UDK 025.323:004

VIAF (Virtualus tarptautinis autoritetinis failas): Vokiečių bibliotekos ir Kongreso bibliotekos vardų autoritetinių failų susietis

Rick BENNETT
OCLC, Dublinas, Ohio valstija, JAV, el. p. rickbennett@oclc.org
Christina HENGEL-DITTRICH
Vokiečių biblioteka, Frankfurtas prie Maino
Edward T. O'NEILL
OCLC, Dublinas, Ohio valstija, JAV, el. p. oneill@oclc.org
Barbara B. TILLETT
Kongreso biblioteka, Vašingtonas, DC, JAV, el. p. btil@loc.gov

Vokiečių biblioteka, JAV Kongreso biblioteka ir OCLC kuria asmenų vardų Virtualų tarptautinį autoritetinį failą (Virtual International Authority File – VIAF), kuriame susiejami pasaulio valstybinės bibliografijos tarnybų autoritetiniai įrašai ir jie bus laisvai prieinami žiniatinklyje. Projekto tikslas – išmėginti, ar įmanoma automatiškai susieti nacionalinių autoritetinių failų autoritetinius įrašus, ir parodyti tokios susieties naudą. Kongreso bibliotekos ir Vokiečių bibliotekos autoritetiniai ir bibliografiniai failai buvo naudojami kuriant pradinį VIAF, apimantį per šešis milijonus vardų ir daugiau nei pusę milijono saitų. Projekte pagrindinis dėmesys skirtas automatinio vardų sutapties algoritmo sukūrimui, kuris naudotų ir autoritetinių įrašų, ir susijusių bibliografinių įrašų duomenis. Parodomas asmenų vardų algoritminės susieties su nacionaliniais autoritetiniais failais praktiškumas; 70 proc. bendrų abiem failams vardų autoritetinių įrašų buvo automatiškai susieta mažesniu nei 1 proc. klaidų dažniu. Strateginis VIAF projekto tikslas – sujungti daugelio nacionalinių bibliotekų ir kitų svarbių šaltinių vardų autoritetinius įrašus į bendrai naudojamą pasaulinę autoritetinių įrašų paslaugą.

Reikšminiai žodžiai: Virtualus tarptautinis autoritetinis failas (VIAF); asmenų vardų tapatinimas.

Įvadas

Virtualaus tarptautinio autoritetinio failo (VIAF)[1], kuriame autoritetiniai įrašai, rodantys tą patį entitetą pasaulio valstybinės bibliografijos tarnybose, turėtų būti susieti ir prieinami internete, galimybes pripažino keletas IFLA Katalogavimo sekcijos grupių. Toks VIAF būtų visuotinės bibliografinės apskaitos koncepcijos praktine plėtra ir remtųsi darbu, kurį atlieka kiekviena valstybinės bibliografijos tarnyba. Tai leistų koegzistuoti šalies arba regiono aprobuotos formos variantams ir tuo pačiu atitiktų viso pasaulio vartotojų poreikius vartoti variantus norima kalba, rašmenimis ar rašyba.

Žiniatinklio ateities plėtra siejama su ontologijų siūlomomis galimybėmis; jų dėka žiniatinklis taptų imlesnis kompiuteriniam ir automatiniam apdorojimui. VIAF, susietas su kelių šaltinių, tokių kaip referavimo ir indeksavimo tarnybos, archyvai, muziejai, leidėjai ir t. t., kontroliuojamais žodynais ir autoritetiniais failais, galėtų tapti „semantinio žiniatinklio"[2] konstrukcijos elementu. Bibliotekos turi galimybę dalyvauti tokios ateities kūrime ir padėti įgyvendinti šią viziją. Svarbu plėtoti tokią bendrą viziją, kad VIAF taptų laisvai prieinamas viso pasaulio vartotojams.

Asmenų vardų saitus autoritetiniuose failuose buvo siūloma nagrinėti ir kituose projektuose. LEAF (*Linking and Exploring Authority Files*) projekte[3] bandyta susieti autoritetinius įrašus iš skirtingų šaltinių, tarp jų – bibliotekas,

archyvus, dokumentavimo ir mokslinius centrus. Šie įrašai yra įvairiais formatais, labai skiriasi jų rūšiniai ypatumai ir turinio apimtis. LEAF projekte siūloma juos automatiškai susieti su įrašais, įkeliant į sistemą. Dėl vardų autoritetinių įrašų skirtingų šaltinių buvo pripažinta, kad vienintelė bendroji informacija, leidžianti sukurti saitus, yra vardas su nuorodomis „žr." ir su juo susijusios datos. Kadangi dabartinių dalyvių autoritetiniuose įrašuose dažnai nėra datos, vardų autoritetiniuose įrašuose tikėtinos nepriimtinai dažnos nesutapties klaidos.

InterParty projektas[4] – tai ES finansuojamas parodomasis projektas, skirtas sukurti saitus tarp skirtingų organizacijų autoritetinių failų. Pagrindinis jo tikslas – remti skaitmeninių teisių valdymą. Pasiūlyta InterParty sistema turėtų suteikti vieną kreipties elementą į daugelį ją sudarančių duomenų bazių, taigi pirmiausia suteikti centralizuotos paieškos paslaugą. Kadangi saitai tarp vardų bet kurioje bazėje identifikuojami rankiniu būdu, tai ryšius nustatantis asmuo gali įvesti ir saitą. Vėliau saitai gali būti naudojami automatiškai. Saitų patikimumas priklauso nuo juos kuriančios organizacijos. Kiti sistemos dalyviai neturi patvirtinti atskiro dalyvio saito teiginio. Projekte numatyta algoritminės sutapties galimybė, bet nenustatyti techniniai arba duomenų reikalavimai, būtini susieties gebai palaikyti.

