

Petra Vide Ogrin

DIGITALNA KNJIŽNICA IN NJENA ORODJA

Raziskovalno področje digitalne knjižnice je široko in lahko obsega zelo različne vidike: od organizacije podatkov, poizvedovalnega sistema, arhiviranja in hranjenja zbirk, digitalizacije in konverzije raznovrstnega gradiva v digitalno obliko do interoperabilnosti, evalvacije, standardizacije, ekonomskih, družbenih in pravnih vidikov ter mnogih drugih. Prispevek se osredotoča na tehnološki vidik digitalne knjižnice s poudarkom na pomenu, ki ga ima razvoj orodij za gradnjo in obdelavo digitalne knjižnice z besedilnim gradivom. Prispevek je razdeljen na štiri sklope: prvi poskuša omejiti pojem digitalne knjižnice in področje, ki ga pokriva; drugi poda nekaj temeljnih misli o metapodatkih in označevanju ter njihovi vlogi v procesih poizvedovalnega sistema; sledi poglavje, ki je posvečeno pregledu računalniških postopkov, potrebnih za obdelavo besedilnih zbirk, ter (ne)dosegljivosti orodij za te postopke v slovenskem jezikovnem prostoru. Zadnje poglavje poda taksonomijo temeljnih sklopov besedilnih zbirk in razloge za njihovo gradnjo, obenem pa osvetli pomen digitalne knjižnice kot raziskovalnega področja.

Digital Library and Its Tools

The area of digital library research is broad and may encompass rather diverse aspects: from data organization, the information retrieval system, digital archives and preservation, digitization and conversion of various materials into digital form up to interoperability, evaluation, standardization, economic, social and legal aspects and so on. This article focuses on the technological aspects of the digital library, with an emphasis on the significance of tool development for processing the textual material comprised by the digital library. The paper consists of four parts: in the first part an attempt is made to define the limits of the notion of the digital library and the area this covers; the second part contains some fundamental thoughts concerning metadata, mark-up and their role in the processes of the information retrieval system; the following chapter is devoted to surveying the computer procedures necessary for processing textual collections and the (in)accessibility of tools for these procedures in the Slovenian linguistic environment. The last chapter offers the taxonomy of fundamental complexes of the textual collections and the reasons for their construction, at the same time highlighting the importance of the digital library as an area of research.

Uvod

Raziskovalno področje digitalne knjižnice je raznovrstno in lahko obsega organizacijo podatkov, poizvedovalni sistem, arhiviranje in ohranjanje zbirk, digitalizacijo in konverzijo raznovrstnega gradiva v digitalno obliko, interoperabilnost, evalvacijo, standardizacijo, ekonomske, družbene ter pravne vidike in še mnoge druge. Pričujoči prispevek se osredotoča na tehnološki vidik digitalne knjižnice, s poudarkom na osvetlitvi pomena, ki ga ima razvoj orodij za gradnjo in obdelavo digitalne knjižnice z besedilnim gradivom. Pri tem se ozira zlasti na pojem digitalne knjižnice, ki temelji na principu prostega dostopa do informacij in neovirane širitve znanja, kar pomeni, da sta tako gradivo takšne zbirke kakor tehnologija, ki je potrebna za njeno gradnjo, obdelavo in delovanje, prosto dostopna.

Besedilo prispevka je razdeljeno na štiri sklope: prvi poskuša omejiti pojem digitalne knjižnice in področje, ki ga pokriva; drugi poda nekaj temeljnih misli o metapodatkih in označevanju ter njihovi vlogi v procesih poizvedovalnega sistema; sledi poglavje, ki je posvečeno pregledu računalniških postopkov, potrebnih za obdelavo besedilnih zbirk, ter (ne)dosegljivosti orodij za te postopke v slovenskem jezikovnem prostoru. Zadnje poglavje poda taksonomijo temeljnih sklopov besedilnih zbirk in razloge za njihovo gradnjo, obenem pa osvetli pomen digitalne knjižnice kot raziskovalnega področja, kjer se kaže vedno večji pomen koordiniranega razvoja tehnologije. Sklep povzame razvojne možnosti digitalne knjižnice.

1 Brezmejnost znotraj meja

Predstava o digitalni knjižnici ni nova, marveč je nastajala postopoma. Enega njenih vidnih začetkov lahko postavimo že v čas neposredno po 2. svetovni vojni, ko je Vannevar Bush v *Atlantic Monthly* objavil članek z naslovom *As We May Think*.¹ V njem je Bush predvidel sistem, ki ga je poimenoval »memex« (skovanka iz »memory extender«); ta sistem naj bi združeval razne vrste osebnih datotek kakor tudi raznovrstno gradivo knjižnic ter možnosti medsebojnih povezav. Memex bi posamezniku, zlasti znanstveniku, omogočal shranjevanje vseh njegovih knjig, spisov ali pisem, obenem pa bi bil povezan z drugimi viri – knjigami, filmi, fotografijami, umetniškimi slikami, zemljevidi – ter bi omogočal hiter dostop do iskanih informacij v shranjenem gradivu; memex naj bi bil pravzaprav nekakšen

¹ VANNEVAR BUSH, *As We May Think*, *Atlantic Monthly*, Boston 1945, 176, str. 101–108. <http://www.theatlantic.com/doc/194507/bush> (Vsi spletni naslovi v prispevku so bili dostopni dne 7. 1. 2008); Bush (1890–1974), ki je bil inženir elektrotehnike in tedaj direktor U.S. Office of Scientific Research and Development, velja za pionirja ideje o WWW.

