UDK 004.65:004.774.2

CONTENTUS – link semantinių multimedijos bibliotekų

Jan NANDZIK
Acosta Consult, Frankfurtas prie Maino, Vokietija, el. p. jn@acosta-consult.de
Andreas HEß
Vokietijos nacionalinė biblioteka, Frankfurtas prie Maino, el. p. a.hess@d-nb.de
Jan HANNEMANN
Vokietijos nacionalinė biblioteka, Frankfurtas prie Maino, el. p. j.hannemann@d-nb.de
Nicolas FLORES-HERR
Acosta Consult, Frankfurtas prie Maino, Vokietija, el. p. nf@acosta-consult.de
Klaus BOSSERT
Acosta Consult, Frankfurtas prie Maino, Vokietija, el. p. kb@acosta-consult.de

Vis plečiantis internete skelbiamam turiniui ir žinioms, bibliotekoms atsiveria galimybės teikti savo duomenis ir naujoviškai pristatyti savo rinkinius. Konceptualiai gimininga informaciją galima sieti semantiškai, taip praturtinant vartotojus gausesniais duomenų rinkiniais ir naujoviškomis paieškos galimybėmis, atsirandančiomis dėl medijoms, vietiniams metaduomenims ir išoriniams informacijos šaltiniams būdingų tarpusavio ryšių.

Šiame straipsnyje pristatomi CONTENTUS projekte plėtojami bibliotekoms ir multimedijos archyvams skirti sprendimai, susiję su įvairialypių duomenų šaltinių integravimo ir inovatyvių semantinės paieškos koncepcijų teikimo iššūkiais.

Reikšminiai žodžiai: multimedijos rinkiniai; metaduomenys.

Įvadas

Vokietijoje apie 30 000 kultūros institucijų yra sukaupusios neįtikėtinai turtingą multimedijos archyvą įvairiomis laikmenomis – knygų, vaizdų, juostų ir filmų. Šioms kultūros paveldo organizacijoms iškyla būtinybė teikti visuomenei interneto prieigą prie žinių, sukauptų gausiuose multimedijos rinkiniuose. Technologinį vartotojams skirtos prieigos prie skaitmeninių rinkinių pagrindą ateityje sudarys naujoviškos semantinės multimedijos paieškos paslaugos. Šių technologijų taikymo išankstinė sąlyga yra *bibliografinių metaduomenų, automatiškai kuriamų metaduomenų* ir *išorinių informacijos išteklių integravimas* žinių bazėje. Straipsnio tema – metodologinė ir techninė šios srities plėtra, įgyvendinama Vokietijos nacionalinės bibliotekos ir partnerių CONTENTUS projekto kontekste[1].

Paieška multimedijos rinkiniuose: kultūros paveldo organizacijoms kylantys iššūkiai

Siekiant įgalinti semantinę multimedijos paiešką, gausūs tokios informacijos rinkiniai turi būti aprūpinti pakankamu skaičiumi tinkamų aprašomųjų metaduomenų. Tačiau iki šiol dauguma multimedijos objektų anotuojami ir kataloguojami rankiniu būdu informacijos specialistų. Kadangi didesniojoje dalyje tokių išteklių, pvz., garso ir vaizdo medijoje, gausu informacijos, rankiniu būdu kurti metaduomenis labai sudėtinga, brangu ir užima daug laiko.

Rankiniu būdu indeksuojant stambius nestruktūruotus multimedijos failus, faktiškai neįmanoma objektą išsamiai aprašyti arba pavyksta išsamiai indeksuoti tik jo fragmentus. Su tokia nepageidaujama situacija dažnai susiduria kultūros paveldo organizacijos, nes nepakanka žmogiškųjų išteklių norint apimti sparčiai

[1] http://www.theseus-programm.de/en-us/theseus-application-scenarios/contentus

I schema. CONTENTUS apdorojimo grandinė

gausėjančius multimedijos rinkinius.

Atsiradus atitinkamiems svarbiems interneto ištekliams (pvz., Vikipedijai ar *GeoNames*) bei bendrai prižiūrimiems duomenų rinkiniams, tokiems kaip autoritetiniai failai, multimedijos objektų indeksavimas neturėtų apsiriboti vien tik jų turinio aprašymu. Kadangi šie išoriniai ištekliai potencialiai gali semantiškai praturtinti rinkinius sudarančius vienetus, multimedijos objektai kaip visuma arba tokį objektą sudarantys atskiri entitetai (pvz., asmenys, vietos, įvykiai) turi būti semantiškai susiejami su atitinkamais internetiniais duomenų rinkiniais, taip papildant jau esamus metaduomenis.

Vis dėlto indeksuoti multimediją ir susieti jos rinkinių entitetus su atitinkamais išoriniais ištekliais yra nelengva ir tai netapo kultūros paveldo organizacijų įprasta praktika – tai yra vienas iš CONTENTUS projekto sprendžiamų uždavinių.

CONTENTUS vizija: naujos kartos multimedijos bibliotekos

CONTENTUS – tai mokslinės ir technologinės plėtros projektas, kuriam vadovauja Vokietijos nacionalinė biblioteka ir kuri yra Vokietijos vyriausybės finansuojamos mokslinės iniciatyvos THESEUS dalis[2]. Jis siūlo kultūros institucijoms ir kitiems turinio valdytojams visą paketą sprendimų, palengvinančių vientisą perėjimą iš neapdorotų skaitmeninių duomenų į semantinę multimedijinę paieškos terpę [Bossert, Flores-Herr, Hannemann, 2009].

CONTENTUS modelio ir projekto metu plėtojamų metodologijų bei koncepcijų pagrindu kuriama sistema, leidžianti kultūros institucijoms teikti galutiniams vartotojams plačią prieigą prie multimedijos rinkinių. Galutiniams vartotojams siūlomos inovacinės paieškos pasirinktys, kurių atsiradimą skatina multimedijos objektų gausa ir metaduomenys iš įvairių šaltinių, tarp jų – „tradiciniai“, intelektiniu būdu sudaromi duomenys, automatiškai kuriama informacija ir interneto ištekliai.

Siekiant apimti bendruomenes, kuriančias ir tobulinančias semantinių žinių tinklus, CONTENTUS vykdomas glaudžiai bendradarbiaujant su ALEXANDRIA ir *Mediaglobe* – THESEUS projektais, skirtais *Web 3.0* ir medijų technologijų plėtrai. Tokia bendra veikla galiausiai padės sukurti atvirųjų žinių tinklus, kuriuose bus galima kultūros paveldo institucijų multimedijos objektus susieti su socialinio saityno ištekliais: *naujos kartos multimedijos bibliotekomis.*

CONTENTUS siekia sukurti kultūros paveldo organizacijų infrastruktūrą, leidžiančią veiksmingai apdoroti stambius multimedijos rinkinius ir susieti juos su išoriniais metaduomenų ištekliais. Atskiros projekto pakopos sudaro apdorojimo grandinę, kaip parodyta I schemoje.

