1. Kiekvienas sąvokos URI turi užtikrinti turinio atranką, mašinoms teikdamas formalius RDF aprašus, o žmogui vartotojui – HTML.
2. Žodynus pateikite parsisiuntimui skirtais paketais, kiekvienai sąvokos schemai skirdami atskirą paketą.
3. Pateikite SPARQL galinį tašką, kad vartotojai galėtų išskirti specifines reikmes, atitinkančias žodyno dalis.
4. Atskleiskite žodyno turinį, pavyzdžiui, pasitelkdami VoID [19] ontologiją.
5. Viešinkite skelbiamą žodyną įvairiuose semantinio saityno forumuose.

8. Naudokite semantinio saityno programinę įrangą žodynui tvarkyti

Svarbu ir tai, kad, plėtojant semantinio saityno kalbas palaikančią programinę įrangą, atsiranda naujų žodynų tvarkymo priemonių, palengvinančių tokias užduotis, kaip nuoseklumo valdymas, išraiškos tobulinimas naudojant ontologijas, apibrėžiančias santykius, kurie yra specifiškesni už hierarchinius ir asociacinius, versijų tvarkymas, sąvokų atsisakymas ir plėtra, sutapties su kitais žodynais nustatymas, importas ir eksportas RDF formatu savoje aplinkoje, SPARQL galinių taškų įterpimas.

Puikus šios veiklos pavyzdys yra Europos bendruomenės atliekamas semantinio saityno programinės įrangos arsenalo pritaikymas tvarkant ir skelbiant EUROVOC, o ateityje – ir daugelio kitų Europos Sąjungos leidinių biuro skelbiamų žodynų. Ši veikla atitinka didelius tvarkybos ir procedūrų eigos reikalavimus: žodynai yra daugiakalbiai (daugiau negu dvidešimt kalbų); terminai pasižymi išsamia struktūra, apimančia sinonimus, akro-

nimų vertimus; informacija organizuojama remiantis mikrotezaurais, kuriems būdingi įvairūs ribojimai; ir, žinoma, žodynai skelbiami SKOS formatu skiriant unikalias versijas ir įterpiant laiko žymas.

Dabarties veikla ir perspektyvos

Pabaigai pateiksime trumpą dabartinių semantinio saityno bibliotekų iniciatyvų apžvalgą. Ji nėra baigtinė – iniciatyvų sąrašas neabejotinai ilgės su kiekvienu mėnesiu ir metais.

Kongreso bibliotekos autoritetiniai duomenys ir žodynai

Kongreso biblioteka pirmoji pradėjo skelbti savo žodynus formatais, atitinkančiais semantinio saityno standartus ir susietų duomenų pažangiausią praktiką. 2009 m. buvo paskelbtas pirmasis žodynas – LCSH, 2010 m. sekė kiti. Kaip minėta, automatinis esamų žodynų perkėlimas į SKOS kelia tam tikrų problemų, todėl šią iniciatyvą būtina išsamiai įvertinti. Tačiau kreiptis kisti negali – *id.loc.gov* vardų erdvė turi likti, ir vėliau, skelbiant žodyną semantiniame saityne, turi būti atsižvelgiama į šią vardų erdvę bei su ja siejama.

Europos žodynai

TelPlus projekto ribose atliktų tyrimų rezultatas buvo keitimasis duomenimis tarp Prancūzijos nacionalinės bibliotekos ir jos techninių partnerių, siekiant nustatyti RAMEAU perkėlimo į semantinį saityną kelią. Tikimės, kad ši veikla plėsis, įtraukdama ir partnerius iš *Europeana* ir Europos bibliotekos. Daugiakalbė ir daugiakultūrė aplinka kelia labai rimtų iššūkių, tačiau kartu ji atveria ir plačias galimybes. Minėjome Europos Sąjungos leidinių biuro iniciatyvas. Neabejotina, kad glaudi visų šių Europos iniciatyvų sąveika ir keitimasis gerąja patirtimi galiausiai leis atsirasti turtingam daugiakalbiam susijusių viešosios prieigos žodynų debesiui, suteikiančiam prieigą prie viešųjų Europos vertybių – kultūrinių, įstatyminių, ekonominių.

