

Susieti duomenys ir bibliotekos

Jan HANNEMANN

Vokietijos nacionalinė biblioteka, Frankfurtas prie Maino, el. p. j.hannemann@d-nb.de

Jürgen KETT

Vokietijos nacionalinė biblioteka, Frankfurtas prie Maino, el. p. j.kett@d-nb.de

Semantinis saitynas ir ypač Atvirų duomenų sieties iniciatyva skatina institucijas skelbti savo duomenis, jais keistis ir juos susieti. Tai atveria plačias galimybes bibliotekoms – jos gali papildyti savo duomenis, susiedamos juos su kitais, išoriniais duomenų šaltiniais.

Šiame straipsnyje išsamiai aptariama pirmą kartą Vokietijos nacionalinės bibliotekos teikiama susietų atvirų duomenų paslauga. Daugiausia dėmesio skiriama iššūkiams, su kuriais susidurta steigiant šią paslaugą. Atsižvelgiant į sukauptą patirtį, straipsnyje aptariama keitimosi bibliotekų duomenimis perspektyva ir tarptautiniu lygiu susietų bibliotekų duomenų kūrimo galimybės. Nusakomi galimi šio proceso palengvinimo būdai ir naujos paslaugos, kurias būtų galima siūlyti besiplečiančių metaduomenų sankaupių pagrindu.

Reikšminiai žodžiai: semantinis saitynas; atvirų duomenų sieties iniciatyva; susietų duomenų paslaugos.

1. Įžanga

Šiuo metu keitimasis bibliotekų duomenimis yra didžiaja dalimi izoliuotas, nes bibliotekų kaupiami duomenys pirmiausia skirti tik bibliotekoms. Keitimosi duomenimis su kitomis institucijomis ir jų bendro naudojimo procesas dar tik pradedamas plėtoti. Šioje srityje bendradarbiauja daugiausia tik bibliotekos, o bibliotekų duomenys dar nėra integruoti į saityną. Taip yra todėl, kad bibliotekų duomenų rinkiniai ir kitų sričių duomenys nėra iki galo susieti, taip pat dėl dabar taikomų duomenų kaupimo procesų bei duomenų formatų, kurie, savaime suprantama, neperžengia klasikinės bibliotekose įprastos praktikos ribų.

Semantinis saitynas ir ypač Atvirų duomenų sieties iniciatyva skatina institucijas skelbti savo duomenis, jais keistis ir juos susieti saityne. Šiuos duomenis padaryti labiau matomus galima susiejant juos su kitais informacijos šaltiniais. Tai aktualu tiek ne pelno, tiek komercinėms institucijoms. Integracija į susietų duomenų tinklą, arba „semantinį debesį“, taip pat reiškia ir geresnį vartotojų interesų tenkinimą, pavyzdžiui, informacija nuolat pateikiama formatu, kuris suprantamas ir ne bibliotekų specialistams. Darbas su besiplečiančia semantinio debesies žinių baze gali padėti spręsti daugelį sudėtingų uždavinių, su kuriais dabar susiduria bibliotekos, prižiūradamos ir tobulindamos savo vietinius duomenų rinkinius. Tai tokie

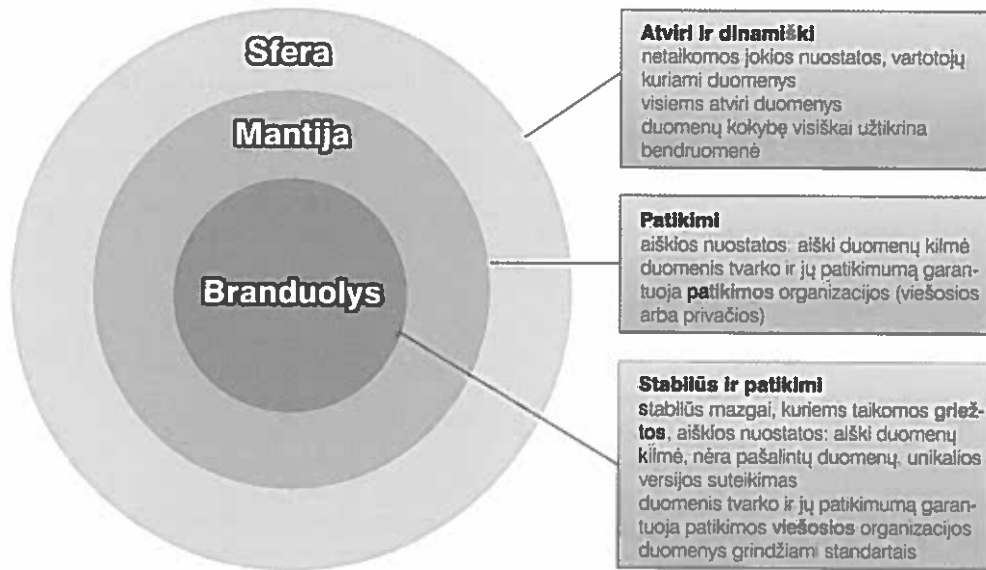
svarbūs uždaviniai, kaip besidubliuojančių duomenų aptikimas, dviprasmiškumo šalinimas, individualizavimas, duomenų kokybės valdymas ir jos gerinimas. Tai suteikia pagrindą atsirasti naujoms paslaugoms, naudojančioms daugiau negu vienos institucijos duomenis. Tokios bibliotekų ir kitų kultūros paveldo institucijų pastangos naudingos ir susietų duomenų bendruomenei. Bibliotekų duomenys paprastai yra labai aukštos kokybės, nes juos kaupia, peržiūri ir tvarko kvalifikuoti specialistai. Todėl jie gali tapti besiplečiančiam semantiniam saitynui taip reikalinga patikima atrama.