1. *Skaitmeninimas.* Daugelis archyvų tebėra analoginės formos, todėl objektų apdorojimas neįmanomas. Tokių archyvų atveju pirmasis medijų atvėrimo išsamiai semantinei paieškai žingsnis yra stambiu mastu atliekamas skaitmeninimas.

2. *Kokybės kontrolė.* Siekiant neatsilikti nuo šiuolaikinių skaitmeninimo aparatų (pvz., knygų skenavimo robotų) tempo, būtina automatinė kokybės analizė (pvz., knygos puslapių skenavimo kokybės patikra). Šiuo atveju siekiama geresnės vartotojui ir turinio analizei skirtos (žr. toliau) medijų kokybės.

3. *Turinio analizė.* Kartais nepakanka rankiniu būdu sudaromų metaduomenų aprašų, kad būtų įmanoma rezultatyvi multimedijos objektų paieška. Kita vertus, išsamiam teksto, garso ir vaizdo turinio anotavimui, pvz., transkribuojant pasisakymus ar indeksuojant žinių transliacijas, reikia didžiulių žmogiškųjų pastangų ir finansinių išteklių. CONTENTUS projekto ribose siekiant palengvinti su paieška susijusios informacijos kūrimą, svarbų vaidmenį atlieka paslaugos, automatiškai analizuojančios įprastus multimedijos išteklius, tokius kaip vaizdai, muzikos arba vaizdo įrašai.

4. *Semantinė sietis.* Siekiant praturtinti esamus metaduomenis, automatiškai sukurta informacija gali būti susiejama su bibliografiniais metaduomenimis ir interneto ištekliais. Pavyzdžiui, dokumentinio filmo autorius gali būti taip pat ir knygos, kuri savo ruožtu susieta su Vikipedijos straipsniu ar autoritetinio failo įvediniu, autorius. Be to, panaikinamas išvestinių entitetų, tokių kaip vietos, asmenys, įvykiai ir t. t., homonimiškumas (pvz., automatiškai atskiriamos sąvokos *apple* (obuolys) kaip vaisius ir *Apple* kaip korporacija) ir jie susiejami su *susietų atvirų duomenų debesimi*, t. y. gausių informacijos šaltinių, saityne pateiktų kaip susiję duomenys, visuma.

[2] http://www.theseus-programm.de

5. *Atvirų žinių tinklai.* Šioje pakopoje multimedijos objektai išorinių bendruomenių gali būti toliau papildomi išoriniais ištekliais.

6. *Semantinė paieška.* CONTENTUS siūlo galutiniams vartotojams inovacines multimedijos paieškos funkcijas suvienijant tekstų, vaizdų, garso, garso ir vaizdo turinio paieškos galimybes bendroje semantinėje vartotojo sąsajoje.

Toliau aptarsime duomenų integravimo iššūkius ir semantinės paieškos tobulinimo koncepcijas.

Duomenų integravimas

Vienas iš CONTENTUS projekto sunkumų yra būtinybė integruoti duomenis ir metaduomenis iš įvairių skirtingų šaltinių. Tokie duomenys paprastai apima skaitmeninimo veiklos produktus, sukurtus kaip skaitmeniniai dokumentus, vartotojų pateiktą informaciją ir daugelį kitų dalykų. Kadangi mūsų siekis yra suteikti prieigą prie skirtingų šaltinių multimedijos turinio per integruotą sąsają, sunkumų kelia atitinkamų metaduomenų integravimas ir derinimas. Kitaip negu tradicinėse katalogavimo sistemose, mes praturtiname duomenis išoriniais šaltiniais, nes tikime, kad šių šaltinių informacija naudinga vartotojui. Netgi jei išorinių metaduomenų kokybė kartais (bet nebūtinai) žemesnė už kokybę, kurios paprastai tikimasi iš bibliotekininkų kuriamų metaduomenų, jie, būdami išsamesni arba apimdami aspektus, į kuriuos nebuvo atsižvelgta, gali papildyti esamus duomenis. Tarkime, *Deutsches Musikarchiv*[3], kuriame saugomas pagrindinis Vokietijos natų leidinių ir garso įrašų rinkinys ir kuris atlieka su muzika susijusios bibliografinės informacijos centro funkciją, katalogas neapima atskirų dainų ar garso takelių. Susiejus katalogo duomenis su muzikos duomenų baze, vartotojas gali pasiekti išsamesnę informaciją, pvz., garso takelių sąrašus.

Bet kokių duomenų paslaugų, integruojančių skirtingų šaltinių informaciją, atveju būtina skirti du naudojimo aspektus:

1) siejimą;

2) integravimą.

Pirmuoju atveju skirtingų šaltinių metaduomenys nekaupiami toje pačioje duomenų bazėje, o tik laisvai susiejami. Šaltinių, kurie nepavaldūs paslaugos teikėjui, metaduomenys naudojami tik prireikus. Tokio metodo privalumas yra tai, kad paslaugos vartotojui pateikiami (meta) duomenys visada yra kiek įmanoma atnaujinti, netgi jei jie yra iš nepavaldžių paslaugos teikėjui šaltinių. Trūkumas būtų tai, kad niekas negali garantuoti, jog išoriniai šaltiniai bus visada prieinami.

Antruoju atveju visų šaltinių metaduomenys integruojami toje pačioje paslaugos teikėjui priklausančioje duomenų bazėje arba ontologinių duomenų saugykloje. Todėl jų prieinamumas priklauso tik nuo paties teikėjo paslaugų sistemos. Tačiau būtina įdiegti naujinimo metodiką, kad būtų užtikrinama, jog paslaugos vartotojui pateikiami duomenys nėra pasenę. Kitas aspektas, į kurį būtina atsižvelgti, yra licencijavimas, nes kitų šaltinių duomenys ne tik naudojami, bet ir kopijuojami.

Abiem atvejais pagrindinis techninis iššūkis yra rasti atitiktį tarp skirtingų schemų. Tai galima padaryti:

1) rankiniu / intelektiniu būdu;

2) automatiškai / pusiau automatiškai.

CONTENTUS taikomi abu būdai, priklausomai nuo duomenų šaltinio. Būtina pažymėti, kad ir patys duomenų šaltiniai gali būti kuriami intelektiniu arba automatiniu būdu. Pavyzdžiui, skenuojamuose tekstuose pasitaikantiems asmenims, organizacijoms, vietoms ir temoms rasti bei jų homonimiškumui pašalinti mes taikome automatinius informacijos išskyrimo algoritmus, tačiau inkorporuojame ir intelektiniu būdu sukurtų autoritetinių sąrašų duomenis.

