ПРЕТРАЖИВАЊЕ ДИГИТАЛНЕ БИБЛИОТЕКЕ

Драгана Столић
Универзитетска библиотека
„Светозар Марковић”,
Београд

Александра Тртовац, *Проналажење информација у дигиталним библиотекама*, Београд: Универзитетска библиотека „Светозар Марковић”, 2017.

УДК: 02:004.9(049.32)
001.103.2(049.32)

Публикација *Проналажење информација у дигиталним библиотекама* Александре Тртовац представља прерађену докторску дисертацију „Дескриптори метаподатака и дескриптори садржаја у проналажењу информација у дигиталним библиотекама”, коју је ауторка одбранила 2016. године на Филолошком факултету у Београду. Објављена је у оквиру пројекта „Примена напредних технологија у библиотекама“, који је подржан од стране Министарства културе и информисања у 2017. години.

Монографија пружа систематизован преглед основних појмова окупљених око појма дигиталне библиотеке и интегрише сазнања о актуелним питањима везаним за оптимално коришћење дигиталних садржаја. Ипак, она се не фокусира само на алате и технологије који су данас у употреби, и који практично чине стандард у дигитализацији и претраживању дигиталног текста, већ пружа и историјску перспективу развоја неких библиотечких система, показујући опсег и природу нужних промена у библиотечкој делатности. Посебна вредност ове публикације јесте иницијално опредељење да се традиционална библиотека не запостави, а још мање дискредитује. Стога се и специфична улога дигиталне библиотеке посматра пре свега у односу на традиционалну, како би се јасно повукла линија разграничења, што не значи

умањење улоге класичне библиотеке и њених информационих система већ указивање на неопходно прилагођавање захтевима времена.

Као научноистраживачки рад, публикација једним делом обухвата и конкретно испитивање појединих електронских система, њихових могућности када је у питању приступ информацијама садржаним у дигиталним објектима, детаљну анализу добијених резултата и извођење закључака у складу са тим. Најзад, на самом крају, пружен је и могући модел изградње једне дигиталне библиотеке, као смерница за све оне који намеравају да се упусте у овај изазовни и захтевни посао.

Изложене дефиниције, ставови и чињенице стоје у функцији доказивања основне хипотезе да се применом напредних метода и технологија брже и прецизније проналазе информације у дигиталним библиотекама, за разлику од традиционалних и оних дигиталних библиотека у којима те методе и технологије нису примењене. Могућности за приступ информацијама у дигиталном објекту су енормно порасле, али то још увек не подразумева да су оне саме по себи доступне, те да је довољно да садржај само постоји у некаквом дигиталном облику како би информације могле да буду искоришћене. У публикацији су стога детаљно описане напредне методе које помажу да информације постану видљиве, односно да дигитални објекат буде у потпуности претражив.

ОД ТРАДИЦИОНАЛНЕ ДО ПРЕТРАЖИВЕ ДИГИТАЛНЕ БИБЛИОТЕКЕ

У анализи појединих аспеката теме полази се од дефиниција и један од првих, на овај начин протумачених појмова је онај који стоји у наслову – *проналажење информација* (енг. *information retriaval*), који укључује коришћење различитих референтних извора, нарочито оних који су „ускладиштени у рачунарским системима" (Оксфордски речник енглеског језика). Од ограничења лисних каталога дошло се до проширених могућности електронског каталога који је, захваљујући стандардним форматима за библиографски опис, структурираним кроз систем поља и потпоља, омогућио потпуније описивање публикације, те стога једноставније долажење до информација.

Ипак, у сваком од тих система, без обзира на њихову природу, важно је присуство критеријума који одређују ефикасност и употребљивост или успешност комуникације између корисника и система: *одзив* (енг. *recall*) и *прецизност* (енг. *precision*). Према наведеним дефиницијама, ови појмови говоре о односу добијених и релевантних резултата претраживања, при чему је „неопходно успостављање компромиса јер се истовремена оптимизација одзива и прецизности не може постићи". Добра структурираност одређеног система подразумева адекватну меру између ових појмова и пружа могућност да се уравнотеженост лако успостави.

Основни елементи тог процеса су атрибути који се приписују добијеном материјалу како би он био претражив и видљив. То су *дескриптори*, појединачне речи или фразе које служе да према одређеном критеријуму идентификују јединицу која се претражује, и који се деле на *дескрипторе метаподатака* и *дескрипторе садржаја*. Од посебне је важности правилно разумевање појма *метаподатак*, који се у једном делу библиотечке заједнице погрешно разуме као термин везан искључиво за електронске документе. Како А. Тртовац нарочито истиче: „каталогизација у библиотекама представља један облик доделе метаподатака, а формати UNIMARC и MARC21 са правилима која их прате (ISBD стандард, Англо-америчка правила за каталогизацију AACR2) су стандарди за метаподатке".

Метаподаци, или „информације о информацијама", могу се поделити у три групе: на *описне* или библиографске, *административне* и *структуралне*. Описни метаподаци су „најсличнији библиографским подацима у системима за каталогизацију" и омогућавају идентификацију публикације према формалним и садржинским критеријумима; административни подразумевају назнаке о правним и техничким условима коришћења објекта, као и о динамици коришћења; структурални говоре о односу појединачног објекта у односу на надређене или подређене објекте, као и о вези између различитих верзија истог документа. Када је реч о форматима метаподатака, истакнут је XML или Прошириви језик за обележавање, који има хијерархијску структуру, који је прилагодљив и применљив у различитим рачунарским системима и чије познавање представља императив у савременом библиотекарству, али и RDF, који садржи напреднију синтаксу и наглашава везу између елемената података, што чини основу развоја семантичког веба. Поред тога, детаљније се описују најраспрострањенији и кориснички оријентисани формати који су прерасли у стандарде метаподатака, као што су Даблинско језгро (Dublin Core), MODS, METS, LOM, TEI и метаподаци у пројектима Европеана. За све њих публикација *Проналажење информација*, детаљним описом и дефинисањем, олакшава правилно разумевање и разграничавање.