»povečan osebni nadomestek njegovemu spominu.« Bush je tako, sicer na osnovi nekaterih že pred 2. svetovno vojno obstoječih zamisli, prvi oblikoval jasno izraženo idejo o potrebi po informacijski povezanosti in izmenjavi znanstvenih izsledkov. Hkrati je predvidel razvoj tehnologije, ki bo omogočala zbiranje, shranjevanje in hitro dostopanje do shranjenih informacij. Dvajset let kasneje je J. C. R. Licklider napisal knjigo *Libraries of the Future*,² v kateri je razmišljal o potrebnosti raziskovanja in razvoja za zares uporabno digitalno knjižnico; Lickliderjeve napovedi o razvoju digitalnega sveta so se izkazale za nenavadno natančne. Oba, tako Bush kakor Licklider, sta z razvojem podobe digitalne knjižnice postavila konceptualne začetke informacijske družbe, ki se danes živahno uresničuje v svetu WWW, kjer se srečujejo potrebe po najraznolikejših informacijah, zabavi ali znanju. Zlasti slednje je zanimivo kot skupen prostor, v katerem je mogoče najti dosežke znanstvenega raziskovanja tako s področij naravoslovnih kakor družboslovnih in humanističnih ved z vsega sveta ter jih na enem mestu vzporejati in tehtati. Tovrstna možnost ponovno generira vzpodbudo za nadaljnje raziskovanje in doseganje novega znanja, pri čemer igrajo pomembno vlogo prav digitalne knjižnice. V njihovem okrilju namreč poteka intenzivna izmenjava informacij, ki služijo znanstvenemu delu. Izkušnja svetovnega spleta in znotraj njega digitalne knjižnice implicira brezmejnost prostora, količine shranjenih podatkov in s tem povezanih možnosti širjenja informacij, pa četudi je bistvo digitalne knjižnice v njenih mejah, namreč v odločitvi za gradnjo določene zbirke (ali več zbirk).³

Pojem digitalne knjižnice ni homogen in ga je v (strokovni) literaturi mogoče zaslediti v več pomenih. Uporabljajo ga za opisovanje različnih zbirk in tako ali drugače organiziranih vsebin, vendar je ustreznost njegove rabe pogosto nejasna. Borgmanova⁴ ugotavlja, da pregled definicij izraza *digitalna knjižnica* pokaže, za kako zelo različne vsebinske predstave se uporablja. Poleg tega je svetovni splet prostor, kjer zavoljo vsakodnevnega nastajanja velikih količin novih digitalnih

² JOSEPH CARL ROBNETT LICKLIDER, *Libraries of the Future*, Cambridge 1965. Licklider (1915–1960) je ena pomembnejših osebnosti v zgodovini razvoja računalniške znanosti in prav tako eden od začetnikov razvoja ideje o internetu; po l. 1950 je bil zaposlen na Massachusetts Institute of Technology (MIT), kjer je mdr. raziskoval tudi, kako razvoj računalništva vpliva na razvoj digitalne knjižnice.

³ Zbirka: kar je nabrano, zbrano; skupek, izbor del, sistematično urejenih po določenih merilih (prim. ANTON BAJEC et al., *Slovar slovenskega knjižnega jezika*, Ljubljana 1970–1991, str. 836), za razliko od tistega, kar ni izbrano, kar ostane zunaj meja izbranega.

⁴ CHRISTINE L. BORGMAN, *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*, Cambridge 2000, str. 35.

vsebin ni mogoče imeti pregleda nad njimi.⁵ Zelo obsežna definicija, ki povzame tako tehnološki kakor uporabniški vidik, je nastala v okviru *IEEE CAIA Workshop on Intelligent Access to On-line Digital Libraries* in pravi: »Digitalna knjižnica je ustroj mehanizmov programiranja, shranjevanja in posredovanja skupaj z vsebino in programsko opremo, potrebno za reproduciranje, posnemanje in povečanje storitev, ki jih nudijo konvencionalne knjižnice, ki temeljijo na papirju in drugih sredstvih za zbiranje, katalogiziranje, iskanje in širjenje informacij. Popolna služba digitalne knjižnice mora zagotoviti vse bistvene naloge tradicionalne knjižnice in poleg tega izkoristiti dobro znane prednosti digitalnega hranjenja, iskanja in posredovanja.«⁶ Mlajšo po nastanku, pa tudi bolj jedrnat in izrecno definicijo najdemo pri Wittnu in Bainbridgeu, kjer je digitalna knjižnica pojmovana kot organizirana zbirka informacij oz. kot »osredotočena zbirka digitalnih objektov, tako besedilnih, video- ali avdio- dokumentov,⁷ skupaj z metodami za dostop in poizvedovanje ter za izbor, organizacijo in vzdrževanje zbirke.«⁸

Težavnost definicije pojma digitalne knjižnice izhaja torej iz heterogenosti področja, ki ga pojem označuje. Težko bi govorili tudi o dosledni, enotni in konsenzualni klasifikaciji zvrsti digitalnih knjižnic. Digitalne knjižnice so lahko izvorno digitalne (angl. *born digital*) ali posledično nastale s procesom digitalizacije tiskanega gradiva. Če povzamem po knjigi Chowdhuryja in Chowdhuryjeve,⁹ se digitalne knjižnice med seboj lahko razlikujejo glede na nastanek: njihova gradnja lahko poteka v okviru določenega programa ali ustanove (nacionalne knjižnice, univerzitetne knjižnice) ali v določenem časovnem obdobju ali določeni državi; razlikujemo jih lahko tudi po vrsti gradiva, ki ga predstavljajo (besedila,

⁵ Pri tem je zanimivo, da so nekatere zelo znane digitalne knjižnice obstajale že pred pojavom svetovnega spleta, npr. ena zgodnejših, nastala sredi 1980., je bila Perseus, ki je sprva uporabljala HyperCard in zatem CD-ROM, ali nekateri lokalno omejeni sistemi vzajemne katalogizacije oz. javno dostopnih katalogov (OPAC).