Bibliotekos suvokė šią galimybę; keletas institucijų ketina skelbti savo duomenis kaip „susietus duomenis“. Tačiau praktiškai tai įgyvendinti labai sudėtinga. Be organizacinių kliūčių, kultūros paveldo institucijoms, tokioms kaip bibliotekos (ypač toms, kurių IT biudžetas yra ribotas) gana problemiškas yra techninis duomenų skelbimo semantiniame saityne ir jau paskelbtų duomenų naudojimo aspektas.

Šio straipsnio tikslas – aptarti, kaip susieti duomenys susiję su bibliotekomis ir kitomis kultūros paveldo institucijomis bei išsamiai išdėstyti Vokietijos nacionalinės bibliotekos patirtį steigiant tokią paslaugą.

2. Vizija: globali kultūros schema

Pagrindinė susietų duomenų saityno problema yra duomenų patikimumas: ar duomenys teisingi ir ar sukur-



1 schema. Sluoksninis patikimumo ir stabilumo modelis

ti procesai, užtikrinantys aukštą duomenų kokybę? Kas už tai atsakingas? Lygiai taip pat svarbu ir patikimumas laikui bėgant: ar išteklius pakankamai stabilus, kad būtų galima jį teikti kaip nuorodą? Tai ypač svarbu mokslinių tyrimų srityje, kur nuorodos yra būtinas dalykas, ir aukštesniojo lygio paslaugoms, kurių pagrindą sudaro susieti duomenys.

Nors duomenų rinkinys gali būti naudingas ir nebūdamas visiškai patikimas, vis dėlto manome, kad susietų duomenų saityno pagrindą turi sudaryti stabilūs duomenys, kurie būtų patikima atrama, ir kad kultūros paveldo institucijos unikalios tuo, kad kai kurie šio pagrindo segmentai gali sudaryti būtent iš jų parengtų duomenų: sujungiant visas vietines kultūros paveldo institucijų žinių bazes įmanoma sukurti milžinišką globalią kultūros schemą, kurią sudarytų patikima ir stabili informacija.

Siekiant atspindėti skirtingus duomenų kokybės ir patikimumo lygmenis, globalios kultūros schemos duomenis geriausia vaizduoti pasitelkiant sluoksninį modelį (žr. 1 schemą). Kiekvienam modelio sluoksniui taikomi atitinkami metodai, kurie artėjant prie branduolio, t. y. patikimiausių žinių bazės, griežtėja. Branduolio elementai yra stabilūs ir patikimi, todėl gali būti teikiami kaip nuorodos. Kiekvienam šių elementų aprašui turi būti suteikiamas unikalus versijos numeris, kad būtų galima nuolat registruoti duomenų pokyčius ir jų kilmę. Kiekvienam sluoksniui taikomos aiškios nuostatos, grindžiamos priimtais ir oficialiai paskelbtais standartais ir katalogavimo taisyklėmis. Siekiant užtikrinti duomenų kokybę ir stabilumą, branduolio duomenų

patikimumą turi garantuoti viena arba kelios patikimos viešosios institucijos.

Šis modelis gali apimti net ir automatinio būdu sukurtus metaduomenis, jeigu tik dokumentiškai užfiksuota jų kilmė. Labai svarbu tai, kad informacija, tenkinanti būtinus reikalavimus, gali būti perkeliama arčiau branduolio. Taip patikimos informacijos branduolys, o tuo pačiu ir visos duomenų sancaupos vertė, laikui bėgant gali plėstis.

Susiklostė palankios sąlygos šios vizijos įgyvendinimui: kultūros paveldo institucijos jau taiko bendrai parengtus ir oficialiai paskelbtus standartus, tokius kaip MARC21 ar RAK-WB, bei duomenų kūrimą ir tvarkymą reglamentuojančias taisykles, nors keičiantis duomenimis saityne, būtina padaryti šių standartų ir taisyklių pakeitimus. Vokietijos bibliotekų bendruomenėje keičiantis duomenimis ir bendrai tvarkant centrinės duomenų bazes susiformavo darbų srautai, galintys būti tvirtu stabilių duomenų rinkinių teikimo saitynui pagrindu. Puikus to pavyzdys – bendrai tvarkomi vokiečių autoritetiniai failai asmenims, kolektyvams ir dalykinėms rubrikoms (atitinkamai PND, GKD ir SWD). Griežtai apibrėžti šalinimo, naujinimo ir dublikatų suliejimo procesai kartu su taikoma identifikavimo schema užtikrina aukštą stabilumo lygį. Todėl Vokietijos nacionalinė biblioteka, steigdama savo pirmąją susietų duomenų paslaugą, pasirinko būtent šiuos duomenų rinkinius.