Mūsų dabartinė sistema integruoja metaduomenis iš tokių šaltinių:

– Vokietijos nacionalinės bibliotekos: autoritetiniai failai ir katalogų duomenys;

– Vikipedijos: asmenų nuotraukos (planuojama: papildoma „foninė“ informacija apie asmenis ir vietoves);

– *MusicBrainz*: kompaktinių diskų takelių sąrašai;

– automatiškai išskiriama: asmenys, organizacijos, vietos ir temos iš tekstų ir garso įrašų, muzikos takelių panašumai.

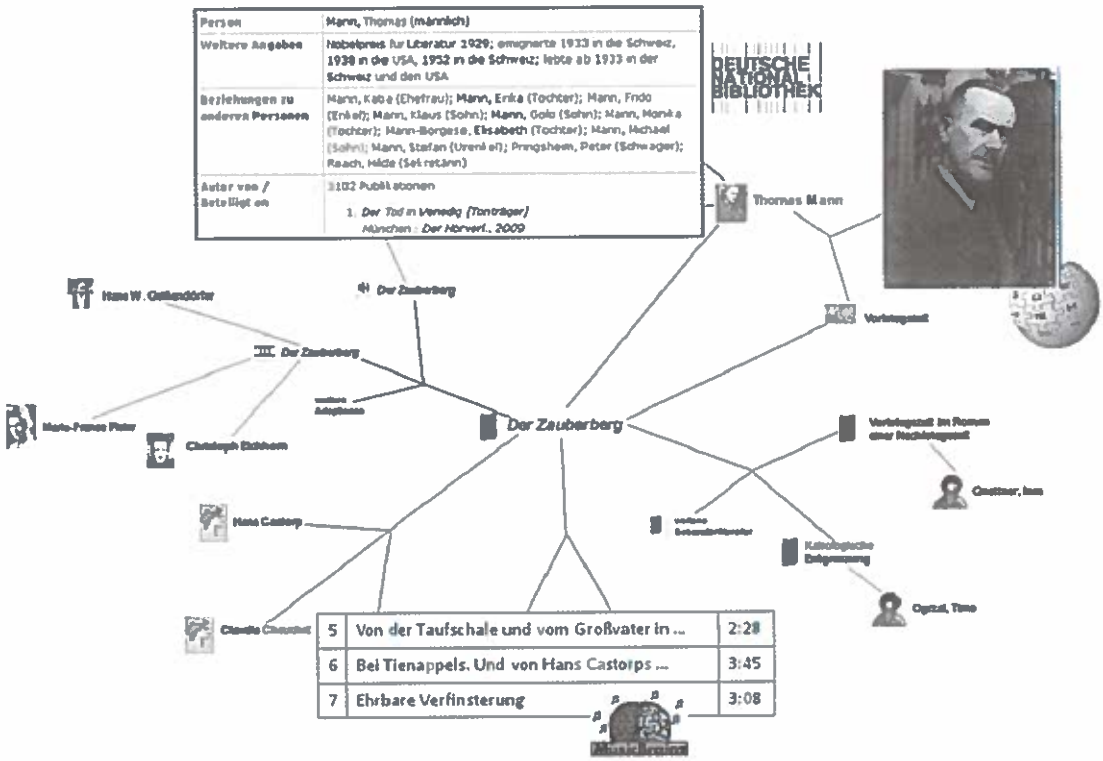
Autoritetiniai ir katalogų duomenys naudojami kaip pamatinė informacija. Vikipedijos ir autoritetinių failų sutaptis nustato savanoriai rankiniu būdu. Ši sutaptis jau naudojama vokiškojoje Vikipedijoje ir Vokietijos nacionalinės bibliotekos katalogų sistemoje. Šiuo metu rankiniu būdu rengiama ir *MusicBrainz* garso takelių sąrašų bei autoritetinių failų sutaptis.

Skirtingų šaltinių duomenų integracija CONTENTUS vaizduojama 2 schemoje. Joje matyti, kad CONTENTUS saugomi metaduomenys susisteminti tinkle susiejant autorius ir kūrinius bei papildomą informaciją, pvz., apie atitinkamas vietoves, temas ar istorinius laikotarpius.

URI naudojimas

Atsižvelgiant į susietų atvirų duomenų principus, rekomenduojama kaip identifikatorius naudoti susietiems duomenims skirtus ir nuolatinius URI. URI leidžia su-

[3] Vokietijos muzikos archyvas buvo įkurtas 1970 m., jis tęsia Vokietijos muzikos fonotekos (*Deutsche Musik-Phonothek*, 1961–1969) veiklą. *Musikarchiv* yra Vokietijos nacionalinės bibliotekos padalinys Leipcige. Žr. taip pat: http://www.d-nb.de/eng/sammlungen/dma/samml_bestaende



2 schema. Skirtingų šaltinių duomenų apie vokiečių autoriaus Tomo Mano kūrinį „Der Zauberberg“ integravimas

jungti skirtingus informacijos šaltinius, susijusius su mus dominančiais entitetais (asmenimis, vietomis, organizacijomis ir t. t.), todėl jie ypatingai svarbūs derinant duomenų rinkinius ir integruojant įvairiarūšių duomenų šaltinių informaciją CONTENTUS žinių bazėje.

Asmenų vardų homonimiškumo šalinimas

Siekiant patobulinti paieškos funkcionalumą, svarbu susieti medijas, pvz., tekstinius dokumentus, su kita informacija, tokia kaip autoritetinių failų duomenys. Šiuo metu tai atliekama asmenims, kolektyvams, vietoms ir t. t. Vienas sunkumų, su kuriais susiduriama, yra identifikuoti asmenį esant keletui asmenų tokiu pačiu vardu. Taipogi kartais neįmanoma iš pirmo žvilgsnio atskirti asmenų vardų nuo kitų žodyne pateikiamų žodžių. CONTENTUS projekte su autoritetiniais failais susiejami ne tik dokumentų autoriai ar kiti metaduomenyse minimi asmenys, bet ir asmenys, minimi pačiuose iš dokumentų išskirtuose tekstuose. Tam būtina taikyti automatinius algoritmus – asmenų vardų išskyrimui ir jų homonimiškumo pašalinimui.

Šis CONTENTUS taikomas metodas sukurtas Pilz ir Paaß [Pilz and Paaß, 2009]. Norint pašalinti asmens vardų homonimiškumą, lyginamas vardo kontekstas ir dokumentas, kuriame identifikuojama tikroji asmens tapatybė. Mūsų pavyzdyje tekstas lyginamas su šio asmens aprašymais Vikipedijoje. Asmuo, apie kurį Vikipedijos straipsnis labiausiai tinka atitinkamam dokumento tekstui, identifikuojamas kaip asmuo, minimas originaliame dokumente.