VIAF projektas

2003 m. vykusio IFLA Pasaulio bibliotekų ir informacijos kongreso Berlyne metu Vokiečių biblioteka, Kongreso biblioteka ir OCLC sutarė kurti asmenų vardų Virtualų tarptautinį autoritetinį failą (VIAF)[5]. VIAF projekto tikslas – išmėginti, ar įmanoma automatiškai susieti nacionalinių autoritetinių failų autoritetinius įrašus ir parodyti VIAF naudą. VIAF projektas susies Vokiečių bibliotekos ir Kongreso bibliotekos vardų autoritetinius failus į vieną virtualią vardų autoritetinių įrašų sistemą. OCLC kuria programinę įrangą, kuri leis palyginti dviejų autoritetinių failų vardų autoritetinius įrašus. Strateginis VIAF projekto tikslas – sujungti daugelio nacionalinių bibliotekų ir kitų svarbių šaltinių autoritetinius vardus į bendrai naudojamą pasaulinę asmenų, kolektyvų, konferencijų, vietų ir t. t. autoritetinių įrašų paslaugą.

VIAF projektą sudaro penki etapai:

1. Sukurti vadinamuosius „praplėstus autoritetinius" įrašus iš *Personennormdatei* (PND) ir Kongreso bibliotekos autoritetinių įrašų. Šis etapas apims tam tikrų autoritetinių įrašų identifikavimą, jų įtraukimą į praplėstus autoritetinius įrašus ir bet kokių specialių naujų failų apdorojimo poreikių nustatymą.

2. Sukurti sutapties algoritmus ir palyginti PND ir Kongreso bibliotekos praplėstus autoritetinius įrašus, siekiant sukurti pradinę VIAF versiją. Čia pasikartojo 1-ojo etapo procesai, nes tarpiniai sutapties rezultatai išryškino papildomos informacijos, kuri turėtų būti išrinkta ir įtraukta į praplėstus autoritetinius įrašus bei leistų patobulinti sutapties procesą, poreikį.

3. Atvirųjų archyvų iniciatyvos (*Open Archive Initiative* – OAI)[6] pagrindu sukurti serverį ir suteikti prieigą prie VIAF.

4. Siekiant palaikyti VIAF duomenų bazę būtina, kad visos dalyvaujančios tarnybos papildytų ir keistų autoritetinius ir bibliografinius įrašus. Ši duomenų naujinimo ir priežiūros sistema bus kuriama pagal protokolus, kuriuos naudoja OAI informacijai apie duomenų naujinimą gauti.

5. Prieigai prie VIAF įrašų atvirajame žiniatinklyje bus sukurta vartotojo sąsaja. Ilgainiui duomenų bazė palaikys Unikodą, įvairias kalbas ir rašmenis. Vykdant tiesiogiai į duomenų bazę siunčiamas užklausas, pavyzdžiui, surasti Kongreso bibliotekos vardo variantą, sutampantį su PND esančiu vardu, gali būti naudojamas paprastas hipersaitas, palaikantis semantinio žiniatinklio galimybes.

Iš pradžių projektas buvo sutelktas parodyti VIAF galimybes susieti vardų autoritetinius įrašus tarp PND ir Kongreso bibliotekos vardų autoritetinių failų (*Library of Congress Name Authority File* – LCNAF). 2005 m. gruodžio 31 d. LCNAF buvo 4,2 mln. asmenų vardų autoritetinių įrašų. Iki to paties laikotarpio Kongreso biblioteka sukūrė ir išplatino 9,3 mln. bibliografinių įrašų.

2005 m. pabaigoje PND failą sudarė 2,6 mln. asmenų vardų autoritetinių įrašų. PND autoritetinis failas naudojamas tiek Vokiečių bibliotekos, tiek Bibliotheksverbund Bayern bibliografiniuose įrašuose. Šiuose dviejuose bibliografiniuose failuose yra 15 mln. bibliografinių įrašų, susietų su PND autoritetiniais įrašais.

Vardų tapatinimo problema

Iš pradžių VIAF funkcionuos kaip vokiečių-anglų ir anglų-vokiečių kalbų asmenų vardų žodynas. Pavyzdžiui, Amerikos vartotojas ieško J. P. De Valk (vardo forma, suteikta Kongreso bibliotekos), vardas gali būti automatiškai „išverstas" į Johannes P. De Valk (forma, suteikta Vokiečių bibliotekos). Įprasta, kad, kaip ir šiuo atveju, skirtingos katalogavimo tarnybos nustato nevienodą to paties asmens vardo formą arba, priešingai, keliems autoriams vartoja tą pačią vardo formą. Tikėtina, kad Vokiečių biblioteka J. P. De Valk formą gali nustatyti visai kitam autoriui.

Tam pačiam asmeniui gali būti vartojamos skirtingos vardo formos arba tą pačią vardo formą gali turėti skirtingi žmonės, dėl to sunku patikimai palyginti skirtingų autoritetinių failų vardus. Dviejų autoritetinių failų aprėptis labai skiriasi; tik nedidelė asmenų vardų dalis pateikiama abiejuose failuose. Todėl siekiant užtikrinti patikimą sutaptį reikia vartoti ir kitą informaciją, ne tik patį vardą. Asmenų vardų autoritetiniuose įrašuose dažnai pateikiamos asmens

gimimo ir (arba) mirties datos. Paprastai atskirti to paties vardo žmones pakanka gimimo ir mirties datų derinio.

Šiai vardų tapatinimo autoritetiniuose įrašuose problemai išspręsti iš Kongreso bibliotekos ir Vokiečių bibliotekos autoritetinių failų buvo parinkti bendri vardai be papildomos informacijos. Po to šių autoritetinių įrašų poros rankiniu būdu buvo peržiūrėtos, siekiant nustatyti, ar jos atitinka tą patį asmenį. Šio patikrinimo metu nustatyta, kad apie 10 proc. vardų porų priklauso skirtingiems žmonėmis. Taigi nustatytos vardo formos sutapties klaidų dažnis nepriimtinai didelis. Kadangi dviejuose autoritetiniuose failuose vardo formos ne visada tapačios, parenkant panašias, bet netapačias vardų poras susidarytų didesnis klaidų dažnis. Šiuo paprastu būdu taip pat nepavyksta susieti daugelio vardų, kuriems buvo suteiktos skirtingos formos.