Labai svarbus žingsnis įgyvendinant minėtą viziją yra savo vietinių žinių bazių, kaip susietų duomenų, skelbimas. Toliau apibūdinami šios veiklos iššūkiai ir mūsų patirtis.

3. Susietų duomenų paslaugos steigimo iššūkiai

Steigti susietų duomenų paslaugą nėra taip paprasta, kaip dažnai teigiama. Kultūros paveldo institucijos, siekiančios integruotis į semantinį debesį, susiduria su įvairiausiais sunkumais. Galima išskirti atskiras šių sunkumų kategorijas.

3.1. Techniniai

Norint įsteigti susietų duomenų paslaugą, būtina tam tikra infrastruktūra. Dažniausiai tai duomenų saugojimo priemonės (paprastai tai speciali RDF metaduomenims skirta („triplestore“) arba įprasta duomenų bazė), saityno serveris ir susiejimo sistema, interpretuojanti gaunamas saityno užklausas, nukreipianti jas į duomenų saugojimo vietą ir pateikianti rezultatą.

Susietų duomenų iniciatyvos yra santykinai naujas reiškinys, todėl dauguma jų technologinių sprendimų tebėra parengtinio lygmens ir ne visada būna išsamiai aprašyti. Institucijoms – susietų duomenų srities naujokėms – ne visada aišku, kuri iš daugybės technologinių priemonių geriausiai tiktų jų tikslams įgyvendinti.

3.2. Konceptualūs

Kita svarbi problema yra duomenų modeliavimas. Galima rinktis iš didelio skaičiaus daugiau ar mažiau pripažintų ontologijų, kurių kiekvienai būdingi savi privalumai ir trūkumai. Tokiame kontekste svarbu atsižvelgti į tai, kaip apibrėžiamos atskiros savybės: ar jos tinka modeliuojamiems duomenims? Jeigu nepavyksta išrinkti optimaliai tinkančios ontologijos, gali tecti taikyti keleto ontologijų derinį ir (arba) praplėsti jas sau pritaikytomis savybėmis.

Tam tikrą informaciją labai sunku modeliuoti, pavyzdžiui, duomenis apie duomenis. Pavyzdžiu galėtų būti tam tikrų duomenų kilmės arba procesų ar taisyklių, kurias taikant duomenys buvo sukurti, nurodymas. Pastarasis atvejis ypač svarbus duomenims, kurie sukurti naudojant automatinius algoritmus, tačiau kuriems būtina nurodyti taikytas katalogavimo taisykles ir standartus. Ši problema sprendžiama įvairiai: naudojant *N*-ary santykius¹, OWL 2 aksiomų anotacijas², reifikaciją³, kvadrupletus ir savo reikmėms pritaikytas ontologijų praplėstis. Kiekvienam sprendimui būdingi savi privalumai ir trūkumai. Šioje srityje nėra nustatytų pažangiausių metodų ir bendro sutarimo, kas labiausiai tinka bibliotekoms. Iš esmės iki šiol

trūksta bibliotekų duomenų modeliavimo taikant apibūdintus metodus patirties.

Kita problema susijusi su URI nustatymu. Paprastai savo duomenis skelbiančioms organizacijoms (ne bibliotekoms) priklauso izoliuoti duomenų rinkiniai, kuriems nepriskirti jokie vieši identifikatoriai ir netaikomi keitimosi duomenimis metodai. Nėra būtina įvedinėti naujus URI entitetams ir jų aprašams – bibliotekos savo duomenims ir aprašomiems entitetams naudoja daug viešų identifikatorių, keitimosi duomenimis tarp bibliotekų mastai irgi didžiuliai. Mes laikomės nuomonės, kad geriau neatskirti susietų duomenų nuo tradicinių keitimosi duomenimis metodų, o atvirkščiai – juos sujungti. Pasirinkta identifikavimo schema turėtų tikti visiems darbų srautams: identifikatoriai neturėtų skirtis, atsižvelgiant į naudojamą keitimosi duomenimis protokolą ar duomenų formatą (pvz., RDF ir MARC21).

3.3. Teisiniai

Svarbūs du su teisiniais klausimais susiję dalykai: publikavimo teisė ir susietų duomenų licencijos. Kultūros paveldo institucijos, tokios kaip bibliotekos, dažniausiai kaupia duomenis, paprastai bendradarbiaudamos su kitomis įstaigomis. Tokiu atveju turi būti aiškiai apibrėžiama, kokie duomenys gali būti skelbiami kaip susieti duomenys, t. y. tapti viešai prieinami.

Privatumo problema gali iškilti skelbiant tiek bibliografinius, tiek autoritetinius duomenis. Jeigu biblioteka kaupia informaciją apie autorius, gali būti taip, kad kai kurie autoriai nenorės, kad jų asmeninė informacija (gimimo data, gimimo arba gyvenamoji vieta, santykis su organizacijomis ir t. t.) būtų skelbiama.