Objektų ir schemų sutaptis

Sutapčių tarp įvairiarūšių duomenų rinkinių rengimas yra viena iš seniausių informacijos mokslo problemų. Čia reikia skirti du dalykus: objektų sutaptis ir duomenų struktūrų sutaptis. Automatinio objektų sutapčių rengimo problemos sprendimas prilygsta dubletų aptikimui. Įprasta kaip identiškų objektų nustatymo priemones taikyti atstumo tarp eilučių metrikas, pvz., plačiai žinomas Levenshtein [Levenshtein, 1965] arba Jaro-Winkler [Winkler, 1999] metrikas, ir (arba) fonetinio panašumo matavimo algoritmus, tokius kaip *Soundex* [Russell, 1918]. Faktiškai šiuolaikiniai atitikties nustatymo algoritmai naudoja keleto metrikų derinį (žr., pvz., [Johnston and Kushmerick, 2004]).

Moksliniuose darbuose ir literatūroje automatinio schemų derinimo problema nagrinėjama nuo to laiko, kai buvo pradėtos kurti duomenų bazės [Melnik et al., 2002], ji iškyla derinant XML schemas, ir visai neseniai – derinant ontologijas [Shvaiko et al.; 2009, Heß, 2006]. Taikant algoritmus paprastai remiamasi struktūriniais ir

leksiniais panašumais, kai kuriais atvejais – žinomomis objektų pateiktimis abiejose derinamose schemose.

Vietovių sutapties nustatymas

Kai kada galima pasikliauti intelektiniu būdu parengtomis schemų ar objektų sutaptimis (žr. anksčiau), nes jos buvo kuriamos kolektyviai (Vikipedijos atveju) arba jų rengimas yra nesudėtingas (*MusicBrainz*). Tačiau atliekant sudėtingesnes užduotis, esminis dalykas yra galimybė taikyti tikslius automatinius sutapčių rengimo algoritmus.

Tolesnei CONTENTUS semantinės paieškos plėtrai planuojama įdiegti grafines valdymo priemones (žr. kitą skirsnį): geografinei informacijai apie vietoves, kurios aptinkamos, pavyzdžiui, visatekščiuose medijų dokumentuose arba sietis su kuriomis įgalinama per metaduomenis, atvaizduoti. Siekiama įtraukti sutaptis su geografinių duomenų bazėmis, tokiomis kaip *GeoNames*.

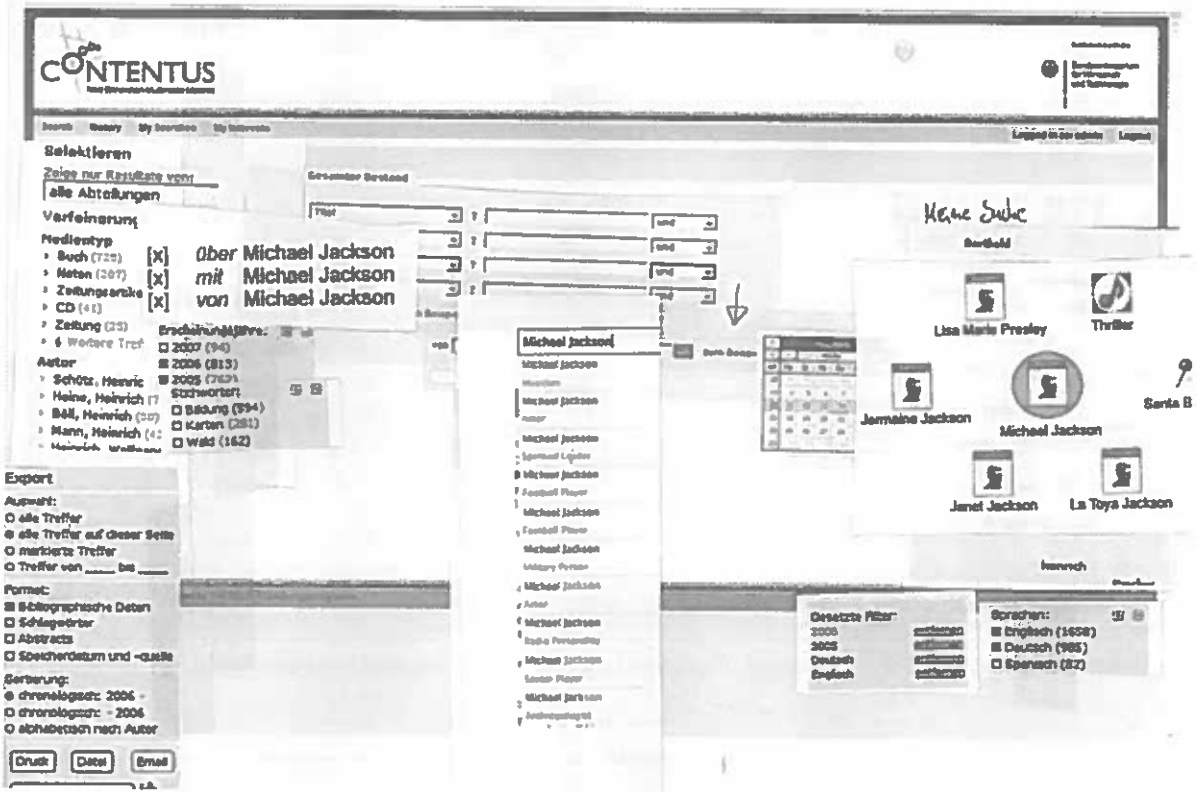
Šiam tikslui ketinama derinti euristinius metodus ir panašumo nustatymo metrikas. Autoritetiniame faile, sudarančiame sutapties pagrindą, paprastai pateikiama informacija apie šalį ir (jei tokia yra) federalinę valstiją arba sritį, kurioje yra miestas. Miesto vardui nesant unikaliam (pvz., Paryžius Teksase, JAV, ir Paryžius Prancūzijoje), šią informaciją galima naudoti homonimiškumui pašalinti. Panašūs metodai buvo sėkmingai taikomi autoritetinių failų informacijos homonimiškumui šalinti Vokietijos nacionalinės bibliotekos pirmojo susietų duomenų projekto kontekste [Hannemann et al., 2010].

Paieška ir orientavimas

CONTENTUS projekto metu sukurta paieškos sistema jungia du informacijos šaltinius: tradicinę visatekstę optiškai atpažintų tekstų ir garsinių tekstų transkripcijų rodyklę, taipogi ontologijų semantinę informaciją. Pagrindinės šios Semantinės multimedijos paieškos sistemos (*Semantic Multi-Media Search* – SMMS) medijos yra garso, garso ir vaizdo medžiaga, skenuota spausdinta medija, sukurti kaip skaitmeniniai tekstiniai dokumentai.

CONTENTUS paieškos tikslas – užtikrinti prieigą prie visų šių informacijos šaltinių per bendrą sąsają. Todėl kuriant vartotojo sąsają (UI – *User Interface*) sunkiausiai sprendžiami toliau išvardyti uždaviniai.

1. Aiškus skirtingų duomenų šaltinių derinimas.
2. Vientisas multimedijos duomenų ir metaduomenų integravimas.
3. Vartotojui palanki prieiga prie semantinės paieškos priemonių.