⁶ HENRY M. GLADNEY, *Digital Library: Gross Structure and Requirements: Report from a March 1994 Workshop*, <http://www.ifla.org/documents/libraries/net/rj9840.pdf>

⁷ Dokument je posamezna enota v digitalni zbirki, ki je opremljena z ustreznimi metapodatki in se običajno prikaže na zaslonu kot zadetek po izvedeni uporabniški iskalni zahtevi: t.j. bodisi pesem, pismo, esej, knjiga, slika/fotografija, avdio- ali videoposnetek; obseg dokumenta oz. granulacija dokumentov v digitalni zbirki je odvisna od urednikove odločitve o arhitekturi podatkov v tej zbirki glede na namen.

⁸ IAN H. WITTEN, DAVID I. BAINBRIDGE, *How to Build a Digital Library*, Amsterdam 2003, str. 6.

⁹ GOBINDA G. CHOWDHURY, SUDATTA CHOWDHURY, *Introduction to Digital Libraries*, London 2004.

slike/fotografije, avdio- ali videoposnetki); zelo pogoste so hibridne knjižnice, kjer v okviru klasične knjižnice in na temelju digitalizacije lastnega tiskanega ali rokopisnega gradiva nastane spremljajoča digitalna knjižnica (primer Kongresne knjižnice in zbirke *American memory*).¹⁰ Kljub raznovrstnosti področja je poglobitve naloge in lastnosti digitalne knjižnice mogoče strniti v naslednje:

- a) neomejen dostop: uporabniki s kateregakoli dela sveta lahko kadarkoli dostopajo do velikih količin raznovrstnih podatkov: besedila, slike, avdio- in videoposnetki (pogoj: infrastruktura, interoperabilnost);
- b) istočasen dostop več uporabnikov do istega vira;
- c) dostop do strukturiranih podatkov: glede na metapodatkovni model in označenost gradiva uporabnik lahko dostopa ne le do knjižničnega kataloga, temveč do celotnih besedil oz. posameznih delov besedilnih ali drugih dokumentov;
- č) sistem za poizvedovanje, ki lahko omogoča več načinov iskanja, npr. enostavno ali zahtevno iskanje, bodisi po metapodatkih ali podatkih (celotno besedilo, slike/fotografije, avdio- ali videoposnetki), po eni ali več zbirkah hkrati, lahko nudi večjezikovni uporabniški vmesnik, večjezikovno iskanje ipd.;
- d) neomejenost digitalnega prostora (glede na tiskani medij);
- e) medsebojna (potencialna) povezanost digitalnih knjižnic.

Diahrono gledano so sprva digitalne knjižnice obsegale zlasti elektronske knjižnične kataloge, torej bibliografske zapise oz. podatke o posameznih enotah gradiva, ki so ga hranile na svojih policah (npr. o avtorstvu, času in kraju nastanka, obsegu); tako je uporabnik dobil informacijo, v kateri knjižnici je dosegljivo gradivo, ki ga potrebuje, in nekatere druge podatke o njem: ali si ga je mogoče sposoditi na dom ali pa je izposojeno oz. je njegova uporaba morda omejena na čitalniško branje. Takšni katalogi so tudi danes nepogrešljiv segment sodobne klasične knjižnice. Bistveni premik v pojmovanju digitalne knjižnice pa predstavlja dostop do celotnih besedil, ko si torej uporabnik lahko na domačem računalniku v celoti ogleda gradivo, ki ga zanima. Kakor beremo v *How to Build a Digital Library*: »Hitro iskanje besed, skupin besed ali sosledij besed v celotnih besedilih obsežnih zbirk je temeljna funkcija digitalnih knjižnic, po kateri se razlikujejo od fizičnih knjižnic. Zmožnost iskanja po celotnem besedilu prispeva k večji vrednosti dokumentov, namenjenih raziskovalnemu delu, četudi je redko uporabna za prostočasno branje.«¹¹ Okolje digitalne knjižnice je zaradi prizadevanj za

¹⁰ <http://memory.loc.gov/ammem/index.html>

¹¹ IAN H. WITTEN, DAVID I. BAINBRIDGE (kot op. 8), str. 157.

učinkovitejše delovanje in pestrejšše storitve za svoje uporabnike tako postalo zelo dejavno raziskovalno področje za razvoj orodij za digitalno knjižnico, ki pa se ni ustavilo pri zmožnosti iskanja po celotnih besedilih, temveč napreduje v smeri iskanj vedno novih tehnoloških pristopov (prim. 3. poglavje).