Gali iškilti ir su licencijomis susijusių problemų. Turi būti iš anksto susitariama dėl susietų duomenų naudojimo sąlygų, nes teisinis tikrinimas užima laiko. Gali būti naudinga skirti komercinį ir nekomercinį naudojimą; pastaruoju atveju duomenys dažnai pateikiami neatlygintinai.

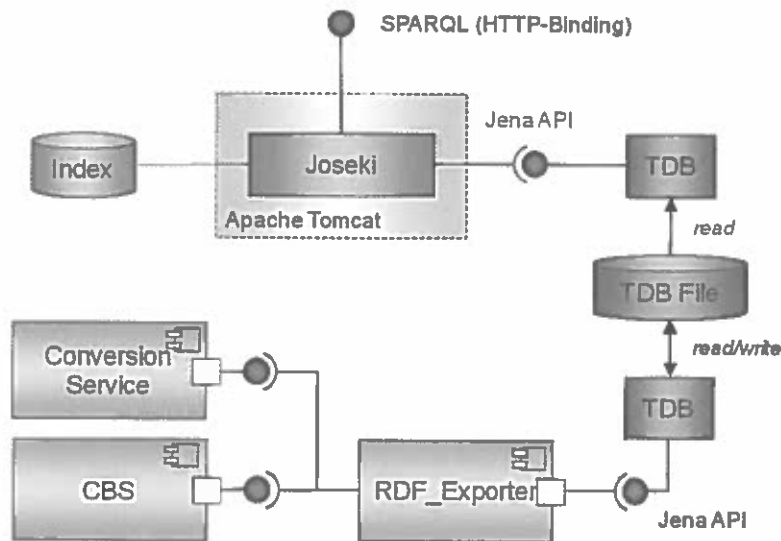
3.4. Bendri

Bendra visoms apibūdintoms kategorijoms yra informacijos apie susietų duomenų paslaugų steigimo patirtį (būtent šiam dalykui ir skirtas šis straipsnis) stygius. Beveik nėra nuoseklių patarimų, kaip tai padaryti; be to, juose neraskime atsakymų į daugelį klausimų. Šios srities pažangiausia praktika siūlo integruotis į semantinį saityną siekiančioms institucijoms taip reikalingą lankstumą, tačiau nepateikia tikslių standartams būdingų nurodymų.

¹ www.w3.org/TR/swbp-n-aryRelations

² www.w3.org/TR/owl2-syntax/#Axioms

³ [en.wikipedia.org/wiki/Reification_\(computer_science\)](http://en.wikipedia.org/wiki/Reification_(computer_science))



2 schema. Dabartinė sistemos struktūra

4. Vokietijos nacionalinės bibliotekos susietų duomenų paslauga

Šiame skyriuje aptariama Vokietijos nacionalinės bibliotekos patirtis steigiant pirmąją susietų duomenų paslaugą. Jai skirto projekto mastai specialiai nebuvo plečiami, kad būtų išvengta nepagrįstų planų ir siekių. Daugiausia dėmesio buvo kreipiama į atskirus bibliotekos duomenų rinkinių segmentus, o ne ieškota visiems duomenims tinkančių universalių sprendimų.

Techninei ir konceptualiai paslaugos plėtrai buvo pasirinktas iteracinis metodas, kurio kiekvienas ciklas truko 1–2 mėnesius ir kurį įgyvendinus kiekvieną kartą būdavo parengiamas vis išsamesnis funkcionalus paslaugos prototipas. Tai suteikė galimybę kolegoms ir galimiems vartotojams jau pradinio etapu teikti pasiūlymus, į kuriuos buvo atsižvelgta rengiant galutinį paslaugos projektą.

4.1. Techninis įgyvendinimas

Sistemos struktūrą vaizduoja 2 schema. Visi bibliotekos duomenys saugomi centrinėje duomenų bazėje, kurioje jie nuolat naujinami ir papildomi. Kai kuriuos duomenis pateikiant kaip susietus duomenis, programos modulis *RdfExporter* iš Bibliotekos centrinės katalogavimo sistemos⁴ išrenka metaduomenis, konvertuoja juos į RDF ir įkelia į *Jena TDB*⁵ RDF saugyklą. Kitu šio proceso etapu duomenys praturtinami nuorodomis į išorinius šaltinius. Būtiną duomenų transformavimą atlieka keletas konversi-

jos modulių (kurių kiekvienas skirtas tam tikram duomenų tipui, t. y. atskiri moduliai asmenims, kolektyvams ir dalykinėms rubrikoms), teikiamų bibliotekos Konversijos centro. Prieiga prie duomenų teikiama per *Joseki* serverį.

Buvo pasirinkta būtent tokia duomenų saugojimo ir prieigos technologija, nes ji buvo taikyta ankstesnio projekto metu. Reikia pasakyti, kad ši technologija tebėra aktyviai tobulinama, todėl nėra išsami ir jos dokumentacija.