Vardų tapatinimo sprendinys

Patvirtinant ar atmetant galimus vardų atitikimus būtina papildoma tapatinimo informacija. Pavyzdžiui, panagrinėkime Kongreso bibliotekos Diane Glynn autoritetinio įrašo duomenis:

100 10 $a Glynn, Diane, $d 1946-
400 10 $a O'Connor, Diane, $d 1946- $w nna
670 $a Country western dancing, 1994: $b CIP t.p. (Diane Glynn) pub. info. (an avid country w. dancer & co-author of How to make your man more sensitive)

Vardai ir gimimo data – vieninteliai tiesiogiai vartojami duomenys. Kompiuterinio apdorojimo metu galėtų būti išrinktos dvi antraštės, įtrauktos į 670 lauką (Duomenų šaltinis). Iš tikrųjų iš šių laukų patikimai išrinktos gali būti tik kelios antraštės.

Akivaizdu, kad bibliografiniai įrašai yra papildomų duomenų apie asmenį šaltinis. Šie bibliografiniai įrašai gali pateikti apie asmens kūrinį papildomų žinių, padedančių atskirti asmenis, turinčius panašius vardus. Viename bibliografiniame įraše nurodyta:

100 1 $a Glynn, Diane, $d 1946- -
245 10 $a How to make your man more sensitive / $c by Diane and Dick O'Connor.
700 1 $a O'Connor, Dick, $d 1938- $e joint author –

Bibliografiniuose įrašuose yra dviejų rūšių papildomų duomenų. Paprastai bibliografiniuose įrašuose yra konkretūs kūrinio duomenys – antraštė ir apraiškos konkretūs duomenys – ISBN. Tapačios antraštės užtikrina beveik tikslią vardų sutaptį. Bibliografiniame įraše taip pat yra papildomų duomenų, kurie gali būti taikomi daugeliui asmens kūrinių. Šie duomenys gali padėti palyginti autorius, jei nėra konkrečios antraštės sutapties. Toks pavyzdys gali būti bendraautoris Dick O'Connor. Dick O'Connor daugiau nei vienos Diane Glynn knygos bendraautoris, tai labai svarbus faktas tapatinant vardus autoritetiniame faile. Net jeigu tas pats kūrinys įtrauktas į abi duomenų bazes, bet tik vienoje iš jų pateiktas kūrinio vertimas, antraštės sutaptis gali būti sunkiai automatiškai nustatoma. Tuo atveju bendraautorio vardas duomenų bazėse gali būti panašus ir tai patvirtintų sutaptį.

Visų esamų bibliografinių įrašų, kuriuose vardas yra kaip pagrindinis ar papildomas pradmuo arba kaip dalykas, duomenys transformuojami, kad būtų sukurtas tarpinis įrašas, vadinamasis „išvestinis autoritetinis". Tada šie išvestiniai autoritetiniai įrašai, sujungti su originaliu autoritetiniu įrašu, tampa praplėstu autoritetiniu įrašu. Kadangi praplėsti autoritetiniai įrašai apima papildomus, su vardu susijusius bibliografinio įrašo duomenis, jie gali padėti tiksliau atlikti tapatinimo procesą nei patys autoritetiniai įrašai.

Vardų sutapties patvirtinimas

Paprastas dviejų autoritetinių failų vardų sugretinimas yra priimtinas būdas surasti tą patį asmenį. Galimi vardo formos skirtumai mažina tikimybę, kad tai bus tas pats asmuo. Kad automatiškai būtų patvirtinta šių asmenų sutaptis, šiuo atveju 1) vardai privalo būti suderinami ir 2) turi būti pakankamai papildomų duomenų, patvirtinančių sutaptį.

Suderinamumas reikalauja, kad nebūtų jokių vardų skirtumų, kurie sukliudytų pateikti tą patį asmenį. Vardai gali skirtis savo išsamumu, pavyzdžiui, John A. Smith ir John Allen Smith. Šie vardai suderinami, kadangi „A" gali būti Allen. Tačiau John A. Smith ir John B. Smith nesuderinami dėl skirtingo antrojo inicialo. Tikrinant suderinamumą atsižvelgiama tiek į vardo aprobuotą formą, tiek į formos variantus.

Jei nustatoma, kad vardai suderinami, šiems vardams surandami sutaptį patvirtinantys papildomi duomenys. Bibliografiniuose failuose gali būti daug skirtingų, bet panašių antraščių ir daug skirtingų, bet panašių vardų. Jei vardo ir antraštės poros abiejuose failuose sutampa, tikėtina, kad vardas rodo tą patį asmenį. Toks pagrindinis būdas taikomas ir kitiems iš bibliografinių įrašų gaunamiems duomenims.

Kaip teigiama koreliacija, datos atidžiai įvertinamos atskirai. Jei datos skiriasi daugiau nei metais, vardai laikomi nesuderinamais ir sutaptis atmetama. Tarp datų leistinas vienerių metų skirtumas. Kuriant VIAF buvo palyginti paprasta rasti nežymius kelių datų nesutapimus, o sutapties patvirtinimui net nežymių datos nuokrypių atveju pakako papildomų tapatinimo duomenų.