2 Metapodatki in označevanje

Bistveni postopek v digitalni knjižnici je iskanje, ki ga omogoča njen iskalni ali poizvedovalni sistem. Prvi stik tega sistema z uporabnikom je uporabniški vmesnik, kamor uporabnik vpiše svojo iskalno zahtevo. Nato steče računalniški proces, ki iskalno zahtevo obdela in vrne zadetke. Koliko je zadetkov, kakšni so – kako natančni ali obsežni – in kako so prikazani na zaslonu, je odvisno od poizvedovalnega sistema. Ta pa je v temelju odvisen od tega, kako so dokumenti, ki predstavljajo osnovne enote digitalnega gradiva, organizirani in opremljeni s podatki o dokumentih. Z opremljanjem dokumentov s podatki o njih te dokumente oplemenitimo z neke vrste dodatno vsebino. To dodatno vsebino lahko dodajamo na dveh ravneh:

- a) na ravni opisa dokumenta: dokumente opremljamo s podatki o avtorju, letnici nastanka, kraja nastanka, dolžini dokumenta, in tem podatkom o dokumentu pravimo *metapodatki*; takšni metapodatki so bibliografski zapis v sistemu COBISS, zapis Dublin Core ali kolofon TEI v zapisu XML za določen dokument;¹²
- b) na ravni besedila dokumenta: celotnim besedilom (angl. *full-text*) dodajamo podatke o strukturi besedila in ta postopek imenujemo *označevanje* (angl. *mark-up*); to ne pomeni zgolj označevanja strukturnih enot besedila v smislu poglavij ali odstavkov, temveč tudi nekaterih drugih vsebin, za katere želimo, da so posebej dosegljive in jih zato izdvojimo kot strukturirane (prim. sliko 1); s strukturiranjem besedilnih podatkov omogočimo:
 - ustrezen izpis zadetkov oz. njihov prikaz na zaslonu po končani izvedeni iskalni zahtevi;
 - nadaljnje procesiranje strukturno označenih informacij: bodisi za zahtevni (angl. *advanced*) način iskanja,¹³ avtomatsko gručenje, klasifikacijo/kategorizacijo dokumentov ali razne vizualizacijske postopke.

¹² Za natančnejšo opredelitev metapodatkov na deskriptivne, strukturne in administrativne metapodatke prim. ARLENE G. TAYLOR, *The Organization of Information*, Westport 2004.

¹³ Prim. PETRA VIDE OGRIN, TOMAŽ ERJAVEC, Towards a Digital Edition of the Slovenian Biographical Lexicon. V: SELJAN, S., STANČIČ, H. (ed.): *Digital Information and Heritage: the Future of Information Sciences*, INFUTURE 2007, Zagreb 2007, str. 115–124.

```

<div corresp="MP35.f MP35.c" id="MP35.d" n="35" type="dip" lang="de">
  <head>M P 35<note place="inline" resp="ed" id="note.001" n="1"> Osnutek Zoisovega pisma Kopitarju je
    napisan s svinčnikom in težko čitljiv. 2 preganjena lista, 4 s svinčnikom popisane strani.
    Prepis Ivana Prijatelja.</note> (prepis)</head>
  <page corresp="MP35.f.1 MP35.c.1" id="MP35.d.1" n="1">
    <p corresp="MP35.c.1.1" id="MP35.d.1.1" n="1">
    <p corresp="MP35.c.1.2" id="MP35.d.1.2" n="2">Laibach 6. April 1812</p>
    <p corresp="MP35.c.1.3" id="MP35.d.1.3" n="3">
    <p corresp="MP35.c.1.4" id="MP35.d.1.4" n="4"> Auf ihr letztes von 5. Febr. und 10. Merz an
      mich, und von 21. d. M. an Karnoffl </p>
    <p corresp="MP35.c.1.5" id="MP35.d.1.5" n="5">Die Frachtspesen für die Bücher an Repitsch und an
      Supan sind eine Kleinigkeit und bezahlt, dagegen ist es wichtiger, daß die Partheien
      befriedigt, und Sie selbst beruhigt seyen! Supan hat besonders viel Freude über die
      Serbischen Editionen, und über das Ms<hi rend="sup">ter</hi> Frag. bezeigt, welches er
      fleisig mit der Gr. Version ausfüllen will. Des mehrern werden Sie wohl bald von ihm selbst,
      sowie von Seite Ravnikers, und der andern Bibelübersetzer, denen er ihren Auszug aus Dubr.
      Brief mittheilen wird, vernehmen. Er ist nun unbeschreiblich voll Eifer. Das Anekdoten von
      Bjelik ist goldwerth. Den Spruch = <foreign>homo prop. deus. disp. </foreign> werden die
      Krainer schwärzlich geben. Ihren Rath hat Supan sogleich bewilligt, und Ihr Vorhaben hat uns
      alle erfreut. Mir war von Bütners Tabellen nichts bekannt. Das Zusammentreffen ist
      wahrhaftig entzückend. Wir erwarten mit größter Sehnsucht ihre Analyse der slavischen
      Alphabete, und den projektirten Stich der neuen Buchstaben!</p>
    <p corresp="MP35.c.1.6" id="MP35.d.1.6" n="6"> Werden Sie zweyerley e und o - werden Sie die
      jerisirenden Werthe mit Buchstaben oder Akzent Zeichen - vorschlagen? Werden Sie absolut
      neue Schriftzüge für die mangelnden Buchstaben erfinden? oder doch einige des Cyr., Glagol.,
      Serb., Russischen - beibehalten? Diese und dergleichen Fragen dringen sich freilich auf!
      Aber ich gestehe noch einmal, daß mit delibiren, zweifeln und Hülfe suchen, nichts
      zustanden kommen kann. Es muß Blitzen, einschlagen und treffen! <del>Das Ungewitter zieht
      sich ja schon zusammen</del>, der entscheidende Augenblick für die slavische Muse ist
      nahe! Indeßen gewinnt ihr Vorläufer Zeit in den Annalen - oder umgekehrt geht alles zu
      Grundel</p>
    <p corresp="MP35.c.1.7" id="MP35.d.1.7" n="7"> Der Himmel erhalte ihre Augen, und Jenko's
      Geduld, beim lesen der Griechischen Ms<hi rend="sup">te</hi>. <foreign>Maneo male,</foreign>
      das die Mühe nicht unbelohnt bliebe!</p>