Priimtas sprendimas pasiteisino, tačiau tikėtina, kad ateityje jis bus keičiamas. Ypatinę dėmesį būtina skirti išplečiamumui. Jau dabar kyla sunkumų dėl duomenų gausos: šiuolaikiniu kompiuteriu išsamiai konvertuoti tris duomenų rinkinius užima maždaug dvi dienas (be papildymo išorinėmis nuorodomis ir jomis praturtintų duomenų rinkinių sujungimo). *Joseki* serveris naudoja daug pagrindinės atminties (keletą gigabaitų; duomenų bazei plečiantis, šis skaičius didėja) ir kol kas neradome būdo, kaip to išvengti.

Kita problema susijusi su tuo, kad dabar taikoma technologija reikalauja rankiniu būdu pradėti duomenų transformavimą. Idealiu atveju duomenų konversija ir praturtinimas išorinėmis nuorodomis turėtų vykti pagal pageidavimą (t. y. gavus užklausą) arba būti susiejamas su automatinio naujinimo priemone, kuri pradėtų veikti įvykus reikšmingiems CBS duomenų bazėms pasikeitimams.

4.2. Duomenų atranka

Kuriant susietų duomenų paslaugą, labai svarbu atsakyti į klausimą, kokie duomenys bus skelbiami ir kokie

⁴ OCLC CBS: oclc.org/cbs/default.htm

⁵ openjena.org/TDB/

išoriniai duomenų šaltiniai bus siejami.

Mes nutarėme susikoncentruoti ties autoritetinių failų duomenimis, ypač informacija apie 1,8 mln. asmenų (iš PND), 160 000 dalykinių rubrikų (iš SWD) ir 1,3 mln. kolektyvų (iš GKD). Tokį pasirinkimą didžiaja dalimi sąlygojo išorinės minėtų duomenų užklauso, duomenų aprėptis ir jų savybės (žr. taip pat 2 skirsnį):

- duomenis jau naudoja daugelis organizacijų;
- duomenų tvarkymo darbų srutai gali būti taikomi ir skelbiant duomenis semantiniame saityne;
- sukurta ir pripažinta duomenų identifikavimo schema;
- susiedami su išoriniais duomenų šaltiniais, mes galėjome pasiremti keleto projektų rezultatais ir bibliotekų bendradarbiavimo derinant duomenų rinkinius patirtimi. Mes teikiame nuorodas į vokiškąją *Wikipedia*⁶ ir *DBpedia*⁷, VIAF⁸, LSCH⁹ ir RAMEAU¹⁰.

Mūsų bibliografinių duomenų aprėptis yra daug didesnė, todėl jiems numatome skirti būsimą projektą.

4.3. Ontologijų pasirinkimas

Semantinio saityno bendruomenė pataria kaip duomenų modeliavimo pagrindą naudoti pripažintas ontologijas, nes tai palengvina keitimąsi duomenimis. Todėl mes laikėmės būtent tokios nuostatos.

Tačiau iš tikrųjų paaiškėjo, kad esamos ontologijos tik dalinai tinka mūsų duomenims modeliuoti. Jų atskirų savybių apibrėžtys neatitinka mūsų duomenų rinkinių. Todėl nė viena iš esamų ontologijų nebuvo pasirinkta. Mums teko kruopščiai išrinkti keletą ontologijų ir tam tikras jų dalis kartu taikyti daugumai mūsų duomenų. Likusiems duomenims sukūrėme savas savybių apibrėžtis, tikėdamiesi, kad ateityje paskelbsime atskirą ontologiją.

Modeliuojant duomenis asmenų ir kolektyvų pateiktims naudojamos esamos ontologijos: RDA elementų rinkiniai, FOAF ir RELATIONSHIP žodynai; RDA sudaro mūsų duomenų modeliavimo pagrindą, nes jį galima taikyti FRBR entitetams (t. y. asmenims ir kolektyvams). Šias ontologijas papildėme Vokietijos nacionalinės bibliotekos sukurtomis savybėmis (*Gameinsame Normdatei* (GND) žodynu). Dalykinių rubrikų duomenų modeliavimas grindžiamas SKOS ir *Dublin Core* elementais, kuriuos irgi papildėme specialiomis GND savybėmis.

Išsamus mūsų pasirinktų ontologijų aptarimas ir galutinio pasirinkimo argumentai įtraukti į mūsų susietų duomenų paslaugos dokumentaciją¹¹.

4.4. Pavyzdžiai

Projekto veiklą iliustruoja šie pavyzdžiai:

– vokiečių autoriaus *Bertolt Brecht* (<http://d-nb.info/gnd/118514768>) XML/RDF pateiktį galima rasti <http://d-nb.info/gnd/118514768/about>

– kolektyvo *IFLA / Section of Public Libraries <The Hague>* XML/RDF pateiktį galima rasti <http://d-nb.info/gnd/10352988-3/about>

– dalykinę rubriką *Führungskraft* galima rasti <http://d-nb.info/gnd/4071497-4>, o jo XML/RDF pateiktį <http://d-nb.info/gnd/4071497-4/about>

