PREDMETNI PRISTUP INFORMACIJAMA NA INTERNETU I KNJIŽNIČNA KLASIFIKACIJA

SUBJECT APPROACH TO INFORMATION ON THE INTERNET AND LIBRARY CLASSIFICATION

Aida Slavić
Filozofski fakultet, Odsjek za informacijske znanosti, Zagreb
E-mail: aida.slavic@ffzg.hr

UDK/UDC 025.4.036
004.738.5
Pregledni rad/Review
Primljeno/Received: 5.2.2001,

Sažetak

Članak se bavi evolucijom informacijskog prostora na internetu, njegovim karakteristikama i različitim pristupima pronalaženju informacija. Predmetni pristup informacijama postaje sve važniji s rastom globalne mreže, otvorenog pristupa informacijama dostupnim u različitim digitalnim formatima koji mogu sadržavati tekst, zvuk, sliku ili skupove podataka. Među pomagalima za pronaženje informacija na internetu razlikuju se opći informacijski servisi i specijalizirani informacijski servisi. Razvoj specijaliziranih informacijskih servisa stavlja naglasak na važnost standarda za metapodatke elektroničke grade i ulogu jezika za indeksiranje u opisu izvora informacija. Knjižnična klasifikacija medu prvim je tradicionalnim knjižničnim pomagalima koja su uspješno primijenjena u unapređenju pronalaženja informacija na internetu.

Ona se može koristiti u metapodacima ugrađenim u dokument kao i u samostalnim metapodacima. Isto se tako može koristiti za pretraživanje informacija kao i za pregledavanje i navigaciju kompleksnim prostorom prezentacije znanja.

Ključne riječi: internet, pronalaženje informacija, predmetni pristup informacijama, knjižnična klasifikacija, bibliografska klasifikacija, specijalizirani informacijski servisi, predmetni direktoriji, metapodaci

Summary

The paper deals with the evolution of Internet information space, its characteristics and different approaches to resource discovery. Subject approach to information becomes increasingly important with the growth of the global network, open access to information resources available in file formats that may contain text, sound, visual or data sets. Tools for finding information on the Internet are divided into two main groups, namely general

information services and quality information services. Development of quality information services stresses the importance of metadata standards and the use of indexing languages in resource description. Library classification is one of the first traditional library tools that has been successfully applied to improve resource discovery on the Internet.

It can be applied both in the embedded and stand-alone metadata systems. It can also be applied for both resource discovery and navigation and exploration through the complex space of knowledge presentation.

Keywords: Internet, resource discovery, subject approach to information, library classification, bibliographic classification, quality information services, subject gateways, metadata

Obilježja informacijskog okruženja na internetu

Od trenutka kada je, već početkom devedesetih, internet postao javni medij za publiciranje i komuniciranje informacija i izvan znanstveno-istraživačke zajednice, opći programi za pretraživanje *{general search services - search engines),* postaju posve nedostatni u lociranju relevantnih izvora informacija iz specijaliziranih područja. Istraživanje i vrednovanje pretraživača koje je potom uslijedilo, potvrdilo je sve slabosti u pronalaženju informacija pretraživanjem slobodnog teksta, a razvoj programa za pretraživanje tijekom devedesetih samo je prividno poboljšao postojeću informacijsku zagušenost (Bharat; Broder, 1997., Lawrence; Lee Giles, 1998.). Informacije koje kolaju webom, nalaze se na statičkim stranicama, tj. stranicama koje obično vidimo dok krstarimo internetom, i na puno većem broju dinamičkih stranica, tj. stranica koje nastaju pretraživanjem baza podataka ili se generiraju programima poput Jave ili Perla (Sherman, 1999.). Kada se govori o pretraživanju stranica na internetu, najčešće se misli samo na pretraživačima dohvatljiv, "vidljiv", tj. statički dio weba. Iako je to manji dio onoga što se u pogledu informacija nalazi na internetu, prema podacima iz veljače 1999., statički se web sastojao od preko 800 milijuna stranica s indeksom rasta iznad 100% u godini (Green, 1999.).

Velik priljev web-stranica i različiti multimedijski formati, neizvjesna trajnost te nedostatna strukturiranost elektroničke građe na webu, uvelike otežavaju automatsko lociranje relevantnih informacija. No unatoč činjenici daje s razvijenim generacijama[1] programa za pretraživanje, pronalaženje relevantnih informacija

'Prema D. Greenu, iako nije posve ispravno inzistirati na stupnjevima razvoja jer su preklapanja velika i predstavnici jedne vrste programa često se nadograđuju i kombiniraju, moguće je ugrubo razlikovati nekoliko generacija pretraživača. Prva generacija pretraživača obuhvaća jednostavne programe za prikupljanje s direktorijima poput AltaViste, WebCrawlera itd., koji su potom služili kao podloga metapretraživačima koji se zasnivaju na pretraživanju više jednostavnih servisa za pronalaženje informacija. Druga je generacija nastupila s pojavom DirectHit pretraživača 1998. godine, sa sposobnošću redanja rezultata prema popularnosti. Iste godine u lipnju Ask Jeeves, a osobito tada javno dostupni The Electric Monk, obilježili su napredak prema pretraživačima treće generacije koji koriste algoritme za obradu prirodnog jezika i obavljaju pretraživanje upita postavljenog prirodnom rečenicom (engleskog jezika). Četvrta generacija uslijedila je potom s pretraživačima Google i Clever s vrlo uspješnom metodom redanja rezultata po relevantnosti na osnovi analize broja veza *(link-based analysis)* koje na neku stranicu vode s drugih stranica. Na prvome mjestu u nizanju rezultata upita naći će se ona stranica na koju upućuje najviše drugih

postalo bitno lakše i sigurnije, opći servisi za pretraživanje daleko su od profesionalne razine koju u pojedinim područjima, disciplinama ili istraživačkim domenama traže korisnici kojima je internet dio svakodnevnog radnog okruženja.