Sugretinus du praplėstus autoritetinius įrašus, kiekvienas sutampantis elementas yra laikomas tapatinimo tašku. Tapatinimo taškai suskirstyti į tris kategorijas: stiprūs, vidutiniai ir silpni. Sutampantiems vardams stiprus sutapties taškas laikomas pakankamu patvirtinti, kad tai yra tas pats asmuo. Stiprūs sutapties taškai yra antraštės, ISBN, gimimo ir mirties datos arba bendraautoriai. Vien tik gimimo datos nepakanka vardams atskirti, ir geriau jas priskirti prie vidutinio tapatinimo taško. Vidutiniais tapatinimo taškais laikoma asmens kūrimo aplinka, t. y. leidėjai, dalykinė sritis arba asmens vaidmuo (pvz., iliustruotojas arba kompozitorius). Stambūs leidėjai leidžia daugelio autorių kūrinius ir kai kurie vardai gali sutapti. Tapatinimas pagal daugelį vidutinių taškų yra pakankamas sutapčiai patvirtinti. Silpni tapatinimo taškai laikomi pakankamais neaiškių, kitaip sudarytų sutapčių atskyrimui. Tokie silpnų tapatinimo taškų pavyzdžiai yra kalba, dalykinė sritis, publikavimo šalis.

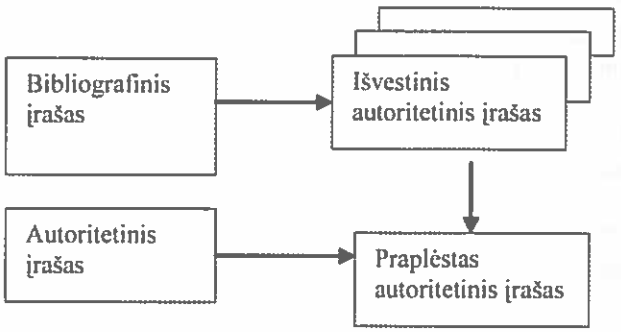
Siekiant sujungti tapatinimo taškus, kiekvienam jų priskirtas taškų skaičius. Numeriui, tokiam kaip ISBN, sutaptis yra arba tiksli, arba jos nėra, tada taškų skaičiaus rezultatas – vienetas sutapčiai ir nulis – nesutapčiai. Tekstui, pavyzdžiui, antraštei, taškų skaičiaus rezultatas nuo nulio iki vieneto priskirtas priklausomai nuo to, koks teksto panašumas. Teksto panašumui apskaičiuoti naudojama trigrama, kuri remiasi taškų skaičiaus metodu. Pavieniai taškų skaičiai pakeičiami priklausomai nuo svorio, grindžiami stiprumu (stiprus, vidutinis ar silpnas) ir sumuojami. Jei bendras taškų skaičius viršija nustatytą ribą bandomojo proceso metu, sutaptis yra patvirtinta. Dabartiniame sutapties algoritme daugelio įrašų bandymas leidžia papildyti šių kategorijų rezultatus. Tikimasi, kad toks papildymas bus atliekamas toliau, daugiau autoritetinių įrašų failų bus įtraukiama į sistemą ir įgyjama patirties.

Praplėsto autoritetinio įrašo sudarymas

Anksčiau aprašyti būdai buvo naudojami ir PND, ir Kongreso bibliotekos praplėstiems vardų autoritetiniams įrašams sukurti. Kongreso bibliotekos bibliografiniai failai buvo apdoroti siekiant gauti išvestinius autoritetinius įrašus Kongreso bibliotekos praplėstam autoritetiniam failui, o Vokiečių bibliotekos ir *Bibliotheksverbund Bayern* bibliografiniai failai buvo apdoroti siekiant išplėsti PND autoritetinius įrašus. 1 pav. pateikta bendra informacijos, naudojamos sudaryti praplėstą autoritetinį įrašą, schema.

Kongreso bibliotekos praplėstame autoritetiniame faile, iš 4,2 mln. autoritetinių įrašų gali būti praplėsta 3,8 mln. (90%). Tik 2,6 mln. (60%) buvo papildyti duomenimis iš bibliografinių įrašų, iš viso 7,4 mln. antraščių. Kiti papildymai buvo atlikti naudojant 4,1 mln. antraščių, paimtų iš autoritetinių įrašų 670 laukų (Duomenų šaltinis). Baigiamojoje straipsnio dalyje bus parodyta, kad, siekiant gauti sutapčių, antraštės yra svarbiausias papildomas elementas.

1 pav. Praplėsto autoritetinio įrašo sudarymas



Pralėstam PND autoritetiniam failui 2,4 mln. iš 2,6 mln. (90%) autoritetinių įrašų buvo šiek tiek papildyti, 2 mln. (80%) buvo praplėsti duomenimis iš bibliografinių įrašų. Likusieji 400 tūkst. įrašų buvo papildyti antraštėmis iš pačių PND autoritetinių įrašų.

Sutapties patikrinimo metodika

VIAF dalyviai palaikė tapatinimo proceso kūrimą kruopščiai peržiūrėdami ir komentuodami rezultatus. Pavyzdžiui, iš pradžių buvo naudojamos serijų antraštės, bet tapatinant vardus dažnai buvo randama netikslumų. Po kiekvienos peržiūros buvo atliekami pakeitimai, dėl kurių sutapčių skaičius padidėdavo arba sumažėdavo klaidingų sutapčių. Per tą laiką buvo sukurtas tikslios taškų skaičiaus ribos ir taškų skaičiavimo algoritmas. Čia pateikti tik patvirtinti galutiniai patikrinimo rezultatai.

Siekdami patvirtinti tapatinimo proceso tikslumą ir efektyvumą, patyrę Vokiečių bibliotekos ir Kongreso bibliotekos kataloguotojai peržiūrėjo bandomuosius vardų sutapties pavyzdžius. Pirmasis bandomasis pavyzdys turėjo du tikslus: nustatyti iš dalies sutampančius vardus dviejuose autoritetiniuose failuose ir kokia šių vardų porų dalis gali būti identifikuota tapatinimo metu. Antrasis bandomasis pavyzdys panaudotas nustatyti sistemines klaidas arba trūkumus, kurie gali būti ištaisyti, ir įvertinti bendrą klaidų dažnį.