3 schema. CONTENTUS popierinio prototipavimo pavyzdys. 2010 m. prototipavimo sesijos antrojo bandymo metu vartotojams buvo suteikta galimybė laisvai pasirinkti ir išdėstyti valdiklius

Semantinės informacijos naudojimui paieškoje būdingi trys pagrindiniai pranašumai lyginant su tradicinėmis paieškos sistemomis.

1. Vartotojai gali naršyti informacijoje, sekdami semantiniais ryšiais tarp medijų objektų ir informacijos šaltinių.

2. Asmenų ir reikšminių žodžių homonimiškumas gali būti pašalinamas pagal jų reikšmę.

3. Tampa matomi paieškos rezultatų santykiai.

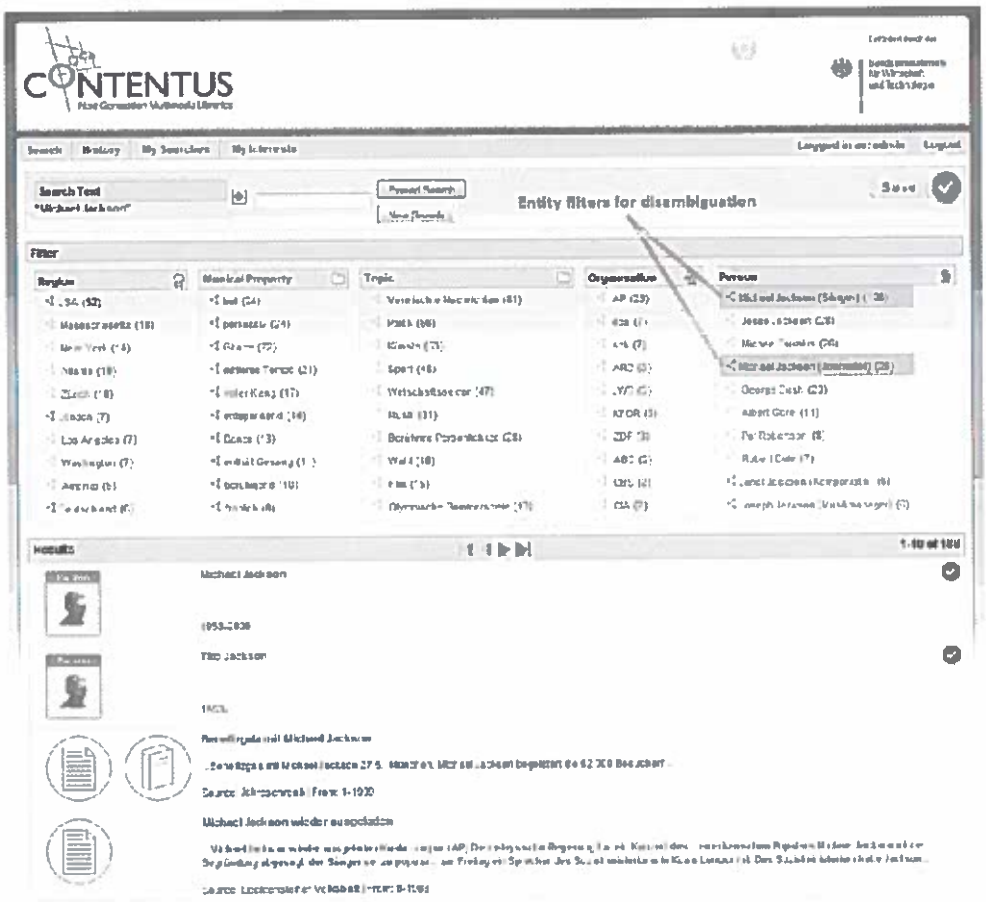
CONTENTUS santykis su UI projektu

Siekiant vartotojams suteikti galimybę iki galo išnaudoti integruotus metaduomenų šaltinius ir skirtingas medijas, būtina pateikti tokią paieškos sąsają, kuri būtų intuityvi ir pasižymėtų naujoviškomis paieškos galimybėmis. CONTENTUS projekto metu buvo sukurtos dvi veikiančios saitynu grindžiamos prototipinės SMMS iteracijos. Apie šiuos du prototipus nuo 2008 m. kaupiami vartotojų atsiliepimai demonstracijų prekybos mugių metu – Frankfurto knygų mugėse 2008 m. ir 2009 m. bei 2009 m. Amsterdamo Tarptautinėje transliuotojų konferencijoje (IBC).

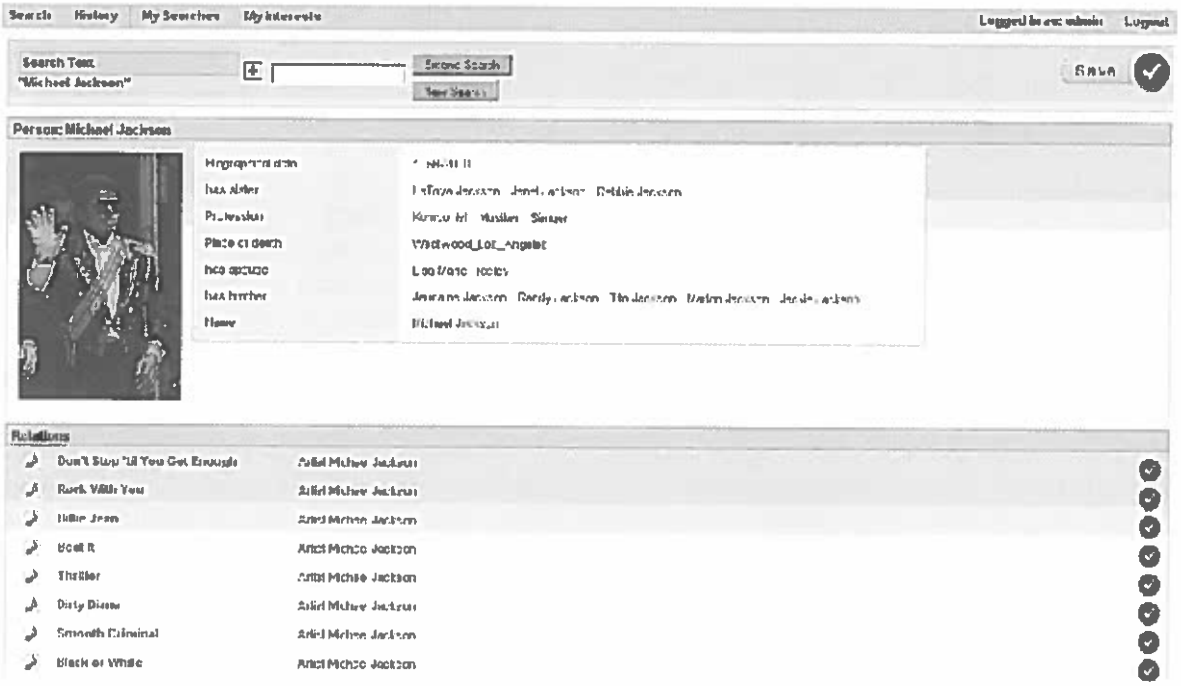
Prieš prasidedant trečiosios demonstracinės versijos projektavimo etapui, siekdami iš naujo patvirtinti ankstesnius (teigiamus) vartotojų atsiliepimus, gautus prekybos mugių metu, 2010 m. surengėme dvi popierinio prototipavimo (žr., pvz., [Maaß, 2008]) sesijas Miuncheno transliavimo technologijos institute, šį kartą įtraukdami vartotojus iš archyvų ir bibliotekų srities.