5. Patirtis

Prieš pradėdami kurti savąją susietų duomenų paslaugą, kurį laiką dalyvavome semantinio saityno veikloje. Todėl manėme, kad esame pasirengę iššūkiams. Tačiau kad ir kokia teigiama būtų mūsų praktinė patirtis, joje būta ir ne visai sėkmingų dalykų bei klaidingų lūkesčių. Ypač svarbu nepriimti be išlygų semantinio saityno bendruomenėje paplitusios nuomonės, kad susijusių duomenų paslauga nesunkiai sukuriamą. Mūsų išvados būtų tokios:

– *Isteigti tokią paslaugą nėra paprasta*. Susijusių duomenų iniciatyvos yra palyginti naujas reiškinys. Todėl tam būtina pagrindinė programinė įranga (priemonės) dar nėra iki galo ištobulinta. Be kitų dalykų, tai reiškia, kad jų dokumentacija irgi gali būti neišsami. Kad paslauga veiktų, būtina įdiegti ir suderinti įvairius programinės įrangos komponentus (pvz., žr. 2 schemą), o tam reikia atitinkamos kvalifikacijos. Duomenys greičiausiai bus konvertuojami į atitinkamą formatą (RDF). Tam būtina ne tik tinkamai modeliuoti (transformuoti) duomenis (tai gali būti gana sudėtinga, jei būtina tiksliai sumodeliuotų ir originalių duomenų sutaptis), bet ir sukurti konvertavimo programas ar eksporto filtrus. Galutinei duomenų pateikčiai koduoti būtina naudoti UTF-8, net jeigu biblioteka savo vidaus reikmėms naudotų kitokią koduotę.

– *Duomenų modeliavimas gali būti sudėtingas procesas*. Skelbiant duomenis saityne, palanku naudoti esamas užregistruotas ontologijas. Deja, jos ne visada tinka atskirų bibliotekų duomenų pateiktims (žr. 3 skirsnį). Tiksliau tariant, gali smarkiai skirtis atskirų savybių apibrėžtys. Paprastai ši problema sprendžiama dviem būdais: tiesiog naudojamos esamos ontologijomis arba jos papildomos naujai apibrėžtomis savybėmis. Pirmasis būdas lengvesnis, tačiau gali iškraipyti duomenis;

⁶ de.wikipedia.org

⁷ wiki.dbpedia.org

⁸ viaf.org

⁹ authorities.loc.gov

¹⁰ rameau.bnf.fr

¹¹ wiki.d-nb.de/display/LDS

antrasis – daug sudėtingesnis, bet jį taikant tiksliai perteikiami duomenys. Nėra paprasta atsakyti, kuris būdas teisingiausias. Siekdami nepakenkti duomenų kokybei, mes pasirinkome antrąjį modelį, kaip labiausiai tinkantį mūsų duomenims.

– *Ne visur laikomasi palankaus požiūrio į atvirą keitimąsi duomenimis.* Net iki atsirandant susietiems duomenims, bibliotekos keitėsi savo duomenų rinkiniais ir juos derino. Tokių projektų rezultatai gali būti svarbūs informacijos šaltiniai saistant susietų duomenų rinkinius. Tik gaila, kad ne visų institucijų požiūris į atvirą keitimąsi yra palankus, o šių rezultatų skelbimą gali apsunkinti bendra duomenų nuosavybė. Šioje srityje mes įgavome tiek teigiamos, tiek neigiamos patirties, todėl rekomenduojame, kad bibliotekos, prieš pasinaudodamos duomenų keitimosi rezultatais, šį savo ketinimą išsamiai aptartų su visomis suinteresuotomis pusėmis.

– *Į pažangiąją praktiką žiūrima kaip į taisykles.* Susieti atviri duomenys didžiąja dalimi pagrįsti pažangiąja praktika, o ne taisyklėmis. Tačiau ne visa susietų duomenų bendruomenė šį pragmatinį aspektą laiko svarbiu. Nukrypimai nuo priimtų standartų dažnai kritikuojami, o tai institucijas – semantinio saityno naujokes – verčia abejoti savo sprendimų teisingumu, nors juos priimančiai organizacijai jie būtų ir prasmingi. Kritiški atsiliepimai neturėtų daryti įtakos bibliotekų iniciatyvai, atvirkščiai – tai turėtų skatinti jas skleisti savo patirtį ir žinias. Rekomendacijas ir pažangiąją praktiką būtina apsvaistyti atsižvelgiant į kiekvienos institucijos poreikius, ypač šiuo ankstyvuojų semantinio debesies formavimosi etapu. Mes sulaukėme kritikos dėl to, kad nesiūlome naudotis SPARQL (*SPARQL Protocol and RDF Query Language*) paslauga. Nors tai tikrai naudinga priemonė, ji jokiū būdu nelaikytina privaloma, ypač esant išbandytoms paieškos alternatyvoms (pvz., paieška naudojant URI¹² (SRU) ir *OpenSearch*¹³) ir bibliotekų duomenų sinchronizavimas (pvz., ORI¹⁴).