W3C bibliotekų susietų duomenų rengimo grupė

W3C bibliotekų susietų duomenų rengimo grupė [20] įpareigota vienerius metus vienyti šios srities pastangas, identifikuoti ir viešinti pažangiausią praktiką bei padėti iki galo realizuoti bibliotekų ir semantinio saityno sąveiką. Ši grupė, jungianti aktyviausius bibliotekų bendruomenės dalyvius, numato glaudžiai bendradarbiauti su ISO darbo grupe, skirta naujajam tezauro standartui (ISO 25964) parengti, ir *Dublin Core* metaduomenų iniciatyva.

Išvada

Tikimės, kad šiuo straipsniu sugebėjome parodyti, jog žodynų tvarkytojams atėjo laikas pradėti aktyviai ir veiksmingai bendradarbiauti su semantinio saityno bendruomene – tam, kad jų vertingas palikimas būtų plėtojamas ir naudojamas naujovišku būdu. W3C parengė dirvą tokiai veiklai; W3C bibliotekų susietų duomenų rengimo grupės chartijoje rašoma: „Siekiant pateikti šį turinį saityne, būtina perorientuoti bibliotekų informacinio suderinamumo perspektyvą, remiantis esama saityno struktūra ir standartais. Bibliotekų sistemose jau yra daug struktūruotų duomenų, kurie gali būti pateikiami kaip susieti duomenys, naudojant semantinio saityno technologijas. Kultūros paveldo institucijos galėtų būti svarbūs autoritetinių duomenų rinkinių (asmenų, temų ir t. t.) teikėjai susietų duomenų saitynui“.

Autoritetinių duomenų rinkiniai neabejotinai priskirtini paveldėtiems duomenims. Semantinio saityno technologijos leistų bibliotekininkams juos veiksmingiau tvarkyti suteikiant prie jų prieigą per semantinių paslaugų sąsajas. Trumpiau tariant, tolesnė raida turėtų skatinti žodynų rengėjus praplėsti savo tradicines veiklos ribas iki viso saityno, pasitelkiant paskirstytąją ir susietą žodyno kaip paslaugos struktūrą.

Iš anglų kalbos vertė T. Auškalnis

Straipsnis parengtas pagal pranešimą, skaitytą 2010 m. Geteborge (Šveicarija) vykusioje 76-ojoje IFLA konferencijoje.

Nuorodos

[1] OCLC WorldCat http://www.oclc.org/us/en/worldcat/default.htm

[2] W3C Semantic Web Activity http://www.w3.org/2001/sw/

[3] TELplus http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/

[4] Library of Congress Authorities and Vocabularies http://id.loc.gov/

[5] EUROVOC http://europa.eu/eurovoc/

[6] Search Engine History http://www.searchenginehistory.com/

[7] Humans Do It Better: Inside the Open Directory Project, Chris Sherman, ONLINE Mag, July 2000 http://www.onlinemag.net/ol2000/sherman7.html

[8] Wikipedia discussion about the "Uncategorized categories" list http://en.wikipedia.org/wiki/Wikipedia_talk:Special:UncategorizedCategories

[9] Representing Knowledge Soup in Language and Logic, John Sowa, 2002 http://www.jfsowa.com/talks/souprepr.htm
[10] The Accidental Taxonomist, by Heather Hedden, Information Today, May 2010, ISBN 978-1-57387-397-0 http://www.hedden-information.com/accidental-taxonomist.htm
[11] The Taxonomy Warehouse http://www.taxonomywarehouse.com/
[12] Simple Knowledge Organization System http://www.w3.org/2004/02/skos/
[13] RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) http://rameau.bnf.fr/
[14] Library of Congress Subject Headings http://id.loc.gov/authorities#ConceptScheme
[15] AGROVOC Thesaurus http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[16] Integrating MACS initial data and new alignments into TEL framework http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/documents/TELplus_D3.4_04012010.pdf
[17] Cool URIs don't change, Tim Berners-Lee, 1998 http://www.w3.org/Provider/Style/URI
[18] How to Publish Linked Data on the Web, C. Bizer, R. Cyganiak, T. Heath. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
[19] Vocabulary of Interlinked Datasets http://semanticweb.org/wiki/VoiD
[20] Library Linked Data Incubator Group charter http://www.w3.org/2005/Incubator/lld/charter