Iš PND į pirmąjį bandomąjį pavyzdį atsitiktine tvarka buvo atrinktas 391 autoritetinis įrašas. Šiems įrašams Kongreso autoritetiniame faile automatiškai ir rankiniu būdu buvo ieškoma sutapčių. Siekiant padidinti automatiškai atliekamo tapatinimo proceso dalį, PND autoritetiniai įrašai buvo suporuoti su visais Kongreso bibliotekos autoritetiniais įrašais, kuriuose buvo ta pati pavardė, todėl buvo parinkta 74 tūkst. porų. Visoms 74 tūkst. vardų poroms ir automatiškai sutapatintoms 79 PND ir Kongreso bibliotekos autoritetinių įrašų poroms buvo taikomas sutapties algoritmas.

Visus 391 PND autoritetinius įrašus peržiūrėjus rankiniu būdu, buvo nustatyti 35 papildomi vardai, kurie atitiko Kongreso bibliotekos autoritetinį įrašą, bet nebuvo nustatyti parenkant pavardžių poras, kai siekiant patikrinti sutaptį naudotas sutapties algoritmas. Rankiniu būdu atliekamos peržiūros metu nustatyta, kad 79 automatiškai parinktos sutaptys buvo teisingos. Remiantis PND bandomuoju pavyzdžiu nustatyta, kad apie 30% PND vardų pateikti ir Kongreso bibliotekos autoritetiniuose įrašuose ir kad algoritmas gali tapatinti apie 70% bendrų vardų. Iš šių rezultatų ekstrapoliacijos apytiksliai 800 tūkst. vardų, turimų abiejuose autoritetiniuose failuose, galima tikėtis, kad automatiškai tapatinant galima identifikuoti 550 tūkst. vardų.

Rezultatai taip pat buvo peržiūrėti siekiant patobulinti vardų porų nustatymo procesą. Vartojant tik pavardes, vos ne tūkstančiui vardų porų būtina atlikti visą tapatinimo procesą kiekvienai porai. Porų parinkimo rezultatų patikrinimas buvo atliekamas rankiniu būdu siekiant įsitikinti, kad tapatinimo strategija pagal pavardę, vardą ir informaciją, apribotą datomis, galėtų būti panaudota apytikriam vardų suderinamumui patikrinti. Šis paprastas rodiklis nustato iki 95% sutapčių tik su viena iš keturių porų. Paprastas rodiklis yra naudingas ir veiksmingas, o jo nežymios pataisos padės pagerinti tolesnius rezultatus.

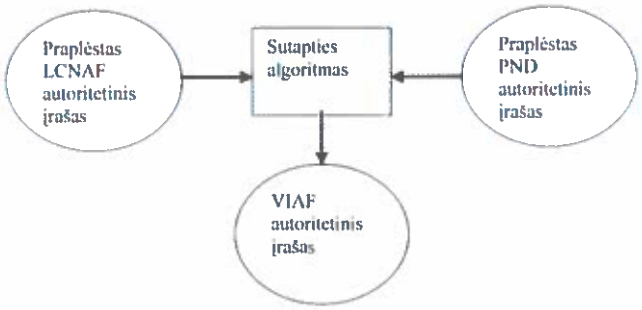
Antro bandomojo pavyzdžio tikslas – įvertinti tapatinimo klaidų dažnį. Viena šio proceso sudėtinių dalių buvo kaip bandomasis pavyzdys atliekamas pirminis taškų skaičiaus ribos patikrinimas ir, jei reikia, jos koregavimas. Naudojant taškų skaičiaus ribą tikimybė, kad ribai artimų taškų sutapties klaidų bus daugiau nei apskaičiuota parenkant poras ir gali būti viršyta riba. Daugumoje autoritetinių įrašų, kuriuose buvo sutampančių vardų, buvo gauti geresni negu ribinės vertės rezultatai. Siekiant visiškai sumažinti klaidų dažnį rankinės peržiūros metu, bandomasis pavyzdys buvo suskirstytas į keturis pogrupius. Rankinės peržiūros metu buvo nustatomos klaidos visoms sutaptims, ir kiekvienam bandomojo pavyzdžio pogrupiui buvo nustatytas klaidų dažnis ir patikimumas. Šie daliniai rezultatai buvo apskaičiuoti ir susumuoti nustatant bendrą tapatinimo metodikos klaidų dažnį. Klaidingų sutapčių skaičius sudarė mažiau nei 1 proc.

Vienas iš bandomojo pavyzdžių pogrupių buvo nagrinėjamas siekiant nustatyti taškų ribą. Jei taškų riba žemėjo, viena klaidinga sutaptis buvo pridedama prie kiekvienos iš trijų teisingų sutapčių. Aišku, kad ribos sumažinimas yra nepagrįstas. Skaičiuojant taškų skaičių virš ribos, klaida buvo tik vienoje sutaptyje iš 25. Kadangi atliekant šiuos skaičiavimus buvo naudojamas palyginti nedidelis sutapčių skaičius, klaidų poveikis bendram dažniui buvo mažas, o tvarkant didesnį skaičių teisingų sutapčių – nedidelis. Taigi išankstinis taškų ribos lygis buvo priimtinas.

Pradinio VIAF kūrimas

Sutapties algoritmas buvo panaudotas abiejų šaltinių praplėstiems autoritetiniams failams patikrinti, ir tokiu būdu gauti sutampantys ir nesutampantys įrašai buvo pakeisti VIAF įrašais. Šis procesas parodytas 2 pav.