– *Dauguma vartotojų pasirenka anonimiškumą.* Siekdami patobulinti savo paslaugą, projekto pradžioje išskėlėme sau du esminius klausimus: kas yra mūsų duomenų vartotojai ir kokiais tikslais naudojami mūsų duomenys? Nors pakvietėme vartotojus teikti atsiliepimus apie paslaugą ir pasidalinti savo lūkesčiais bei patirtimi, atsiliepė tik nedidelė jų dalis. Susietų duomenų anoniminės prieigos koncepcijos taikymo pasekmė yra tai, kad mes žinome tik apie tuos vartotojus, kurie patys užmezga ryšį. Tai reiškia, kad mes neturime galimybės teikti pagalbos likusiems vartotojams. Mes tapome gana neįprasto reiškinio liudininkais: panašu, kad kažkas sukūrė programą, kuri atlieka mūsų susietų duomenų paiešką, taikydama

labai naivų metodą – išbando visus įmanomus skaitinius derinius, galinčius sudaryti IDN, kad sukurtų URI. Daug paprasčiau būtų šį vartotoją nukreipti į mūsų duomenis, skirtus parsisiuntimui, tačiau kadangi jis naudoja telefoniniu ryšiu, mes negalime jo (jos) identifikuoti.

– *Tinkamai sumodeliuoti duomenys yra labai vertingi.* Iki galo sumodeliuoti ir viešai pateikti duomenys tampa visuotinai prieinami. Mūsų kolegos iš Braunšveigo technikos universiteto pavyzdys rodo, kaip naudojant tokius duomenis sukuriama labai naudinga priemonė: per vieną dieną jis importavo mūsų duomenis į duomenų bazę, pritaikė saityno sąsają ir taip sukūrė prie jų paieškos funkcija aprūpintą prieigą.

Nepaisant išvardytų sunkumų, bendras vertinimas būtų teigiamas. Įveikus visas kliūtis, tampa palyginti nesunku pasitelkiant duomenis plėsti sukurtą paslaugą.

6. Ateities veikla

Sukurdami susietų duomenų paslaugą, žengėme žingsnį globalios kultūros schemos vizijos įgyvendinimo link; tačiau mūsų laukia tolesni etapai, kurių ilgalaikiams tikslams pasiekti būtinas kitų institucijų dalyvavimas. Kultūros paveldo organizacijoms, siekiančioms tenkinti saityno reikalavimus, būtina savo techninę infrastruktūrą, verslo procesus, taisykles, nuostatas ir licencijas pritaikyti prie savo metaduomenų. Šie esminiai pokyčiai turi būti realizuojami palaipsniui, taikant lanksčius metodus. Daugiausia dėmesio turi būti kreipiama į kultūros paveldo institucijoms tenkančius sunkumus, o įsibėgėjus semantinio debesies plėtrai gali būti vystomos ir globalinės strategijos.

Šiame skirsnyje pristatome mūsų ateities planus, taip pat ir toli siekiančius tikslus, kurių įgyvendinimas įmanomas tik bendradarbiaujant su kitomis institucijomis.

6.1. Artimiausi uždaviniai

– *Infrastruktūros permainos.* Artimiausias mūsų uždavinys – patobulinti veikiančios paslaugos infrastruktūros ir duomenų pateikimą, papildant ją naujais duomenų rinkiniais. Kita, naujesnė paslaugos versija bus aprūpinta automatinio duomenų ir galbūt nuorodų į kitus duomenų šaltinius naujinimo priemonėmis (žr. anksčiau). Reikės sukurti platesnėmis pritaikymo galimybėmis pasižyminčią paslaugos architektūrą. Siekdami praplėsti prieigą prie mūsų duomenų, sukursime SRU sąsają.

– *Nauji duomenų rinkiniai.* Bus taikomos papildomos klasifikacijos ir tobulinamas jau teikiamų duomenų modeliavimas. Be kitų dalykų, naujieji duomenų rinkiniai

¹² www.loc.gov/standards/sru/

¹³ www.opensearch.org

¹⁴ www.openarchives.org/OAI/openarchivesprotocol.html

apims rinktinį Vokietijos valstybinės bibliografijos duomenų poaibį ir vokišką Dewey dešimtainės klasifikacijos vertimą. Kalbant apie antraštinius duomenis, reikia apsispręsti dėl URI schemos naudojimo ir atlikti kai kuriuos su registracija susijusius veiksmus, kurie atitinka dabartinį keitimosi duomenimis Vokietijoje ir jų daugkartinio naudojimo optimizavimo planą. Šiuo metu dėl to tariamasi su Vokietijos bibliotekų centrais; tikimės, kad greitai laiku bus priimtas bendras nutarimas. Kita svarbi problema – antraštinių duomenų modeliavimas taikant ontologijas. Vėlgi būtinas kompromisas tarp esamų žodynų, tokių kaip *Dublin Core*¹⁵, ir bibliografinė ontologija¹⁶ naudojimo atsižvelgiant į atsirandančius naujus standartus (pvz., RDA) ir esamas mūsų duomenų struktūras.