Посебан задатак представљају дескриптори садржаја којима се отвара нова целина у процесу проналажења информација унутар дигиталног објекта. Да би то било могуће, објекат мора бити *индексиран*, што значи да се неком садржају, или деловима текста, придружују термини односно идентификатори који ће помоћи да корисник дође до те информације приликом претраге. Код индексирања се разликује оно извршено „додељивањем термина и индексирање екстраховањем термина из текста", и уочавају се разлике, односно предности и ограничења ручног и аутоматског индексирања.

Предуслов за успешно аутоматско индексирање је квалитетно пречишћавање текста, како би скенирани садржај, који чини тек груби, при-

марни материјал који треба обрадити, могао да буде претражив, односно како би могао да чини сегмент дигиталне библиотеке. Технологије које се обавезно примењују су: оптичко препознавање карактера (Optical Character Recognition – OCR), оптичка сегментација чланака (Optical Layout Recognition – OLR), нарочито важна за обраду новинских колекција, и препознавање именованих ентитета (Named Entity Recognition – NER), технологија „намењена лоцирању појмова у тексту и њиховом класификовању у одређене категорије". Управо овај последњи подразумева развој, надоградњу и имплементацију одговарајућих језичких алата који су неопходни за успешност препознавања.

ЈЕЗИЧКИ АЛАТИ

Преглед језичких алата, мање или више познатих, и илустрација њихових могућности, обезбеђују овој публикацији посебну информативну вредност, будући да многи од њих нису познати у библиотечкој средини или је њихово познавање недовољно, па стога и нецелисходно. С друге стране, када се узме у обзир да пројекти дигитализације културне баштине, нарочито у библиотекама, постају све опсежнији, јасно је колико је важно потпуније упознавање са напредним методама којима се формирају и развијају стандарди у дигитализацији које треба на време упознати.

Како се напомиње, технологију за препознавање именованих ентитета у српском језику развија Група за језичке технологије Математичког факултета Универзитета у Београду. Неки од таквих алата су: СрпКор или Корпус савременог српског језика, који данас броји око 5.000 текстова и око 122 милиона корпусних речи, као и морфолошки речници српског језика, које је такође развила поменута Група, а који је заснован на стандарду за електронске морфолошке речнике развијене у лабораторији LADL Универзитета у Паризу и који узима у обзир све специфичности српског језика (употребу два писма, деривацију, промене облика и сл.). Добијени речник је настао како на основу традиционалних извора, тако и „екстракцијом информација из процесираних текстова", а укупан број одредница у електронском речнику српског језика је око 125.000, док број облика прелази 4 милиона.

Српски WordNet је лексичка база заснована на светском WordNet систему, развијана у оквиру потпројекта BalkaNet. WordNet је сада глобално усмерена база „која је организована преко чворова и релација између тих чворова творећи тако семантичку мрежу". Речник је, дакле, конципиран као скуп „синсетова" или скупова речи који у неком контексту имају исто значење, и зато представља значајан вишејезички алат. На другој страни од овог речника општег типа, стоје доменски, терминолошки речници, односно речници из библиотекарства и информатике, који имају незамењи-

ву улогу међу језичким алатима. За ово истраживање коришћен је „Библиотекарски терминолошки речник" Љ. Ковачевић, В. Иљац и Д. Бегенишић (2004, 2. изд. 2014), а један део овог рада посвећен је управо предложеним допунама постојећег речника на основу анализе изабраног корпуса (текстови објављени у часопису *Инфотека*).

Поред наведених, међу језичким алатима стоје и Корпуси поравнатих текстова који подразумевају паралелизацију, односно упаривање текстова на различитим језицима, што је значајно јер омогућава „претраживање на различитим језицима које резултира и релевантним погоцима на другим језицима, а доприноси и критеријуму вишејезичности, као и бољој видљивости нашег језика у виртуелном свету". На крају, детаљније је приказан начин рада алата Библиша који је „веб апликација помоћу које се претражује терминима у текстовима чланака објављеним у електронским часописима који излазе на два језика" и у којем су примењене посебне технологије.

Сва изложена појмовна разграничења и прегледи алата омогућавају да на прави начин буде усвојено обимно истраживање које следи, а које је посвећено испитивању изабраних електронских система (каталога, портала Еуропеана и др.) и њиховој процени према усвојеним критеријумима одзива и прецизности. Поред овог, посебан сегмент је посвећен терминологији, са намером да се, после спроведене анализе одређених доменских корпуса, предложи допуна постојећих терминолошких речника неким вишечланим терминима. Оба истраживања, детаљно приказана и јасно образложена, могу послужити као огледни пример спровођења сличних анализа у библиотечко-информационој делатности.

Читав рад усмерен је, пре свега, према оцртавању једног другог модела – могућег модела једне дигиталне библиотеке, који је дат на самом крају и који садржи сажете смернице намењене креаторима дигиталних библиотека са намером да се олакша формирање стратегије за све оне који поседују или намеравају да успоставе дигиталну библиотеку. Не би требало изгубити из вида да, иако проналажење информација преко дескриптора метаподатака углавном даје резултате, оно захтева прецизност у уносу и компетентност онога који метаподатке додељује, а да је потпун приступ информационом потенцијалу једног дигиталног објекта могућ преко дескриптора садржаја, те да у претрази преко комплетног текста кључну улогу играју морфолошки речници.