Prirodna je posljedica tih poteškoća u pronalaženju kvalitetnih informacija paralelni razvoj drugih servisa za otkrivanje i pronalaženje izvora informacija. Spomenuti je trend vezan uz fazu u razvoju interneta koju L. Dampsey naziva "akademskom",[2] a imao je za posljedicu velik broj besplatnih predmetno orijentiranih direktorija *(subject gateways)* od kojih su mnogi prerasli u servise za selekciju, klasifikaciju i prezentaciju informacija namijenjenih pojedinim zajednicama korisnika na nacionalnoj ili međunardnoj razini. Specijalizirani servisi informacija, može se reći, nastali su sredinom devedesetih kao rezultat izazova koje je pronalaženju informacija nametnuo rast interneta (Dampsey, 2000.). Prvi predmetni direktoriji u okviru britanskog projekta eLIB[3] bili su inovacija i poslužili su kao model koji je utjecao na ostale slične projekte u svijetu. Valja, stoga, razlikovati dva osnovna načina i dvije razine u pronalaženju informacija na internetu: *opće servise za pronalaženje informacija {general information services* ili, popularnije, *search engines)* i *specijalizirane servise za pronalaženje informacija {quality information services).** Jasna distinkcija i analiza postupaka otkrivanja, pronalaženja, organizacije i prezentacije izvora informacija na internetu putem kvalitetnih izvora informacija prvi je put opširnije obrađena u izvještaju prve faze europskog projekta DESIRE (Koch, 1997.), a potom u istraživanjima Johna Kirriemuira, Traugotta Kocha i, napose, Lorcana Dampseya (Kirriemuir, 1999., 2000., Koch, 2000., Dampsey, 2000.). Specijalizirani informacijski servisi i dalje su predmetom velikog zanimanja međunarodnih projekata poput DESIRE II[5] i RENARDUS[6] i imaju važnu ulogu u primjeni i razvoju standarda za opis i indeksiranje elektroničke građe te primjeni tradicionalnih knjižničarskih pomagala u upravljanju i organizaciji informacija. Uz standarde za opis ove građe, koji su poznati pod nazivom metapodaci, postoji opravdano zanimanje za jezike za

stranica. Green potom spominje razvoj inteligentnih pretraživača *{intelligent agents* ili *bots)* koje možemo smatrati sljedećom fazom u razvoju, a koji omogućavaju personalizaciju i profiliranje programa za pretraživanja prema korisnikovim potrebama *(user profile)* (Green, 1999.).

[2]Prema L. Dampseyju, internet je počeo kao "ezoterični fenomen", uglavnom vezan uza znanstveno-istraživačko područje fizike, potom je prerastao u "internetsku zajednicu", s porastom broja korisnika, komunikacije kroz raspravišta i druge kanale. Nakon što su pojedine zemlje pokazale zanimanje i počele ulagati u infrastrukturu na nacionalnoj razini te financirati razne projekte na akademskoj razini, internet je prerastao u "akademski" medij. Tu fazu u razvoju obilježava izgradnja mnogobrojnih besplatnih servisa i usluga, predmetnih direktorija i pretraživača.
Trenutačna faza interneta pripada "javnoj informacijskoj infrastrukturi" koju karakterizira pojačana sigurnost i kontrola protoka informacija koje omogućavaju pojavu komercijalnih i financijskih servisa, trgovine itd.

[3]The Electronic Library Programme (http://www.ukoln.ac.uk/services/elib).

"Engleski izraz "quality information services" nije moguće izravno prevesti na hrvatski a da se potpuno očuva smisao. Pridjev "specijalizirani" u ovom kontekstu ukazuje na razinu stručnosti samog servisa i kvalitete informacija koju pruža.

[5]DESIRE - Development of European Service for Information on Research and Education (http://www.lub.lu.se/desire/desireIIIndex.html).

"RENARDUS (http://www.renardus.org/).

označivanje sadržaja, razne vrste kontroliranih rječnika, tezauruse, predmetne sustave, a osobito knjižničnu klasifikaciju.

Opći servisi za pronalaženje informacija

Priroda *općih servisa* za pretraživanje potpuno određuje njihovu primjenu. Ti su programi brojni i tehnološki različito razvijeni i danas se ne može dati općenita ocjena koja bi vrijedila za sve spomenute servise jer se oni međusobno prilično razlikuju kvalitetom i po tome koliki dio interneta pokrivaju. Njihov informacijski profil, međutim, može se sažeti u nekoliko važnih karakteristika: oni su automatizirani i neselektivni u odnosu na web-stranice koje indeksiraju, nisu namijenjeni određenom tipu korisnika i nastoje pokriti što veći broj stranica, iz što više interesnih područja.[7]

Općenito govoreći, servisi kao što su AltaVista, Yahoo, HotBot, MetaCrawler, Direct Hit, Google itd. izvrsno obavljaju funkciju kojoj su prvobitno namijenjeni jer omogućuju točno uspostavljanje veze između tražene riječi i istovjetnog niza znakova na web-stranicama. Iako se korisniku ne mora nužno svidjeti činjenica da se upravo riječ koju traži nalazi na sto tisuća stranica, to se svakako ne može uzeti kao mana tog programa.[8]

Opći pretraživači mogu izvrsno poslužiti onima koji znaju kako program funkcionira, kako se upit sužuje pomoću Booleovih logičkih operatora ili rafinira odaziv kojom od ponuđenih metoda. Posebno su korisni kada treba provjeriti informaciju vezanu uz vlastita imena osoba, ustanova, predmeta ili proizvoda u slučajevima kad će poklapanje traženog iskaza s iskazom pronađenim na web-stranicama uglavnom dati relevantne rezultate. Za one koji se znaju njima služiti, opći su pretraživači nenadomjestiva podrška u obavljanju potrebnih provjera, brzom i jednostavnom pronalaženju informacija čija su pozadina i izvor donekle poznati.