Mes nusprendėme bandomajai vartotojų grupei neteikti dabartinės saityno pagrindu veikiančios paieškos sistemos. Pirmuoju etapu dalyviams pateikėme iš anksto apibrėžtas paieškos užduotis ir paprašėme jų išsakyti savo idėjas, kaip UI galėtų jas nesunkiai išspręsti. Antruoju etapu bandomajai grupei parodėme kontrolinių elementų rinkinį, norėdami išgirsti bandymo dalyvių atsiliepimus apie tai, kaip jie supranta šiuos elementus ir kokio tikisi interaktyvumo.

Vartotojų testo rezultatai rodo, kad vidutinis vartotojas teikia pirmenybę klasikinei pagal *Google* pavyzdį sukurtai paieškos sąsajai: paieškos laukeliui ir tekstinei paieškos rezultatų pateikčiai. Mūsų manymu, viena tokio pasirinkimo priežasčių yra tai, kad daugelis vartotojų nėra susipažinę su naujoviškesniais ar neįprastais vartotojo sąsajos elementais, todėl nenoriai juos naudoja.



4 schema. Paieškos pagal terminą „Michael Jackson“ CONTENTUS rezultatai iki filtravimo ir asmenų homonimiškumo pašalinimo



5 schema. Dainininko Michael Jackson entitetų puslapis

Kadangi, mūsų nuomone, žvalgymo galimybės yra vienas iš semantinėmis priemonėmis praturtintos paieškos sąsajos privalumų, mums reikėjo pasirinkti sąsają, kuri skatintų vartotojus išnaudoti „pridėtinę semantinę vertę" ir tuo pačiu metu jų nevargintų ir neklaidintų, siūlydama nepažįstamas interaktyvumo galimybes. Dauguma vartotojų teikė pirmenybę fasetinei paieškos sąsajai, leidžiančiai siaurinti pradinius paieškos rezultatus šalinant reikšminių žodžių homonimiškumą, o ne nurodytos paskirties užklausų kalbai ar homonimiškumo nebuvimui prieš įrašant reikšminius žodžius.

Taikymo pavyzdys

Galimas toks naudojimosi dabartine vartotojo sąsaja pavyzdys. Vartotojas ieško knygų, kurias parašė žurnalistas *Michael Jackson*. Paieškos priemonėje jis įrašo terminą „Michael Jackson". Tačiau *Michael Jackson* taip pat yra ir ypač populiaraus dainininko bei muzikanto vardas. Kaip ir tradicinės paieškos sistemos atveju, SMMS iš pradžių pateikia paieškos indekso grynojo teksto atitikmenis, kadangi priemonė negali atspėti, kurį iš dviejų asmenų vartotojas turėjo galvoje.

Dėl vardų panašumo paieškos priemonė pateikia rezultatų derinį, kuriame susipynę norimi ir nenorimi visas medijas apimantys rezultatai. Dauguma jų susiję su menininku Michael Jackson (o ne su žurnalistu), todėl jie vartotojo nedomina.

Be medijų paieškos rezultatų sąrašo, sąsaja pateikia taip pat ir dinaminius filtravimo sąrašus (fasetus), automatiškai generuojamus iš paieškos rezultatų. Pastarieji sąrašai, apimantys relevantiškiausias rezultatų aibės sąvokas ir įvardytus entitetus, sudaromi iš intelektiniu būdu parengtų katalogų metaduomenų ir informacijos, kurią atpažįsta CONTENTUS automatinės turinio analizės moduliai.

Fasetų relevantiškumas grindžiamas ne tik jų pasikartojimo dažnumu rezultatų aibėje, bet ir jų gebėjimu efektyviausiai susiaurinti rezultatų aibės apimtį: kiekvienoje rezultatų aibėje (arba daugumoje iš jų) pasitaikantys fasetai praleidžiami, nes jie nėra reikšmingi filtravimui.

Filtravimo fasetai grupuojami į fiksuotas klasių aibes:

- muzikos sąvokos;
- vietovės;
- temos;
- organizacijos;
- asmenys.

Vartotojas gali pasinaudoti šiais filtravimo fasetais savo paieškai siaurinti – sistemos viduje prie originalaus paieškos termino pridedamas atitinkamas terminas arba entitetas, kurio homonimiškumas pašalintas, ir loginė sąvoka „ir". Kiekvienas filtravimo fasetas turi spalvinę piktogramą, žyminčią duomenų kilmę (žr. 4 schemą) – tai leidžia skirti bibliotekų autoritetiniuose failuose pateikiamus asmenis, kurių homonimiškumas pašalintas, ir daugybinius įvardytus entitetus, randamus tekstinės medžiagos statistinės analizės būdu.

Kadangi mūsų pavyzdyje didžioji dalis paieškos rezultatų susiję su *menininku* Michael Jackson, daugelis temų ir entitetų irgi susiję su muzika. Tačiau matome ir tokias temas, kaip „beer" (alus) ir „whisky" (viskis), kurios būdingos *žurnalisto* Michael Jackson darbams. Asmenų sąrašai apima Michael Jackson iš mūsų asmenų duomenų bazės ir susijusius asmenis, tokius kaip popmuzikos dainininko brolius ir seseris. Vieną kartą spustelėjus pele žurnalistui skirtą fasetą, rezultatų aprėptis susiaurėja – joje lieka tik vartotojui svarbios medijos, neberodomi su dainininku susiję rezultatai.

Įdomu tai, kad termino *Michael Jackson* filtravimo fasetai rodo taip pat ir temas bei organizacijas (pvz., KFOR – Kosovo ginkluotosios pajėgos), nieko bendra neturinčias su akivaizdžiausiai būtinais dviem asmenimis – žurnalistu ir menininku. Nors kai kuriuos vartotojus tai trikdė ir jie praleido šiuos įvedinius kaip visiškai nereikalingus, daugelis iš jų tęsė paiešką, sužinodami apie trečiąjį Michael Jackson, NATO pajėgų generolą – nors ir nelauktą, tačiau vis dėlto naudingą paieškos rezultatą.