2 pav. VIAF autoritetinio įrašo sudarymas



Gautame VIAF faile yra 6,3 mln. praplėstų autoritetinių įrašų, tarp jų – 500 tūkst. susijusių įrašų, 3,7 mln. nesutampančių įrašų iš Kongreso bibliotekos ir 2,1 mln. nesutampančių įrašų iš PND autoritetinių failų. Šie skaičiai panašūs į gautus tikrinant įrašus rankiniu būdu. Apskaičiuota, kad dar yra 250 tūkst. autoritetinių įrašų porų, pateikiančių tuos pačius asmenis, tačiau dėl tinkamų duomenų trūkumo jie negalėtų būti automatiškai sutapatinti. Galutinė sistemos versija leis rankiniu būdu tapatinti tokias poras ir suteiks kitas intelektinio sutapčių nustatymo galimybes. Autoritetiniai įrašai apims nuosekliai suteiktą VIAF įrašo numerį.

3 pav. VIAF įrašas

```
000    nz n
001    viaf30543
005    20050826163535.0
008    050826n| |anannabbn |a aaa
040    VIAF $c VIAF
400 10 $w nnaO'Connor, Diane, $d 1946-
700 17 Glynn, Diane, $d 1946- $2 DLC $0 n 94057411
700 17 O'Connor, Diane $2 DDB $0 108982424
901    052512920 $9 1
901    349917275 $9 1
901    350215532 $9 1
903    75014386 $9 1
910 11 how to make your man more sensitive $9 3
910 11 macht eure manner zartlicher $b liebevolle ratschlage
       fur e neues rollenverhalten $9 1
910 11 macht eure manner zartlicher $b wie e frau ihrem mann
       helfen kann e verstandnisvoll $9 1
919    country western dancing, $9 1
920    0-525 $9 1
```

```
920    3-499 $9 1
920    3-502 $9 1
921    dutton $9 1
921    rowohlt $9 1
921    scherz $9 1
922    gw $9 2
922    nyu $9 1
940    eng $9 1
940    ger $9 2
942    18 $9 1
943    197x $9 3
944    am $9 3
950 11 oconnor, dick $9 2
950 11 oconnor, dick $d 1938 $9 1
999    1 $b 75014386 //r94 $2 DLC
999    1 $b n 94057411 $2 LoCNA
999    2 $b 780147766 $b 790425319 $2 DDB
```

3 pav. parodytas VIAF įrašo MARC 21 formatu pavyzdys. Kadangi svarbiausias VIAF tikslas – pateikti saitus tarp failų, VIAF įrašas apima kiekvieno vardo pradmenį 700 lauke (Pradmenų ryšys), kartu nurodant jo šaltinį. Kadangi nėra vienos aprobuotos vardo formos, 100 laukas (Asmens vardo pradmuo) nevartojamas. Kai sutaptį nustato algoritmas, įraše pateikiami du susiję pradmenys. Jei vardas nesutapatintas, pateikiamas tik vienas 700 laukas.

Papildomi duomenys taip pat įtraukiami į praplėstų autoritetinių įrašų vietinius (9xx) laukus. 4 pav. trumpai aprašyti praplėstuose autoritetiniuose įrašuose vartojami vietiniai laukai. Siekiant supaprastinti tapatinimą, visas tekstas yra sunormintas pagal modifikuotą NACO (*Name Authority Cooperative Program of the Program for Cooperative Cataloging)* taisyklių versiją[7]. Duomenys apie tam tikros sąvokos pasikartojimų atvejus kaupiami polaukyje $9. Kadangi šie duomenys pirmiausia skirti kompiuteriniam apdorojimui, jų nebūtina pateikti galutinio vartotojo matomuose įrašuose. Kai vėliau bus pridedami kiti nacionaliniai autoritetiniai failai, jie pirmiausia bus palyginti su jau esamais praplėstais VIAF įrašais, tokiu pat būdu prijungiant papildomas sutaptis prie VIAF įrašų. Kai sutaptys jau gautos, papildomi duomenys iš sutampančių įrašų taip pat suliejami.

Daugeliu atvejų, parenkant sutampančius vardus, aprobuotas vardas viename faile sutampa su daugeliu aprobuotų vardų kituose failuose. Kadangi VIAF tikslas yra sukurti tiesioginę susieties paslaugą, sutaptis būdavo nepatvirtinama, jei buvo gaunama daug sutapčių. 70 tūkst. algoritminių sutapčių buvo atmesta dėl to, kad buvo gauta daug sutapčių. Buvo nustatytos mažiausiai dvi priežastys, kodėl gaunama daug sutapčių.

Pirma, PND yra daug nediferencijuotų vardų, kiekvienas jų sutapdavo su dviem ar daugiau diferencijuotų vardų Kongreso bibliotekos vardų autoritetiniame faile. Vokietijos katalogavimo praktikoje pagal katalogavimo taisykles RAK-WB buvo leidžiama nediferencijuoti asmenų vardų. Kai Vokiečių biblioteka pradėjo kurti autoritetinius įrašus, šios praktikos buvo atsisakyta ir Vokiečių biblioteka daugiau nekuria nediferencijuotų vardų autoritetinių įrašų. Vis dėlto PND vis dar yra daug nediferencijuotų vardų. Vokiečių biblioteka, remdamasi sutaptimis tarp Kongreso bibliotekos ir Vokiečių bibliotekos antraščių, įrašytų į praplėstus autoritetinius įrašus, kiek įmanoma labiau automatiškai diferencijuos vardus su daug sutapčių; likusieji bus diferencijuojami rankiniu būdu. Pataisymai VIAF praturtins dažnais naujinimais ir padidins tikslių saitų tarp sutampančių įrašų skaičių.