No iako napredni programi poput Googlea uistinu funkcioniraju izvrsno, teško je zamisliti da će se bilo tko tko na internetu traži informacije vezane uz važna pitanja iz područja znanosti, prava, industrije, poslovanja ili obrazovanja osloniti u svom radu *isključivo* na opći servis za pretraživanje. Slično kao što u svakodnevnom životu nećemo važnu informaciju vezanu uz posao, istraživanje, pravna ili obiteljska pitanja tražiti u novinama, na televiziji ili radiju, već ćemo se obratiti kojemu organiziranom i specijaliziranom izvoru informacija od kojega ćemo dobiti brzu i pouzdanu informaciju. Način na koji se internet raslojava u smislu komunikacije

[7]Svi opći servisi temelje se na programima za automatsko prikupljanje i indeksiranje web-stranca, *{web spiders* i/ili *web harvesters),* bazi podataka za pohranu web-kazala te više ili manje naprednim programima za formuliranje upita i redanje rezultata po relevantnosti *{database interrogating software* i *ranking algorithms).*

[8]Internetski pretraživači nisu isto što i sustavi za pretraživanje informacija *{information retrieval systems)* čija je funkcija dvojaka: pronaći sve što je relevantno i filtrirati tj. zaustaviti nerelevantne dokumente. Da bi se to postiglo, sustavi za pretraživanje informacija grade se na dobro strukturiranim surogatima dokumenata (bibliografskim podacima ili metapodacima), u zatvorenim i kontroliranim uvjetima baza podataka čija je veličina ograničena a sadržaj namijenjen određenoj vrsti korisnika.

informacija na različitim razinama i za različite potrebe, upućuje na valjanost takve usporedbe.

Sredinom devedesetih razvoj Dublin Core standarda za opis izvora informacija na internetu,[9] istodobno s novom verzijom HTML-a koja je omogućila da se u zaglavlju izvornog koda elektroničkog dokumenta unese jednostavan opis dokumenta s opisom njegova sadržaja, u prvom se trenutku činio ostvarivim rješenjem. Opis dokumenta bio bi tako istodobno dostupan s dokumentom jer bi autor osobno i/ili uz pomoć nekih od automatskih programa za generiranje Dublin Core metapodataka potpomogao njegovo pravilno lociranje pomoću općih servisa za pretraživanje. Ubrzo se, međutim, pokazalo da oni koji publiciraju dokumente nisu uvijek skloni opremati svoje web-stranice metapodacima. Najmotiviranijima su se pokazali oni koji su zloupotrebljavali metapodatke dajući lažan opis dokumenta kako bi povećali mogućnost indeksiranja nekih nepoželjnih ili uvredljivih sadržaja. Posljedica je toga da veoma mali broj općih servisa za pretraživanje indeksira stranice koristeći se metapodacima unutar samog dokumenta *{embedded metadata)*. Dublin Core standard našao je mnogo uspješniju primjenu u specijaliziranim informacijskim servisima o kojima će ovdje biti posebno riječ.

Specijalizirani informacijski servisi

Iako im je namjena pronalaženje informacija na internetu, "predmetni direktoriji" *{subject gateways),* koji u mnogim slučajevima prerastaju u prave specijalizirane informacijske servise *{quality information services),* po svojoj se informacijskoj prirodi potpuno razlikuju od spomenutih općih servisa. Posve su selektivni s obzirom na kvalitetu, stabilnost, pouzdanost izvora koje uključuju u svoje direktorije. Obično su namijenjeni određenoj profesionalnoj, akademskoj ili interesnoj zajednici korisnika. Iako često usmjereni prema jednome predmetnom području poput MathGuidea, MetaChema, OMNI-ja (Organizing Medical Network Information), GeoGuidea ili EEVL-a (Edinburgh Engineering Virtual Library), ti servisi mogu biti po svom sadržaju općeniti poput CORC-a (Cooperative Resource Cata-

'Dublin Core Metadata Element Set (DCMES) osnovni je standard za opis elektroničkih izvora informacija u svrhu pronalaženja i razmjene informacija. To je međunarodni sporazum o osnovnom skupu elemenata koji je nuždan da se jednoznačno opiše i na osnovi toga na internetu pronađe bilo koji objekt nalik dokumentu *{document-like object)*.

DCMES je prihvatila široka zajednica internetskih korisnika kao jednostavan model opisa sadržaja dokumenta. Za razliku od drugih standarda namijenjenih potrebama određene profesionalne zajednice (poput Encoding Archive Initiative za arhive, Computer Interchange of Museum Information (CIMI) za muzeje ili (UNI)MARC za knjižnice), DCMES je skup općenito razumljivih opisnih elemenata koji služe razmjeni među različitim zajednicama korisnika. Da bi ispunio ovu svrhu, DCMES je ostvaren na sljedećim načelima: jednostavnost, semantička razumljivost, međunarodni konsenzus, proširivost i modularnost (u smislu okvira za podršku različitim potrebama). Zadnja verzija (1.1) standarda Dublin Core Metadata Element Set (DCMES) dostupna je na http://purl.oclc.org/dc/documents/rec-dces-19990702.htm.

Odobreni atributi za opis elemenata dostupni su od 11. srpnja 2000. pod nazivom Dublin Core Qualifiers na adresi http://purl.org/documents/rec/dcmes-qualifiers-200007.

logue) ili Signposta. Bilo da se odnose na jedno područje ili na više područja, mogu indeksirati web-stranice na cijelom internetu ili mogu biti ograničeni na nacionalnu razinu poput GERHARD-a (German Harvest Automated Retrieval and Directory). Svaki od spomenutih servisa obavlja selekciju izvora prema unaprijed utvrđenim pravilima.