Kiekvienas asmuo, atrinktas iš homonimiškų asmenvardžių grupės, turi *entitetų puslapį*, atsiveriantį spustelėjus asmens įrašą rezultatų sąraše. Šiame puslapyje vartotojui pateikiama visa suteikta semantinė informacija: asmenų giminaičiai, jų, kaip kūrėjų, darbai, gimimo datos ir vietovės ir t. t. Entitetų puslapiai praturtinami vaizdais, bibliografine informacija ir Vikipedijos tekstais. 5 schemoje pateiktas *dainininko* Michael Jackson entitetų puslapis.

Entitetų puslapyje vartotojai gali pradėti naują paiešką, spustelėdami bet kurį susietą entitetą, temą, vietovę ir t. t., taip pasinaudodami autentiškomis semantinio naršymo galimybėmis, kurios sudaro integralią visumą su palyginti tradicine sąsajos išvaizda ir atmosfera.

Naujoviški paieškos sąsajos elementai

Mūsų su vartotojais atlikti testai parodė, kad sąveika su *semantinių grafų grafinėmis pateiktimis* nebuvo iki galo suprasta arba laikoma neįgyvendinama. Nors vartotojams buvo aiški asmenų santykių grafinio vaizdo reikšmė, jiems buvo nesuvokiama sąveikos su vizualizuotomis pateiktimis idėja.

Interaktyvus laiko ribų valdymas daugeliui popierinio prototipavimo testų dalyvių pasirodė esąs didžiąja dalimi priimtinas. Rezultatų siaurinimas pažymint laiko kraštinę valdiklyje atrodė kaip intuityvus duomenų paieškos būdas, kuris visuotinai priimtas daugelyje žinių sričių.

Kai kurie vartotojai pasiūlė *hierarchinius filtravimo fasetus* (hiperonimų ir hiponimų hierarchija, pvz., augalas -> gėlė -> rožė); jų tinkamumas bus patikrintas būsimuose prototipuose, nes kai kurios autoritetinių failų dalykinių rubrikų dalys jau pateikiamos hierarchine tvarka.

Interaktyvus žemėlapis. Daugelis testuotų vartotojų teigiamai įvertino grafinį vietovių vizualizavimą paieškos rezultatų aibėje. Vartotojams bus suteikiama galimybė apriboti rezultatus pažymint geografinę sritį žemėlapyje.

Sąsajos testo rezultatai

Mūsų atlikti vartotojų testai parodė, kad:

– semantinės paieškos funkcijos labai padeda greičiau rasti atitikmenis dideliuose multimedijos archyvuose;

– vartotojams labai svarbu suprasti, *kaip* bet koks paieškos bandymas suformuoja rezultatų aibę *ir kodėl*. Antraip semantinis lygmuo gali būti painus, ypač į rezultatų aibę įtraukus tolimesnius ryšius, tokius kaip užklausą atitinkančio asmens giminaičius;

– vartotojai nenoriai naudojasi naujoviškomis vizualizavimo priemonėmis kaip vienintele paieškos pradžios vieta, tačiau priima interaktyvų vizualizavimą kaip paieškos tobulinimo priemonę;

– naršymas dažniausiai atliekamas kaip papildomas žingsnis įvedus vieną ar daugiau reikšminių žodžių. Niekas iš vartotojų, atsakydami į mūsų klausimus, nepasiūlė grynojo naršymo kaip prioritetinio būdo, tačiau, antra vertus, teigiamai atsiliepė apie mūsų prototipų siūlomas naršymo priemones, ypač entitetų puslapius.

Numatomi UI papildymai

Projekto metu mūsų sąsaja bus papildyta bent jau šiomis funkcijomis:

– *entitetų vaidmuo filtravimo fasetuose:* testų dalyviai netiesiogiai pabrėžė, kad būtina suteikti galimybę atskirti filtravimą, pvz., mediją, parašytą asmens A, ir mediją, kurios dalykas yra asmuo A;

– *išsamesnis rezultatų ir fasetų aiškinimas:* rezultatų sąraše turi atsispindėti, kodėl bet kuris elementas pateko į sąrašą, ypač kai rezultatai tik netiesiogiai semantiškai susiję su paieškos terminu;

– *interaktyvus laiko ribų vizualizavimo valdymas:* vartotojams turi būti suteikiama galimybė siaurinti rezultatus vizualizavimo priemonėje pažymint laikotarpį taip, kad būtų rodomi tik šiam laikotarpiui priskiriami rezultatai;

– *interaktyvus žemėlapis:* vartotojams turi būti suteikiama galimybė apriboti gautus rezultatus laisvai žemėlapyje pasirenkama sritimi.

Išvada

Ne tik Vokietijoje, bet ir Europos Sąjungos lygmeniu dedamos didžiulės pastangos, kad būtų užtikrinta prieiga prie skaitmenintos kultūros paveldo medžiagos, kad ir kam ji būtų skirta – ilgalaikiam saugojimui ar skaitmeninių bibliotekų, tokių kaip *Europeana* ar *Deutsche Digitale Bibliothek* kūrimui. Todėl vis daugiau bibliotekų ir archyvų susiduria su skirtingiems skaitmeninimo pro-

jektams priklausančių paveldo objektų, vietinių metaduomenų ir išorinių duomenų rinkinių integravimo iššūkiu. Deja, vis dar nepakanka priemonių, sudarančių galimybę nesudėtingai, tačiau visa apimantį, objektų ir metaduomenų teikimą bibliotekų sistemoms ir katalogams.

Kita vertus, yra didelis bibliotekų ir archyvų vartotojų poreikis gauti prieigą prie skaitmeninės medijos, teikiamą vienodomis sąlygomis visiems medijų tipams. Kaip įprasta šiuolaikiniams medijų vartotojams, tikimasi, kad garso ir vaizdo turinys bus tiesiogiai integruotas į informacijos objektus, todėl bus taip pat indeksuojamas ir bibliotekų bei archyvų paieškos sistemų. Šis poreikis dažnai sukelia būtinybę integruoti ir trečiųjų šalių įrankius ir duomenų šaltinius.

CONTENTUS kuriamos technologijos ir koncepcijos, padėsiančios spręsti šiuos uždavinius ir ženkliai supaprastinti skaitmeninės medijos rinkinių kūrimą ir pateikimą bei naudojimąsi jais. Dvi mūsų įdiegtos saitynu grindžiamos demonstracinės sistemos parodė, kad toks medijų objektų ir metaduomenų agregavimas bei pateikimas įmanomas ir, tai dar svarbiau, naudingas bibliotekų ir archyvų vartotojams. Neabejotinai didės žinių sklaidos ir atskleidimo semantinės paieškos būdu svarba, ir mes tikimės, kad CONTENTUS reikšmingai prisidės prie ateities skaitmeninių bibliotekų kūrimo.

Pamokos

Galima būtų pateikti keletą pagrindinių principų, kurių naudingumas pasitvirtino siekiant projekto tikslų.