```

Slika 1. Izsek strukturno označenega dokumenta XML po shemi TEI
(iz *Korespondence Žige Zoisa* – <http://nl.ijs.si/e-zrc/zois/index-sl.html>)

Če so torej »dokumenti gradniki digitalne knjižnice, potem so označevanje in metapodatki njeni organizacijski elementi.«¹⁴ Metapodatki in oznake prispevajo k celotni podobi metapodatkovne arhitekture digitalne knjižnice. Sprva je zmogljivost računalnikov omogočala poizvedovalne sisteme, ki so omogočali iskanje zgolj po metapodatkih o posameznih dokumentih ali publikacijah, torej le po bibliografskih zapisih; danes je zaradi zmogljivejših računalnikov in čedalje večje hitrosti procesiranja možno tudi iskanje po celotnih besedilih, torej po vseh besedah v dokumentih. Sodobni poizvedovalni mehanizmi lahko torej procesirajo uporabnikovo iskalno zahtevo na ravni metapodatkov in strukturnih oznak ali na ravni celotnega besedila.

¹⁴ IAN H. WITTEN, DAVID I. BAINBRIDGE (kot op. 8), str. 221.

3 Računalniška obdelava besedilnega gradiva

Pripravo besedilnih dokumentov za gradnjo digitalne knjižnice lahko razdelimo na tri vrste postopkov glede na izvor gradiva in glede na stopnjo zahtevnosti obdelave podatkov:

- A. postopki digitalizacije;
- B. jezikovno-tehnološki postopki;
- C. postopki besedilnega rudarjenja.

Nekateri postopki znotraj omenjenih sklopov so neizogibni, nekateri pa pomenijo nadgradnjo običajnim storitvam digitalne knjižnice. Prvi sklop digitalizacije gradiva je nujen za tiste digitalne knjižnice, katerih gradivo ni izvorno nastalo v elektronski obliki, temveč je njihovo izhodišče tiskano gradivo. Tako gradivo je treba najprej digitalizirati. Če gre za besedila, obsega postopek digitalizacije praviloma dve fazi:

1. skeniranje, ki pomeni pretvorbo iz primarnega medija v elektronsko datoteko v grafičnem formatu (npr. TIFF);
2. postopek OCR (Optical Character Recognition), ki pretvori grafični format v besedilni format, ustrezen za nadaljnje postopke (večinoma standardizirani zapis XML), pri čemer je pogosto potrebno delno ročno popravljanje napak, ki ostanejo po avtomatski pretvorbi; čeprav so se postopki optičnega prepoznavanja znakov v zadnjih letih precej izpopolnili, je razumljivo, da je natančnost večja pri mlajših virih.

Rezultat končane prve faze so dokumenti v elektronski obliki, ki v naslednjih korakih postanejo gradniki digitalne zbirke: potrebno je domisliti podatkovno arhitekturo zbirke, metapodatkovni model, granulacijo označevanja celotnih besedil, poizvedovalni sistem, uporabniški vmesnik in indeksiranje besedil. Nato sledi drugi sklop postopkov, ki sodijo na področje jezikovnih tehnologij in prispevajo k večji natančnosti in učinkovitosti poizvedovalnega sistema. Ti so bolj ali manj odvisni od lastnosti posameznega jezika. To pomeni, da je pri razvoju algoritmov za te postopke potrebno upoštevati jezikovne posebnosti jezika, v katerem so napisane besedilne zbirke določene digitalne knjižnice.

- *Tokenizacija* razdeli besedilo na besede ali pojavnice (angl. *tokens*). Pri tem se lahko pojavljajo dvoumnosti, kajti pika ne predstavlja vedno konca stavka, temveč je lahko tudi del kratice ali okrajšave, poleg tega je raba ločil odvisna

od jezikovnih lastnosti.¹⁵ Tokenizacija torej ni jezikovno neodvisen proces in algoritem zanjo napišemo glede na potrebe in namen posameznega projekta. Pomembna je, ker predstavlja temeljno fazo računalniške obdelave besedil in so od nje odvisni nadaljnji postopki.