Broj izvora koji ovi direktoriji nude, može se kretati od nekoliko stotina do nekoliko stotina tisuća web-stranica. Da bi se omogućila organizacija i prezentacija te pregled prikupljenih informacija, predmetni direktoriji, razumljivo je, koriste neku od općeprihvaćenih specijalnih ili općih bibliografskih klasifikacija poput Deweyjeve decimalne klasifikacije, Klasifikacije Kongresne knjižnice ili Univerzalne decimalne klasifikacije.

Nakon automatskog prikupljanja adresa web-stranica pomoću računalnih programa *(harvesters),* ti servisi koriste ljude za daljnju manualnu obradu, intelektualnu analizu i provjeru kvalitete informacija prema unaprijed dogovorenim pravilima. Ako postoje financijske pretpostavke, institucije koje stoje iza ovih servisa potiču istraživanje i napore u iznalaženju pouzdanih poluatomatiziranih i automatiziranih načina da se informacije obrađuju i provjeravaju.

S obzirom na visoke zahtjeve koje ti servisi imaju u pogledu upravljanja informacijama, većina gradi baze podataka s punim opisom internetske stranice prema kojemu od postojećih standarda za izradu metapodataka elektroničke građe, poput spomenutog standarda Dublin Core ili kojega drugoga vezanog za određeno interesno informacijsko područje. Treba spomenuti da je prvi katalog internetskih izvora nastao već u razdoblju 1991.-1993. u okviru OCLC-ova (Online Computer Library Centre) projekta *Internet Resource Project.* Potom su unutar iste istraživačke zajednice uslijedili 1993.-1996. InterCat, a 1999. započet je rad na projektu *Cooperative Resource Catalogue* (CORC).

Koncept metapodatka kao sastavnog dijela elektroničkog dokumenta koji nastaje u procesu kreiranja dokumenta i ostaje integralni i nedodirljivi dio samog elektroničkoga dokumenta, u slučaju informacijskih servisa nužno ustupa mjesto konceptu izrade opisa web-stranice kojim se može neovisno manipulirati i koji se može podvrgnuti rigoroznijim standardima i obogatiti dodatnim informacijama *(stand-alone metadata).* Isto vrijedi kada se radi o pretraživanju izvora u obliku slike, zvuka, animacije, programskog koda ili videa. Iako je tehnički izvedivo publicirati ovu vrstu elektroničke građe s već ugrađenim opisom, neujednačenost standarda za izradu metapodataka (na razini strukture, sintakse i semantike) govori u prilog naknadnim, o izvoru neovisnim, metapodacima.

Održavanje, selekcija, klasifikacija i prezentacija izvora na internetu vezana je uz čitav niz tehnoloških pomagala. Projekt ROADS (Resource Organisation and Discovery in Subject-based Services), primjerice, imao je kao rezultat izradu niza programskih pomagala za organiziranje i održavanje takvih servisa, namijenjenih sudionicima britanskog projekta eLIB. Danas se neki od najvažnijih britanskih specijaliziranih informacijskih servisa poput SOSIG-a (Social Science Information Gateway) koriste ovim programskim rješenjima.

Postojeći specijalizirani informacijski servisi nastoje se danas povezati u federacijsku strukturu koja će omogućiti interdisciplinarno pretraživanje i istodobno

pretraživanje izvora infomacija različitih profesija, domena i različitih zajednica korisnika.[10] Trenutni trend naginje prema komercijalnim uslugama koje bi bile zasnovane na dobavljanju visokokvalitetnih usluga iz različitih izvora, baza podataka i intraneta te drugih specijaliziranih informacijskih servisa. L. Dampsey ističe da će budućnost tih servisa ovisiti o njihovoj sposobnosti da se usmjere na održiv model distribuiranih servisa koji će možda izaći iz sadašnjih okvira institucionalnih identiteta i postati dio, primjerice, nacionalnih obrazovnih servisa ili profesionalnih portala (Dampsey, 2000.).

Napredak u razvoju programa za podršku specijaliziranih informacijskih servisa neposredno je vezan uz razvoj standarda i tehnologije koja će podržavati predmetni pristup informacijama, a time i korištenje i prilagodbu postojećih indeksnih jezika poput klasifikacije.

Važnost predmetnog pristupa informacijama

Najveća je vrijednost interneta jednostavno i lako publiciranje različitih sadržaja koji su dostupni i vidljivi na web-stranicama. Veličina zbirke sadržaja, dostupne pregledavanju i korištenju, ograničena je jedino globalnom mrežom. Unutar tih okvira informacijski prostor dalje raste unutar velikog broja intranet i ekstranet" mreža. S obzirom na različite formate i medije na kojima su razni izvori informacija pohranjeni te katkad nejasno porijeklo, nepostojanu lokaciju ili uvjete korištenja, može se pretpostaviti da će se informacije tražiti i pretraživati ponešto drukčije s dobro strukturiranim, standardiziranim, sadržajno eksplicitnim i kontroliranim okruženjem bibliografskih baza podataka.

Novija istraživanja pokazuju da korisnici pretražuju izravno javno dostupne knjižnične kataloge pretežno tražeći dokumente prema njihovu sadržaju (Long, 2000.). Pogotovo ne treba očekivati da će korisnici interneta izvore informacija tražiti prema autoru, naslovu ili kojoj drugoj formalnoj karakteristici. To jednako vrijedi za tekstualne dokumente i za slike, zvuk, videozapise ili druge vrste informacija.