Antra, daugelis Kongreso bibliotekos autoritetinių įrašų atspindi AACR2 praktiką, kuri leidžia sudaryti atskirus autoritetinius įrašus kiekvienam bibliografiniam identitetui, vartojamam vieno asmens, tokiam kaip slapyvardžiai. Tai visiškai kitoks atvejis nei PND nediferencijuoti įrašai. Šiuo atveju vienam asmeniui sudaroma daug autoritetinių įrašų. Pagal RAK-WB taisykles visiems vardų identitetams į PND įtrauktas tik vienas autoritetinis įrašas. Kaip ir nediferencijuoti vardai, šie bibliografinių identitetų autoritetiniai įrašai sukelia problemas, kurioms tinkamo sprendimo dar nerasta.

Susiję vardai gali būti tiesiogiai naudojami kaip automatinis vertimas iš Kongreso bibliotekos autoritetinio failo į PND arba atvirkščiai. Tai gali atitikti semantinio žiniatinklio arba jungtinių paieškos sistemų reikmes. Bendrųjų nuorodų „žr." jungtis gali leisti vartotojams peržiūrėti papildomą informaciją.

Dalyvaujančių šalių autoritetinių įrašų numeriai arba patys VIAF numeriai taip pat gali būti URI pagrindas. Tai suteiktų galimybę atpažinti URI autoritetiniame įraše. Vartotojui, pradedant bet kuria dokumente, įraše ar žiniatinklyje pateikta URI nuoroda, būtų prieinami visi dokumentai, įrašai, ištekliai ir t. t., susiję su URI pateiktais autoritetinių duomenų šaltiniais ir pačiais autoritetiniais įrašais.

Tolesnė sistemos plėtra

Nacionaliniai vardų autoritetiniai failai ir bibliografinių duomenų bazės nuolat keičiasi. Susietos duomenų bazės kuriamos dviejų ar daugiau kintančių failų pagrindu, saitai turi būti iš naujo įvertinti ir dažnai naujinami. Pradinės VIAF sistemos logika ir programinė įranga keičiama, kad nuolat būtų galima naujinti įrašus. Kai gaunami nauji bibliografiniai ar autoritetiniai įrašai, esantys praplėsti autoritetiniai įrašai tobulinami, ir vardų sutapties rezultatai iš naujo įvertinami. Nuolat bus gaunamos naujos sutaptys, o tos sutaptys, kurios nebus toliau palaikomos dėl pakitimų pradiniuose įrašuose, bus pašalinamos. Jei sutaptis pašalinama, ankstesnės sutapties istorija kiekviename sutampančiame įraše bus saugoma ir išliks prieinama.

4 pav. Praplėsto įrašo formatai

90x Kontroliniai numeriai

901	ISBN	$a ISBN skaitmenys (be kontrolinio skaitmens ir brūkšnelių)
902	ISSN	$a ISSN skaitmenys (be kontrolinio skaitmens ir brūkšnelių)
903	LCCN	$a LCCN skaitmenys (be kontrolinio skaitmens ir brūkšnelių)

91x Antraštės laukai

910	Antraštės forma iš 245 lauko	Polaukiai $a ir $b
911	Sutrumpinta antraštės forma iš 210 lauko	Polaukiai $a ir $b
913	Unifikuota antraštės forma iš 130 ar 240 lauko	Polaukiai $a ir $b
914	Išversta antraštė iš 242 lauko	Polaukiai $a ir $b
915	Siejamoji unifikuota antraštė iš 243 lauko	Visi polaukiai
916	Antraštės variantas iš 246 lauko	Polaukiai $a ir $b
917	Autoritetinio įrašo unifikuota antraštė	Paimta iš vardų ar antraščių autoritetinių įrašų, 100 laukas polaukis $t
919	Iš kito teksto paimta antraštė	Įvairios pastabos arba panašūs laukai

92x Leidėjo laukai

920	Leidėjo numeris	$a Leidėjo numeris iš ISBN
921	Leidėjo vardas	$a Leidėjo vardas iš 260 lauko polaukio $b ar 533 lauko polaukio $c
922	Publikavimo vieta	$a Šalies kodas iš 008 lauko

93x Vartojimas

930	Vardo vartojimas	$a Vardo forma iš atsakomybės duomenų, 245 lauko polaukis $c

94x Atributai

940	Kalba	$a Kalbos kodas iš 008 ar 041 lauko polaukio $a
941	Autoriaus vaidmuo	$a Santykio kodas iš 700 lauko, polaukių $e ir (arba) $4
942	NATC Subject	$a NATC survey line number
943	Publikavimo dekada	$a Publikavimo dekada
944	Formatas	$a Rūšis ir bibliografinis lygmuo (008/06-07)
945	Conspectus Subject	Įprastinis vartojimas, žr. PND diskusiją

95x Bendraautoriai

950	Individualieji autoriai	Polaukiai $a, $b, $c, $d ir $q iš 100 arba 700 laukų
951	Kolektyviniai autoriai	Polaukis $a iš 110 arba 710 laukų

96x Vardas kaip dalykas

960	Vardas kaip dalykas	Polaukiai $a, $b, $c, $d ir $q iš 600 lauko
969	Dalyko vartojimas	„Temos“ tekstas, žymintis autoritetinio įrašo pradmenį, buvo naudotas kaip dalykas ir paimtas iš 600 lauko

99x Specialūs laukai

999	Susiję bibliografiniai įrašai	$a Bendras įrašų skaičius $b Įrašo kontrolinis numeris $2 Įrašo šaltinis

Ateityje VIAF sistema pasinaudos šaltinių duomenų bazių OAI pranašumais, jei jos taps prieinamos. Tuo tarpu daugiau tradicinės failų prieigos priemonės, tokios kaip FTP, bus naudojamos projektui išbandyti.

Dėl didelės duomenų sankaupos vienoje vietoje gali būti numatyta daug įvairių duomenų gavimo ir panaudojimo būdų. Saitai, kaip semantinio žiniatinklio dalis, gali būti naudojami verčiant asmens vardą į galutinio vartotojo pageidaujamą formatą. Galėtų būti sukurtos priemonės, palaikančios automatinę paiešką alternatyviose bibliografinių duomenų bazėse, siekiant surasti tinkamą tai duomenų bazei vardo formą. Panašiai galėtų būti sukurtos katalogavimo ir autoritetinės kontrolės priemonės, kurios identifikuotų esančią įraše tinkamą vardo formą. Be abejo,VIAF duomenų bazėje taip pat bus galima atlikti ir tiesioginę paiešką.