– *Paslaugos galutiniam vartotojams*. Viena priežastis, kodėl mes pateikiame duomenis saityne, yra suteikti sau ir kitiems galimybę šių duomenų pagrindu kurti ir tobulinti paslaugas. Tai daryti turėtų ne vien kultūros paveldo institucijos – kūrybiškai dirbanti saityno bendruomenė gali pasiekti daugiau negu pavienės institucijos. Mūsų tikslas yra teikti bent vieną etaloninę paslaugą, demonstruojančią išsamų susietų bibliotekų duomenų galimybių spektrą. Tokios paslaugos tikslinės vartotojų grupės būtų dvi: ji skatintų kitas kultūros paveldo institucijas (ypač Vokietijoje) prisidėti prie globalios kultūros schemos plėtros ir atkreiptų saityno bendruomenės dėmesį į vertingus duomenų rinkinius. Paslaugos priemonė pagal tam tikrą entitetą (pvz., knygą, kūrinį, kolektyvą, temą ir t. t.) URI suranda tinklavietę su nuorodomis į visus registruotus duomenų rinkinius, kurie (tiesiogiai) susieti su išoriniais šaltiniais.

6.2. Ilgalaikiai tikslai

Neabejotina tai, kad ilgalaikiai tikslai įgyvendinami tik bendrai veikiant kultūros paveldo organizacijoms, programinės įrangos pardavėjams, mokslo įstaigoms ir valstybės institucijoms. Vokietijos nacionalinė biblioteka kartu su partneriais pasirengusi išnaudoti savo įtaką ir techninę kompetenciją tolesnei šios srities plėtrai. Būtiniausi uždaviniai yra šie:

– *Bendras licencijavimo modelis*. Plečiasi bibliotekų kartu atliekamo indeksavimo mastai. Vokietijos bibliotekas, muziejus ir archyvus dar labiau suartins numatoma sukurti *Deutsche Digitale Bibliothek*¹⁷. Bendrai kuriamų metaduomenų kontekste taikyti ir įgyvendinti tinkamus licencijavimo modelius įmanoma tik bendromis pastangomis. Jeigu tai neįmanoma, labai sumažėja daugkartinis semantinio debesies duomenų naudojimas – gali net prireikti apdoroti licencijuotus duomenis ne įrašo, o pa-

vienių metaduomenų laukų lygmeniu (pavyzdžiui, vienai institucijai praturtinus kitos institucijos duomenis). Praktiškai tai būtų per daug sudėtinga ir nerentabili.

– *Darby srautų ir nuostatų priėmimas*. Kultūros paveldo organizacijos turi patvirtinti jų atliekamų veiksmų eigą; tai padėtų sumažinti perteklinių duomenų skaičių ir sustiprintų jų ilgalaikiškumą (ypač antraštinių duomenų kūrinio ir apraiškos lygmenimis). Būtinai viešai nurodomi identifikatoriai, kuriems būtų bendrai pritarta ir kurie būtų bendrai naudojami. Tokių iniciatyvų modeliu galėtų būti vokiečių *Gemeinsame Normdatei* (GND).

– *Praktine patirtimi paremti projektiniai ontologijų pasiūlymai*. Kultūros paveldo institucijos nepajėgios pačios išspręsti su duomenų modeliavimu (pvz., duomenų apie duomenis) ir techniniu jo įgyvendinimu susijusių problemų, tačiau jos gali teikti praktika grindžiamus pasiūlymus. Tinkamus sprendimus priimtų programinės įrangos pardavėjai ir mokslo įstaigos.

– *Pažangūs techniniai sprendimai*. Šiuolaikiška techninė bazė nėra visai tinkama kultūros paveldo organizacijų produktyviai veiklai. Būtina sukurti priemones ir sistemas, kurios palengvintų duomenų pateikimą semantiniame debesyje ir jau pateiktų duomenų naudojimą. Susietų duomenų technologija turi tapti integralia esamos aplinkos ir veiklos procesų dalimi. Todėl naujovės turi nepriklausyti nuo atskirų platformų, atitikti atvirojus standartus ir būti daugiamodulinės lygmeninės struktūros, kuriai būdinga atviroji API (programų sąsaja). Lygiagrečiai turi plėtotis ir esamos bibliotekų sistemos: ateityje jos turės sugebėti apdoroti daugialyę kilmę ir versiją identifikuojančią informaciją.

7. Santrauka

Atvirų duomenų sietis kultūros paveldo institucijoms, įskaitant bibliotekas, atveria didžiules galimybes. Tai ne tik plečia duomenų vartotojų ratą, bet, susiejant duomenų rinkinius ir sudarant jų semantinius debesis, išauga pačių duomenų vertė. Šiuo pagrindu bibliotekos ir kitos organizacijos gali steigti naujoviškas ir visapusiškas paslaugas. Įsteigdami susietų duomenų paslaugą, mes žengėme žingsnį globalios kultūros schemos vizijos įgyvendinimo link, tačiau ilgalaikiai šios vizijos tikslai pasiekiami tik sulaukus visuotinio šios koncepcijos palaikymo.

Iš anglų kalbos vertė T. Auškalnis

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

¹⁵ <http://dublincore.org/documents/dc-rdf/>

¹⁶ <http://bibliontology.com/>

¹⁷ www.deutsche-digitale-bibliothek.de