Posve prirodno, korisnik interneta očekuje da se jedan te isti informacijski objekt pojavljuje nekoliko puta u posve različitom kontekstu: nekad kao samostalni objekt, nekad kao dio veće cjeline. Podaci poput naslova, autora, datuma kad je stavljen na raspolaganje, formata i tome slično mogu se razlikovati u različitim slučajevima. Jedina "opipljiva" karakteristika elektroničkog dokumenta, u ovom slučaju ostaje njegov sadržaj (Chan, 2000.). Korisnik upravo očekuje da će moći birati među različitim instancijama određenog sadržaja, tražeći isti ili sličan sadržaj u različitim elektroničkim formatima (tekst, HTML, pdf itd.) ili datotekama različite veličine.

[10]RDN (Resource Discovery Network) jedna je od incijativa u Velikoj Britaniji s ciljem organiziranja postojećih specijaliziranih informacijskih servisa u centraliziranu mrežu s nizom podorganizacija *(hubs)*.

"Extranet je mreža, za javnost "zatvorenih", intranet mreža koje komuniciraju preko interneta.

Da bi se omogućio prikaz sadržaja dokumenta, tj. prezentacija i pregled elektroničke građe prema sadržaju, važno je raspolagati standardima i pomagalima za izradu formalnog i sadržajnog opisa, a potom treba omogućiti strojnu čitljivost i strojnu razumljivost odgovarajućih metapodataka.

Knjižničarskoj zajednici koja čitavo stoljeće razvija i primjenjuje standarde kako bi omogućila gradnju i razmjenu bibliografskih informacija, to nipošto nije novo. Ali je posve novo i važno razumjeti da je knjižničarska zajednica na internetu tek jedna od mnogobrojnih informacijskih zajednica. Ono što je vrijedilo unutar zatvorenih zidova bibliografske kontrole od standarda do stručnog nazivlja, vrlo je teško primjenjivo u području elektroničke trgovine, posla i financija te pravnih, političkih, obrazovnih i drugih informacijskih servisa. Svaka od spomenutih zajednica razvila je određene standarde za opis elektroničke građe koju komunicira i razmjenjuje. I poput bibliotekarske zajednice, svatko za sebe izvrsno funkcionira. Može se jedino reći da nitko nije imao vremena razviti jezike za označivanje sadržaja i organizaciju i prezentaciju znanja do one razine do koje su to učinili knjižničari. Trend razvoja prema otvorenom pristupu informacijama na internetu, nameće potrebu za pregledavanjem i pretraživanjem svih informacija neovisno o tome iz koje domene dolaze *(cross-domain searching)*. To je područje u razvoju standarda na internetu kojemu bi knjižničarsko iskustvo u gradnji i korištenju indeksnih jezika moglo bitno pridonijeti.

Knjižnična klasifikacija, primjerice, vrlo je rano prepoznata kao korisno pomagalo. Standardi za izradu metapodataka nekih specijalnih domena poput EAD-a (Encoded Archive Description) za arhivske podatke, IMS-a (Learning Object Metadata) ili CIMI-ja (Consortium for the Computer Interchange of Museum Information) naveli su u svojim elementima za opis sadržaja knjižničnu klasifikaciju. Dublin Core kao opći standard za izradu opisa izvora informacija na internetu, također promiče korištenje nekih od postojećih pomagala knjižnične struke. Kao rezultat takva pristupa, korištenje klasifikacije u pretraživanju omogućit će pronalaženje i pregledavanje izvora informacija iz mnogih različith područja ljudske djelatnosti.

Korištenje knjižnične klasifikacije

Postoji više razloga koji idu u prilog primjeni knjižnične klasifikacije na internetu s obzirom na važnost i vrijednost predmetnog pristupa informacijama. Kad govorimo o pretraživanju, imamo na umu dvije osnovne funkcije koje svaki kvalitetan sustav za pretraživanje nudi:

- odaziv na precizan upit o određenom predmetu (što? ~ gdje se nalazi?)
- mogućnost pregledavanja i kretanja kroz predmetno područje (gdje? ~ što se tamo nalazi?)

Prva se funkcija može zadovoljiti korištenjem bilo kojeg abecednoga predmetnog sustava za označivanje poput pojmova označitelja preuzetih iz nekog tezaurusa ili predmetnica nekog predmetnog sustava. Druga se funkcija kretanja kroz velik broj različitih, sustavno organiziranih predmetnih područja od općenitog ka specijalnom i obrnuto, može postići jedino korištenjem klasifikacijske sheme

(Svenonius, 2000.). Štoviše, klasifikacijska shema može služiti i u pretraživanju preciznog zahtjeva bilo pretraživanjem klasifikacijske oznake, bilo pretraživanjem riječima prirodnog jezika, koje su povezane s klasifikacijskim oznakama.

Klasifikacija se može koristiti u obje ove funkcije i pritom ima nekoliko, za internet važnih, dodatnih prednosti. Veličina rječnika velikih općih klasifikacijskih sustava obično zadovoljava preciznost potrebnu u specijaliziranim informacijskim servisima. Deweyjeva decimalna klasifikacija u 21. izdanju ima više od 20.000 pojmova, Univerzalna decimalna klasifikacija nakon revizije 2000. godine ima 63.000 pojmova koji se pritom mogu neograničeno kombinirati, a Klasifikacija Kongresne knjižnice koja je po prirodi enumerativna i teži navesti sve kombinacije pojmova ima, ovisno o specijalnim podjelama koje se broje, između 200.000 i 400.000 pojmova. Predmetni direktoriji koji se koriste nekom od tih klasifikacija, rijetko da koriste više od prve tri osnovne podjele, a osim GERHARD-a[12], nijedan od servisa koji koristi UDK, ne rabi složene brojeve.