- *Stavčna segmentacija* je postopek razdelitve besedila na stavke, pri čemer večinoma predpostavljamo, da so ločnice med stavki ločila, zlasti pika, vprašaj in klicaj. Vendar dvoumnost pike lahko povzroči napake pri procesiranju. Stavčna segmentacija je v določeni meri prav tako odvisna od posebnosti posameznih jezikov, zlasti kar zadeva rabo ločil, zato je algoritem zanjo potrebno napisati glede na potrebe projekta.
- *Blokiranje* izključi iz nadaljnjih postopkov obdelave besedilnih dokumentov besede, ki so pomensko šibke, npr. pomožne glagole, predloge, veznike, nikalnice; algoritem za blokiranje lahko izdelamo s pomočjo seznamov blokiranih besed (angl. *stop words*). S postopkom blokiranja pospešimo delovanje poizvedovalnega sistema; vendar je uporabnost blokiranja v nekaterih primerih vprašljiva, npr. pri procesiranju besednih zvez v angleškem jeziku: iskalna zahteva, ki ne bi vrnila zadetka, če bi bil vključen proces blokiranja, bi bil Shakespeareov verz: »To be or not to be.«
- *Krnjenje* je postopek, v katerem so različne oblike določene besede zvedene na njen krn, npr. hiš za hiša, -e, -o itd. Učinek krnjenja je očiten zlasti pri iskanju informacij kot mehanizem razširitve iskalnega izraza tudi na njegove variantne oblike, ki upoštevajo npr. drug sklon ali število. Krnjenje je jezikovno odvisen postopek, zato je pri algoritmu za slovenski jezik potrebno upoštevati specifične lastnosti slovenskega jezika. Poglavitni izziv predstavlja izjemna pregibnost slovenskega jezika, torej samostalniška in pridevniška sklanjatev ter glagolska spregatev. Nekateri algoritmi za slovenski jezik sicer obstajajo,¹⁶ vendar niso prosto dostopni.
- *Lematizacija* je postopek, ki vsaki obliki besede pripiše ustrezno lemo ali osnovno obliko (npr. berem, lema = brati) in je odvisen od jezikovnih

¹⁵ SABINE SCHULTE IM WALDE, HEIKE ZINSMEISTER, *Introduction to Corpus Resources, Annotation and Access: Tokenisation and Morpho-Syntactic Annotation*. Foundational Course 18th European Summer School in Logic, Language and Information Málaga, Spain, July 31 – August 4, 2006. www.coli.uni-saarland.de/~schulte/Teaching/ESSLLI-06/Slides/morph-syn.pdf

¹⁶ Prim. MIRKO POPOVIČ, PETER WILLETT, *Processing of Documents and Queries in a Slovene Language Free Text Retrieval System. Literary and Linguistic Computing*, Oxford 1990, 5, str. 183–90; JURE DIMEC et al., *Iskalnik za slovenske in angleške dokumente na svetovnem spletu*. V: T. ERJAVEC, J. GROS (ur.): *Jezikovne tehnologije za slovenski jezik: zbornik konference*. Ljubljana 1998. <http://ml.ijs.si/isjt98/zbornik.html#dimec>

posebnosti ter še toliko bolj zanimiv v primeru zelo pregibnih jezikov, kamor sodi tudi slovenščina. Za slovenski jezik sta bili razviti dve orodji za lematizacijo: prva na Inštitutu Jožef Stefan z uporabo metod strojnega učenja na osnovi oblikoslovno označenega dela korpusa MULTEXT-East¹⁷ kot jezikovnega vira za učenje in ovrednotenje sistema.¹⁸ Drugo orodje pa je nastalo na Inštitutu za slovenski jezik Znanstvenoraziskovalnega centra SAZU,¹⁹ in sicer z uporabo pravil. Lematizacija je postopek, ki lahko poveča natančnost poizvedovalnega sistema, od njega pa so odvisni tudi nekateri nadaljnji jezikovno-tehnološki postopki.

- *Oblikoslovno označevanje* (angl. *POS tagging – part-of-speech tagging*) je postopek, ki vsaki besedi v besedilu pripiše njene oblikoslovne lastnosti, npr. samostalnik ženskega spola ednine v dajalniku. Orodja za oblikoslovno označevanje lahko uporabljajo slovnična pravila, statistično metodo ali metodo strojnega učenja oz. kombinacijo metod. Za (pol)avtomatsko označevanje besednih oblik z njihovimi oblikoslovnimi oznakami je potrebno definirati oblikoslovne kategorije, izdelati slovar, ki za posamezne besedne oblike določi njihove možne oznake, sestavljene iz oblikoslovnih kategorij, in nato z njegovo pomočjo polavtomatsko označiti besedilo. Za slovenski jezik obstajata dva nabora oblikoslovnih oznak, prvega so razvili na Inštitutu za slovenski jezik ZRC SAZU, drugega pa v okviru projekta MULTEXT-East na Inštitutu Jožef Stefan, ki upošteva mednarodne standarde.²⁰ Oblikoslovno označevanje lahko predstavlja pomembno stopnjo v pripravi besedil za nadaljnjo obdelavo, npr. postopkov besedilnega rudarjenja.

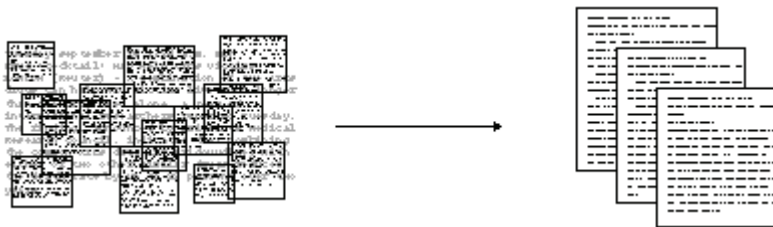
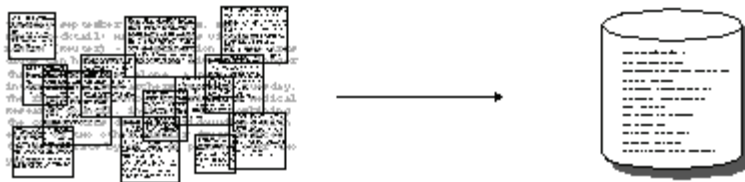
Omenjeni jezikovno-tehnološki postopki predstavljajo predobdelavo podatkov v smislu priprave besedil za določen namen, bodisi za jezikoslovne analize ali za postopke besedilnega rudarjenja. Zlasti slednji so zanimivi za področje digitalne knjižnice in jih lahko uvrstimo v zadnji, tretji omenjeni sklop postopkov, ki jih najdemo predvsem v tistih digitalnih knjižnicah, katerih dejavnost je tudi raziskovanje novih možnosti in funkcij za bolj učinkovito delovanje sistema digitalne knjižnice. Postopke besedilnega rudarjenja razlikujemo od postopkov zgolj luščenja informacij ali tudi poizvedovanja po informacijah. O poizvedovanju