– *Modulinio projekto svarba.* Ne visų bibliotekų ir archyvų poreikiai yra vienodi – kai kurioms bibliotekoms ir archyvams nebūtini skaitmeninimo metodai, kai kada bibliotekos ir archyvai gali pasiūlyti paieškos sąsajas, kurioms būtų reikalingos tik objektų metaduomenų kūrimo ir siejimo technologijos. Todėl CONTENTUS problemoms spręsti specialiai parengtas modulinis projektas. Kiekvienam apdorojimo procesui (žr. „Įvadą“) parengti skirtingi sprendimai, kuriuos suinteresuotos institucijos gali taikyti atskirai arba visus kartu.

– *Atvirųjų standartų ir sąsajų svarba.* Siekdami palengvinti anksčiau aptartą CONTENTUS technologijų integraciją, daug dėmesio skyrėme atviriesiems standartams, sąsajoms ir duomenų formatams. Pavyzdžiui, semantinei multimedijos paieškai CONTENTUS naudojame į paslaugas orientuotą struktūrą (SOA). Sąveikaujant skirtingiems moduliams per *Web Services*, atsižvelgiama į šiuolaikinės bibliotekos infrastruktūros reikmes ir labai lanksčiai integruojami skirtingi duomenų šaltiniai – tiek esantys pačioje bibliotekoje, tiek teikiami trečiųjų šalių paslaugų teikėjų. Integruojant išorinius informacijos šaltinius, jų metaduomenų naudojimą palengvina tipiniai susietų duomenų rinkinių formatai (XML/RDF).

– *URI nauda semantiškai siejant objektus, sąvokas ir informacijos šaltinius.* Žr. ankstesnį skirsnį „URI naudojimas“.

– *Vartotojai pirmenybę teikia gerai struktūriškai suprojektuotoms, tačiau veiksmingoms sąsajoms.* Tai ypač aktualu naujoviškų funkcijų, pvz., būtinų semantinei paieškai atlikti, atveju. Būtina, kad grafinės pateiktys (pvz., sąvokų ir santykių) išliktų intuityvios ir paprastos naudoti, netgi jas taikant skirtingose žinių srityse. Yra didelis poreikis, ypač tarp profesionalių vartotojų, teikti plačias galimybes individualizuotai vartotojo sąsajos sąrankai.

Būsima veikla ir vizija

Dabar projektas pereina kasmetinį pasikartojantį realizavimo ciklą ir sparčiai artėja prie trečiojo etapo – saitynu grindžiamos demonstracinės sistemos, kuri bus pateikta profesionaliai auditorijai IBC konferencijos parodoje 2010 m. rugsėjo mėn. Naujoji demonstracinė versija apima pertvarkytą vartotojo sąsają, taip pat praplėstą semantinę fasetinės paieškos sistemą, pasižymi tobulesniu multimedijos turinio apdorojimu. Iki 2010 m. pabaigos naujoji CONTENTUS SMMS demonstracinė versija bus pristatyta ir stacionariame THESEUS mokslinių tyrimų demonstracijų centre Berlyne ir kai kuriuose bibliotekų bei archyvų bendruomenės renginiuose.

Tolesnės vystymo kryptys apims semantinių galimybių plėtrą integruojant *semantinę medijų žiūryklę*, kuri užtikrins tobulesnę sąsają su sistemos atpažintais įvardytais entitetais. Kitas svarbus spręstinas uždavinys – didesnis pritaikymas asmeniniams ir bendruomenės poreikiams, padėsiantis susiformuoti naujiems *informacijos naudojimo bendradarbiaujant* būdams. Tai leis vartotojams išsamiai sąveikauti su informacijos objektais – tiek tenkinant asmeninius poreikius, tiek bendradarbiaujant su kolegomis ar vartotojų grupėmis. Galiausiai svarbu ir tai, kad mes integruosime vertingus duomenų šaltinius iš susietų atvirų duomenų debesų.

Viena iš CONTENTUS vizijų yra teikti savo metaduomenų integravimo ir semantinės paieškos koncepcijas faktinio istorinių objektų rinkinio kontekste. Tuo tikslu buvo suskaitmeninta didelė dalis buvusios Vokietijos Demokratinės Respublikos *Musikinformationszentrum* (MIZ) archyvo. Įvairūs šio izoliuoto rinkinio medijų objektai bus integruoti į galutinę CONTENTUS demonstracinę versiją, kuri bus baigta rengti 2012 m. pradžioje. Manome, kad šis turinys labai tinkamas mūsų sistemos privalumams specifinėje žinių srityje atskleisti ir kad jis suteiks naujų įžvalgų apie buvusios Vokietijos Demokratinės Respublikos muzikinį gyvenimą.

Iš anglų kalbos vertė T. Auškalnis

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

Nuorodos

Bossert, Klaus and Nicholas Flores-Herr and Jan Hannemann. *CONTENTUS: Technologien für digitale Bibliotheken der nächsten Generation*. Dialog mit Bibliotheken, Bd. 21, p. 14-20. ISSN 0936-1138. German National Library, 2009.

Hannemann, Jan and Jürgen Kett. *Linked Data for Libraries*. In: Proceedings of World Library and Information Congress: 76th IFLA General Conference and Assembly (IFLA 2010), Gothenburg, Sweden.

Heß, Andreas, 2006. An *Iterative Algorithm for Ontology Mapping Capable of Using Training Data*. In: Proceedings of the 3rd European Semantic Web Conference (ESWC 2006), Budva, Montenegro.

Johnston, Eddie and Nicholas Kushmerick, 2008. *Web Service aggregation with string distance ensembles and active probe selection*. Information Fusion 9(4): 481-500 (2008).

Levenshtein, Vladimir I., 1965. *Binary codes capable of correcting deletions, insertions, and reversals*. In: Doklady Akademii Nauk SSSR. 163, Nr. 4, 1965, S. 845-848 (In Russian. English translation in: Soviet Physics Doklady, 10(8) S. 707-710, 1966).

Maaß, Christian and Elica Savova, 2008. *Paper Prototyping in der Softwareentwicklung*. In: Das Wirtschaftsstudium, 11/2008 (In German).

Melnik, Sergey and Hector Garcia-Molina and Erhard Rahm, 2002. *Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching*, In: Proceedings of the 18th International Conference on Data Engineering (ICDE), San Jose CA, USA.

Pilz, Anja and Gerhard Paaß, 2009. *Named Entity Resolution Using Automatically Extracted Semantic Information*. In: Proceedings of workshop Lernen, Wissen, Adaptivität (LWA 2009), Darmstadt, Germany.

Russell, Robert C., 1918. United States Patent 1261167, application filed Oct. 25, 1917, patented Apr. 2, 1918.

Shvaiko, Pavel and Jérôme Euzenat and Fausto Giunchiglia and Heiner Stuckenschmidt and Natasha Noy and Arnon Rosenthal (Editors), 2009. *Ontology Matching (OM-2009), Papers from the ISWC Workshop*. October 2009.

Winkler, W. E., 1999. *The state of record linkage and current research problems*. Statistics of Income Division, Internal Revenue Service Publication R99/04, 1999.