Spomenuti veliki opći klasifikacijski sustavi u ovom pogledu imaju prednost pred drugim, manje poznatim klasifikacijama jer su dostupni u elektroničkom obliku, a ustanove i organizacije koje ih distribuiraju, redovito ih revidiraju i razvijaju. Važna prednost ovdje spomenutih klasifikacija jest njihova raširenost i manje-više standardna primjena u velikom broju zemalja i na velikom broju jezika. Internet na kojemu je u početku prevladavao engleski jezik, sadrži danas velik broj vrijednih informacija na svim jezicima svijeta, čime se u općim servisima za pretraživanje pristup informacijama bitno ograničava na područje jezika na kojem se obavlja pretraživanje. Kad se koriste u okviru metapodataka, klasifikacijske oznake imaju stanovitu prednost u pretraživanju zbog neovisnosti o prirodnom jeziku. Ako pretpostavimo scenarij da dokument na japanskom, ruskom ili kineskom sadrži i opis sadržaja sažet u klasifikacijsku oznaku, onda bilo koji dokumentu nalik objekt može biti lociran, a komunikacija između korisnika i pretraživača može teći na bilo kojem jeziku na kojemu je prijevod te klasifikacijske oznake dostupan. Sličan scenarij, primjerice, podržava GERHARD koji omogućuje predmetno pretraživanje na francuskom, engleskom i njemačkom, okosnica kojega je klasifikacija.

Postojeća i buduća uloga klasifikacije na internetu, pomno je studirana i opisana 1995. godine u prvoj fazi europskog projekta DESIRE.[13] U okviru projekta analizirani su predmetni direktoriji koji su se zasnivali na intelektualnoj odnosno manualnoj klasifikaciji izvora informacija na internetu poput EEVL-a (Edinburgh Engineering Virtual Library, http://www.eevl.ac.uk/), EELS-a (Engineering Electronic Library, Sweden, http://www.ub2.lu.se/eel/), NetFirst Databasea, http://www.oclc.org/oclc/netfirst/netfirst.htm, SOSIG-a (Social Science Information Gateway), NISS-a (National Information Services and Systems, http://www.niss.ac.uk), ADAM-a (Art, Design, Architecture & Media Information

[12]German Harvest Automated Retrieval and Directory - GERHARD servis je za pronalaženje informacija na njemačkom dijelu Interneta (http://www.gerhard.de).

[13]Development of a European Service for Information on Research and Education - DESIRE (http://www.desire.org) veliki je europski projekt koji se odvija u nekoliko faza, od kojih je druga bila od 1998. do 2000. Projekt uključuje deset ustanova iz Nizozemske, Norveške, Švedske i Velike Britanije. Cilj projekta izgradnja je kompleksne informacijske mreže za istraživačku europsku zajednicu.

Gateway, http://adam.ac.uk/advanced/dsearch.html) i OMNI-ja (Organising Medical Networked Information, http://www.omni.ac.uk).

Na osnovi spomenutih prednosti, u izvještaju projekta DESIRE zaključeno je da je klasifikacija nezamjenjiva u pregledavanju i omogućavanju sistematskog pregledavanja i uspostavljanju hijerarhijskih veza među predmetnim područjima (Koch, 1997.). Unatoč nekim površnostima i neujednačenostima u prikazu pojedinih klasifikacija, izvještaj je ovog projekta imao važnu ulogu u popularizaciji knjižnične klasifikacije u široj internetskoj informacijskoj zajednici. Druga faza projekta nastoji potom integrirati manualno izgrađene predmetne direktorije u velike automatski generirane predmetne indekse. Napori oko postizanja jednako kvalitetnih, a pritom automatiziranih specijaliziranih informacijskih servisa koji bi podržavali pretraživanje svih predmetnih područja istodobnim pretraživanjem i pregledavanjem različitih klasifikacijskih struktura, trenutačno je žarište više istraživačkih projekata u svijetu. Da bi ovaj pristup postao održiv, potrebno je dalje razvijati i ujednačavati, ali i primjenjivati standarde za izradu metapodataka koji bi omogućili dovoljno kvalitetan opis izvora informacija i njihovu razmjenu. Trenutačno se dosta studija provodi na povezivanju postojećih standarda *(metadata mapping i classification mapping),* a dosta je dobrih rezultata postignuto u razvoju programa automatske klasifikacije. Jedna od potencijalno najvažnijih karika u lancu internetske tehnologije jest razvoj standarda Resource Description Framework (RDF)[14] koji nudi ljudski i strojno čitljivi razumljivi format koji može podržavati i povezivati više različitih standarda metapodataka i osigurati podršku potrebnu u kontroliranju i sintakse i semantike pojedinih standarda. Da bi to bilo moguće, RDF umjesto HTML-a koristi XML kao precizniji, proširivi i moćniji jezik za kodiranje strukture i izgleda elektroničkih dokumenata. RDF omogućava identificiranje svakoga pojedinog elementa iz određene sheme metapodataka. Istodobno, može se referirati, tj. može se dovesti u relaciju svaki element s nekim unaprijed zadanim semantičkim okvirom ili popisom *(semantic registry)* koji može biti dio samog dokumenta ili smješten na bilo kojem dijelu interneta. To je možda mogući scenarij za implementaciju klasifikacijskih shema pomoću kojih bi se mogli interpretirati i pretraživati prirodnim jezicima sadržaji dokumenata prisutnih na internetu pod uvjetom da oni u svom izvornom kodu sadrže klasifikacijsku oznaku.

Zahvaljujući projektima NORDIC,[15] DESIRE, CORC,[16] SCORPION,[17] knjižnična je klasifikacija prepoznata kao izvrsno pomagalo u sažimanju sadržaja dokumenta u jezično neovisnu oznaku koja, podržana od kvalitetnog sustava za

[14]Resource Description Framework (RDF) model and syntax specification : W3C proposed recommendation.
http://www.w3.org/TR/1999/PR-rdf-syntax-19990105.