¹⁷ <http://nl.ijs.si/ME/>

¹⁸ SAŠO DŽEROSKI, TOMAŽ ERJAVEC, Strojno učenje neznanih slovenskih besed. V: T. ERJAVEC, J. GROS (ur.): *Jezikovne tehnologije za slovenski jezik: zbornik konference*. Ljubljana 2000. <http://nl.ijs.si/isjt00/zbornik/sdjt00-Dzeroski21.pdf>

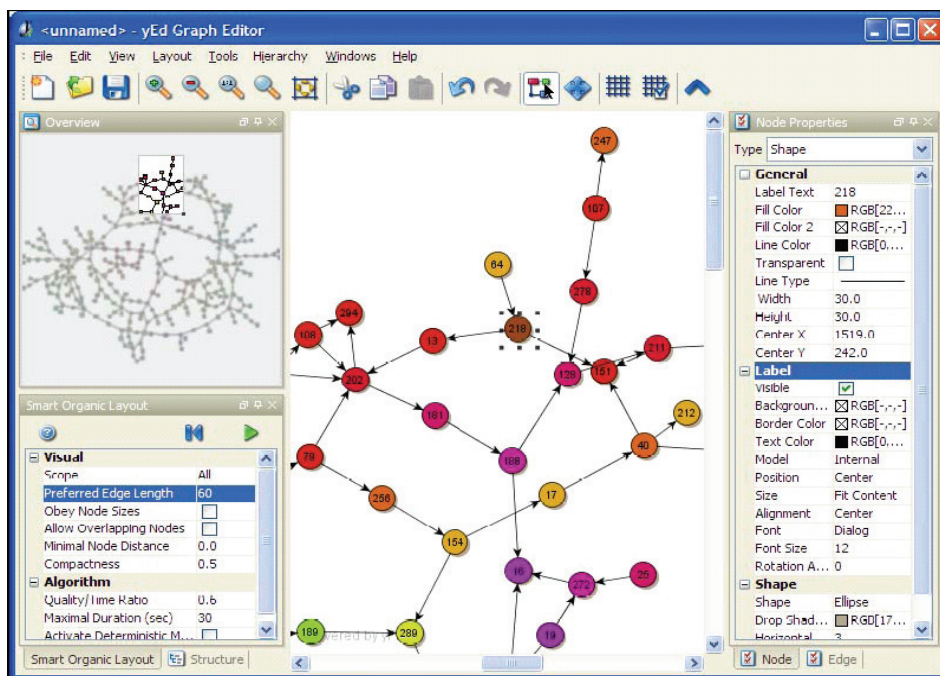
¹⁹ <http://bos.zrc-sazu.si/>

²⁰ <http://nl.ijs.si/ME/V3/msd/>

Slika 2. *Poizvedovanje po informacijah*Slika 3. *Luščenje informacij*

po informacijah (angl. *IR – Information Retrieval*) govorimo zlasti v primeru iskanja informacij prek uporabniškega vmesnika, ko uporabnik v iskalno polje vpiše iskalno zahtevo in poizvedovalni sistem po končani poizvedbi vrne bolj ali manj ustrezen izbor dokumentov, v katerih se pojavljajo iskane informacije (slika 2). Luščenje informacij (angl. *IE – Information Extraction*) pa predstavlja postopke pridobivanja posameznih informacij iz dokumentov, pri čemer so rezultat postopka luščenja ekstrahirana posamezna dejstva o enotah, dogodkih, relacijah (slika 3). Ta dejstva so nato uporabna za nadaljnje procesiranje, mdr. tudi za postopke besedilnega rudarjenja.

Besedilno rudarjenje (angl. *text mining*) pomeni postopke pridobivanja informacij, ki so relevantne, nove ali kako drugače zanimive z določenega vidika, in sicer s pomočjo strukturiranja besedila in identifikacije strukturnih vzorcev ter evalvacije in interpretacije rezultatov. Besedilno rudarjenje nadgrajuje postopke luščenja informacij, ko odkriva povezave ali medsebojne odnose med izluščenimi podatki iz raznih (mnogih) dokumentov ali digitalnih zbirk ter posledično vodi v odkrivanje ali pridobivanje novih informacij in spoznanj. Na področju besedilnega rudarjenja nastajajo orodja za postopke, kakor so klasifikacija oz. kategorizacija dokumentov (angl. *document classification/categorization*), razpoznavanje poimenovanj (angl. *Named Entity Recognition – NER*), povzemanje dokumentov (angl. *document summarization*), zlasti v smislu povzemanja več dokumentov v skupni povzetek (angl. *cross-document/multi-document summarization*), analiza