Išvados

PND failui projektas jau davė apčiuopiamos naudos. Automatinis sutapčių patikrinimas abiejuose failuose leido gerokai patobulinti PND, ir Vokiečių biblioteka tikisi didelio palaikymo diferencijuojant vardus, nustatant praplėsto įrašo porose vienodas antraštes. Tapatinimo procesai ir algoritmai, sukurti projektui, pritaikomi daugelyje kitų programų. Buvo ištirtos galimybės, kaip panaudoti vardų tapatinimo duomenis tobulinant prieigą prie bibliografinės informacijos ir palaikyti dalyvių katalogavimo veiklą.

Projektas parodė praktines asmenų vardų dviejų nacionalinių autoritetinių failų automatinio siejimo galimybes. 70 proc. asmenų vardų autoritetinių įrašų, esančių abiejuose failuose, buvo susieti mažesniu nei 1 proc. klaidų dažniu. Strategija originalius autoritetinius įrašus papildyti duomenis iš bibliografinių įrašų gerokai pagerino sutapčių dažnį ir sumažino klaidingų sutapčių skaičių. Nedideli autoritetinių įrašų pakeitimai galėtų žymiai pagerinti tapatinimo rezultatus. Dėl nevykusiai gramatiškai sutvarkyto 670 lauko (Duomenų šaltinis) buvo gauta daug klaidingų sutapčių. Būtų naudinga papildoma struktūra, kuri leistų išvengti sutrumpintų vardų ir antraščių vartojimo arba kurioje būtų aiškiai parodyti pradinio bibliografinio įrašo saitai. Bendros veiklos arba specialybės (kompozitorius, iliustruotojas, matematikas ir t. t.) nustatymo aiškumas praplėstų tolesnį, tiek automatinį, tiek rankinį tapatinimą, siekiant įtraukti išsamesnes vardų formas dalinėse nuorodose.

Bibliotekoms tyrimas atvėrė plačias autoritetinės kontrolės, autoritetinių įrašų naudojimo, tinklo ir abipusės susieties ir semantinio žiniatinklio kūrimo galimybes. Toms Vokietijos bibliotekoms ir bibliotekų tinklams, kurie gauna ar kaupia bibliografinius įrašus su LCNAF kreipties elementais, VIAF taps platforma pereiti nuo vieno autoritetinio failo prie kito, taip pat pateikti LCNAF ir PND formų sankirtą arba LNCAF bibliografinių įrašų kreipties elementus perrašyti į PND kreipties elementus, arba suteikti galimybę per VIAF ieškoti ir gauti duomenis su PND pradmenimis. Pritaikytas tokiuose daugianacionaliniuose ir daugiakalbiuose portaluose, kaip Europos bibliotekos portalas (*European Library Portal*), VIAF galės automatiškai susieti paieškos užklausas abiejuose LCNAF ir PND failuose, leisdamas naudotis susietais abiejų šaltinių bibliografiniais įrašais.

Įdiegus tapatinimo technologiją, planuojama sukurti naujinamą sistemą, kurioje naudojant OAI galimybes bus kaupiami dalyvaujančių šalių dabartiniai asmenų vardų autoritetiniai ir bibliografiniai duomenys. Sukurta keičiamo dydžio sistema, todėl nauji dalyviai mielai kviečiami teikti savo autoritetinius ir bibliografinius duomenis. VIAF apimties didinimo ribos nebus aiškios tol, kol į projektą neįsitrauks daugiau įstaigų.

VIAF projekto dėmesio centre yra asmenų vardų autoritetinių įrašų tapatinimo problema. VIAF prižiūrėti, plėsti ir taikyti bus reikalingas ilgalaikių paslaugų teikimas ir valdymo strategija. Reikalingi sprendimai dėl projekto plėtros, įtraukiant kolektyvų vardus, unifikuotas antraštes ir t. t., taip pat naujas dalyvaujančias įstaigas. Ketinama plėsti sistemos pajėgumus, į ją įtraukiant Unikodo ženklų rinkinį. Unikodas leis įtraukti nelotyniškus rašmenis. Iššūkiu taps sutapties algoritmo išplėtimas, ypač tokioms rašto sistemoms, kaip korėjiečių, kinų ar japonų.

Iš anglų kalbos vertė N. Bliūdžiuvienė ir L. Buckienė

Straipsnis parengtas pagal pranešimą, skaitytą 2006 m. Seule (Korėja) vykusioje 72-ojoje IFLA konferencijoje.

[1] IFLA Core Activity: IFLA-CDNL Alliance for Bibliographic Standards (ICABS) http://www.ifla.org.sg/VI/7/icabs.htm [May 2006].

[2] Berners-Lee, Tim, Hendler, James, Lassila, Ora. The semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*. May 17, 2001. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C7084A9809EC588EF21 [May 2006].

[3] LEAF Project, http://www.leaf-eu.org [May 2006].

[4] Project InterParty: From Library Authority Files to E-Commerce, Andrew MacEwan, http://www.haworthpress.com/store/E-Text/View_EText.asp?a=3&fn=J104v39n01_11&i=1%2F2&s=J104&v=39 [May 2006].

[5] VIAF: The Virtual International Authority File, http://www.oclc.org/research/projects/viaf [May 2006].

[6] Open Archives Initiative - Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html [May 2006].

[7] Hickey, Thomas B., Toves, Jenny, O'Neill, Edward T. NACO normalization: a detailed examination of the Authority File Comparison Rules. *Library resources & technical services*. Vol. 50, no. 3, p. 18-24.