[15]NORDIC Metadata Project projekt je koji se (u nekoliko uzastopnih faza) bavi izradom pomagala i podrške za primjenu Dublin Core standarda metapodataka u opisu elektroničke građe. U projekt su uključene Norveška, Švedska, Danska, Island i Finska.

[16]Cooperative Online Resource Catalog (CORC) projekt je američkog OCLC-a (Online Computer Library Center) koji je započeo 1999. godine s ciljem da osigura potrebnu podršku za kreiranje metapodataka i opis elektroničke građe prema modelu kooperativne katalogizacije.

[17]SCORPION je OCLC-ov projekt koji se bavi automatskom klasifikacijom elektroničkih izvora informacija prema sustavu Deweyjeve decimalne klasifikacije.

pretraživanje, može omogućiti pretraživanje i pregledavanje izvora informacija. Posebno valja spomenuti postignuća projekata SCORPION, koji na osnovi Deweyjeve decimalne klasifikacije, i osobito GERHARD-a, koji na osnovi UDK-a, daju dobre rezultate u automatskoj klasifikaciji izvora informacija na internetu.

Automatska klasifikacija na internetu

Specijalizirani informacijski servisi zbog velikog priljeva i protoka informacija nastoje automatizirati većinu postupaka u odabiru, vrednovanju, provjeri postojanosti web-adrese i organizaciji informacija. Veliki napori ulažu se u razvoj programa koji mogu ekonomski isplativo podržavati brzinu, kvalitetu i pouzdanost tih servisa. UDK nije klasifikacija koja se najčešće koristi na internetu, ali primjena te klasifikacije u servisima poput SOSIG-a ili NISS-a na većoj je razini no u većine servisa koji koriste Deweyjevu decimalnu klasifikaciju ili koju drugu (Newton, 2000.). Isto vrijedi u području razvoja programa za automatsku klasifikaciju.

Automatska klasifikacija postupak je grupiranja dokumenata *(clustering)* ili njihovih surogata na osnovi doznačenih indeksnih pojmova, ili samih dokumenata na osnovi sličnosti njihova sadržaja. Obavlja se pomoću programa kreiranih da uspoređuju nizove oznaka (riječi) u određenom tekstu. Na osnovi rezultata usporedbe program izračunava stupanj sličnosti među dokumentima (ili surogatima dokumenata). Automatska klasifikacija koristi se raznim stupnjevima sofisticiranosti: ponekad za vrlo jednostavno grupiranje dokumenata na osnovi ključnih riječi u naslovu, a ponekad u sustavima zasnovanim na obradi prirodnog jezika i usporedbi s rječnikom neke postojeće klasifikacije (Ardo; Koch, 1999.).

U potonjem slučaju automatska klasifikacija gotovo je isto što i automatsko indeksiranje (tj. automatski odabir pojma označitelja za sadržaj dokumenta) jer obje metode koriste obradu prirodnog teksta da bi izlučile pojmove koji nose značenje. Dok automatsko označivanje na osnovi korpusa iz teksta automatski izlučenih riječi određuje koje od njih opisuju dokument, automatska klasifikacija uspoređuje dobivene pojmove s abecednim pojmovnikom klasifikacijske sheme i dodjeljuje klasifikacijski broj, a ne riječ.

Knjižnična klasifikacija tako ima uobičajenu primjenu, jedina je razlika u tome što klasifikaciju izvora informacija umjesto ljudi obavljaju napredni računalni programi. Preciznost i pouzdanost sustava ovisi, naravno, o dužini dokumenta i prirodi njegova jezika (znanstveni tekst u odnosu prema literarnom tekstu koji je manje precizan).

UDK pruža dobru podršku za automatsku klasifikaciju jer ima prilično velik broj pojmova. Rječnik Deweyeve decimalne klasifikacije trebao se, primjerice, znatno proširiti da bi poslužio projektu SCORPION i servisu CORC.

UDK je ujedno i prva klasifikacija koja je bila primijenjena na internetu za tu svrhu 1993. godine u okviru projekta Nordic WAIS/World Wide Web za izradu predmetnog direktorija Wide Area Information Servera (WWW Subject Tree of WAIS)(Ardo, 1995.).

Mnogo sofisticiranija automatska klasifikacija riješena je u servisu koji pruža predmetni pristup izvorima informacija na njemačkom dijelu interneta pod nazivom GERHARD. To je ujedno i najuspješniji program automatske klasifikacije, velikim dijelom stoga što je analiza teksta web-stranica njemačkog weba podvrgnuta progamu za obradu prirodnog jezika. Potom je značenje 500.000 jednostavnih i složenih UDK brojeva preuzetih iz kataloga Eidgenossischen Technischen Hochschule (ETH) u Ziirichu podvrgnuto istoj analizi da bi se dobio rječnik UDK-a koji služi kao obrazac za grupiranje i automatsku klasifikaciju. GERHARD nudi pretraživanje na engleskom, njemačkom i francuskom jeziku. Servis omogućava i pretraživanje i pregledavanje izvora informacija. Korištenje UDK-a, iako klasifikacijski brojevi nisu eksplicitno naznačeni, omogućava kretanje kroz hijerarhiju predmetnih područja i upućuje na prednost korištenja klasifikacije u mrežnom okruženju (Moller, 1999.).

Drugi programi koji za podlogu automatske klasifikacije koriste knjižnične klasifikacije, jesu američki SCORPION koji se koristi Deweyjevom decimalnom klasifikacijom i skandinavski All Engineering Web Index u okviru projekta DESIRE II. Treba spomenuti da se automatska klasifikacija nipošto ne razvija kako bi potpuno zamijenila intelektualni rad u specijaliziranim informacijskim servisima, već da bi olakšala selekciju i omogućila bolju kontrolu nad rastućim brojem informacija.

Zaključak

Broj informacija iz znanstveno-istraživačkog područja, obrazovanja, posla i drugih oblika društvene aktivnosti koje se publiciraju na internetu i intranetu, nezaustavljivo raste. Informacije na internetu često se definiraju kao amorfne, slabo strukturirane, nesamostalne, nestabilne i proizvoljne (Chan, 2000.) što bitno otežava točno i precizno lociranje relevantnih izvora. Još početkom devedesetih postalo je jasno da se zadatak otkrivanja kvalitetnih informacija za pojedine interesne zajednice korisnika ne može prepustiti nepredvidljivoj prirodi i kvaliteti općih servisa za pretraživanje informacija.

Rješenje je očigledno u okupljanju i prezentaciji izvora informacija koji se odnose na određeno područje. Posve jasno, knjižnična je klasifikacija pomagalo u organizaciji znanja koje može pomoći u organizaciji i pretraživanju informacija na internetu. Klasifikacijska struktura nudi okvir organizaciji predmetnih područja s nedvosmislenim kontekstom. Veličinom vokabulara, sintetičnim svojstvom i neovisnošću u prirodnom jeziku, opća klasifikacija poput UDK-a pruža izvanredne mogućnosti. Jednako tako, s obzirom na prirodu elektroničke građe i mogućnost da se opis dokumenta s opisom njegova sadržaja publicira i komunicira zajedno s dokumentom kojem pripada, otvara nove načine na koje se knjižnična klasifikacija može koristiti u otkrivanju i pronalaženju informacija.

Razvoj internetske tehnologije, stabilnija i robusnija arhitektura koja podržava različite strukture metapodataka te korištenje preciznih, proširivih i široko primjenjivih standarda za strukturiranje teksta poput XML-a, otvara prostor za primjenu klasifikacije tako da će, iako nevidljiva običnom korisniku, poslužiti kao okosnica za sustavno pregledavanje i navigaciju internetom.

Primjena klasifikacije u metapodacima izvora informacija na internetu tek je jedan od načina na koji se knjižničarsko iskustvo u gradnji i primjeni indeksnih jezika može iskoristiti za bolju organizaciju globalnog informacijskoga prostora.

LITERATURA[18]

Ardo, A. et al. Improving resource discovery and retrieval on the Internet : the Nordic WAIS/World Wide Web Project: summary report. http://www.lub.lu.se/W4/summary.html (1996-02-14).

Ardo, A.; T. Koch. Automatic classification applied to full text Internet documents in a robot-generated subject index. // 23rd International Online Information Meeting, London, 7-9 December 1999 : proceedings / ed. by Brian McKenna. Oxford : Learned Information Europe, 1999. Str. 239-246.

Bharat, K.; A. Broder. A technique for measuring the relative size and overlap of public Web search engines, 1997. http://www7.scu.edu.au/programme/fullpapers/1937/coml937.htm (1998-04-24).

Chan, L. M. Exploiting LCSH, LCC and DDC, to retrieve networked resources : issues and challanges. http://www.loc.gov/catdir/bibcontrol/chan_paper.html (dostupno od 2000-12-19; 2000-12-25).

Dampsey, L. The subject gateway : experiences and issues based on the emergence of the Resource Discovery Network. // Online information review 24, 1(2000), 8-23.

Green, D. The evolution of Web searching. // 23rd International Online Information Meeting, London, 7-9 December 1999 : proceedings / ed. by Brian McKenna. Oxford : Learned Information Europe, 1999. Str. 251-258.

Jenkins, Charlotte et al. Automatic RDF metadata generation for resource discovery, http://www.scit.wlv.ac.uk/~ex 1253/rdf_paper/ (2000-10-11).

Kirriemuir, J. A brief survey of Quality Resource Discovery Systems : commissioned by the JlSC-funded Resource Discovery Network Centre : final report. September 1999. http://www.rdn.ac.uk/publications/studies/survey/ (2000-09-12).

Kirriemuir, J. et al. Cross-searching subject gateways : the query routing and forward knowledge approach. // D-Lib magazine, January 1998. http://www.dlib.org/dlib/january98/01kirriemuir.html (1999-11-12).

Koch, T. Quality-controlled subject gateways : definitions, typologies, empirical overview. // Online information review 24, 1(2000), 24-35.

Koch, T. The role of classification schemes in Internet resource description and discovery : DESIRE delivery, http://www.ukoln.ac.uk/metadata/desire/classifi-cation/(1998-01-25).

Lawrence, S.; C. Lee Giles. Searching the World Wide Web. // Science 280 (1998), 98-100.

[18]Sve web-adrese navedene u ovom popisu literature dodatno su provjerene 14. rujna 2001.

Long, C. E. Improving subject searching in Web-based OPACs : evaluation of the problem and guidelines for design, Internet searching and indexing : the subject approach / ed. by Alan R. Thomas, James R. Shearer. New York [etc.] : The Haworth Information Press, 2000. Str. 159-186.

Moller, G. et al. Automatic classification of the World Wide Web using Universal Decimal Classification. // 23rd International Online Information Meeting, London, 7-9 December 1999 : proceedings / ed. by Brian McKenna. Oxford : Learned Information Europe, 1999. Str. 231-237.

Newton, R. Information technology and new directions. // The future of classification / ed. by Rita Marcella and Arthur Maltby. Aldershot: Gower, 2000. Str. 43-57.

Sherman, C. The invisible Web. // Your About.com guide to Web Search, http://websearch.about.com/library/weekly/aa061199.htm (1999-07-08).

Svenonius, E. The intellectual foundation of information organization. Cambridge, Ma.; London : The MIT Press, 2000.