Slika 4. Primer prikaza vizualizacije dokumentov

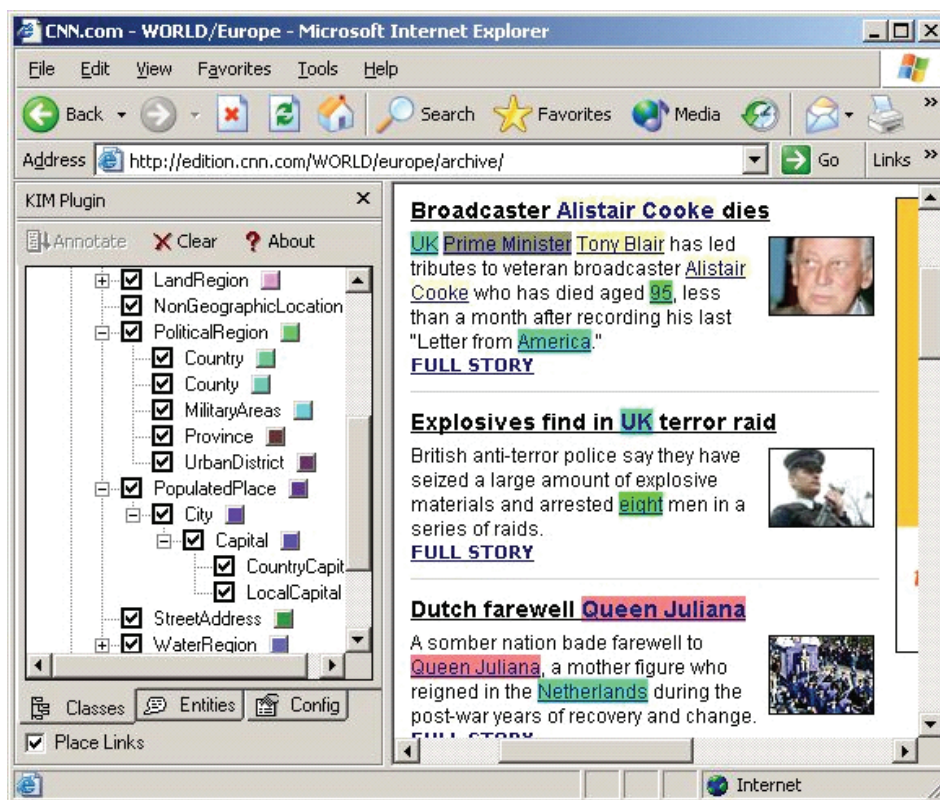
razpoloženja (angl. *sentiment analysis*), modeliranje razmerij med razpoznanimi poimenovanji (angl. *entity relation modeling*) idr.²¹

S pomočjo klasifikacije/kategorizacije dokumentov razvrstimo dokumente v kategorije, denimo glede na podobno vsebino ali skupne poteze v jezikovno-stilističnem pogledu, in jih nato gručimo v ustrezne skupine. Pri tem so lahko zanimive možnosti vizualizacijskih programov, ki uporabniku z ustreznim prikazom na zaslonu ponazorijo rezultate ali vrnjene zadetke po izvedeni iskalni zahtevi²² (slika 4). Avtomatsko razpoznavanje poimenovanj je postopek, ki iz besedil izlušči lastna imena (osebna, zemljepisna, stvarna), izraze za čas, količino, denarne vrednosti, odstotke (prim. *slika 5*);²³ izluščeni podatki lahko služijo raznim

²¹ Za pomoč in prijazen nasvet pri slovenjenju strokovne terminologije se iskreno zahvaljujem Andreji Žele in Špeli Vintar.

²² BLAŽ FORTUNA, MARKO GROBELNIK, DUNJA MLADENIĆ, Visualization of Text Document Corpus. *Informatica*, Ljubljana 2005, 29, 4, str. 497–502.

²³ Glede razvoja orodja za razpoznavanje poimenovanj za nam bližnji hrvaški jezik prim.: Božo BEKAVAC, Strojno obilježavanje hrvatskih tekstova – stanje i perspektive. *Suvremena lingvistika*, Zagreb 2002, 53/54, str. 173–182.



Slika 5. Prikaz označenih razpoznanih poimenovanj

namenom, bodisi avtomatski preslikavi v podatkovne baze ali pa jih je možno uporabiti za druge postopke besedilnega rudarjenja (klasifikacija, gručenje, modeliranje razmerij). Avtomatsko povzemanje vsebine dokumentov, posameznih ali skupine dokumentov, predstavlja zlasti prihranek časa, saj za prvi vtis o relevantnosti nekega dokumenta uporabniku ni potrebno prebrati celotnega dokumenta temveč le povzetek. Postopki za analizo razpoloženja ugotavljajo odnos avtorja besedilnega dokumenta do svojega predmeta, o katerem piše: njegovo presojo in vrednotenje ali njegov čustveni odnos do predmeta oz. čustveni učinek na bralca, ki ga avtor želi doseči. Pri postopkih za modeliranje razmerij gre za ugotavljanje razmerij ali povezav med enotami, ki jih dobimo denimo kot rezultat postopka za razpoznavanje poimenovanj.

Raziskovalno polje besedilnega rudarjenja je interdisciplinarno in sega tudi na področja procesiranja naravnega jezika, strojnega učenja, statistike, jezikovnih

tehnologij in poizvedovalnih sistemov. Čeprav gre za dokaj mlado področje, postaja razvoj orodij za besedilno rudarjenje čedalje pomembnejše tudi zato, ker je večina informacij danes shranjena v besedilni obliki, razvoj tovrstnih orodij pa povečuje učinkovitost dostopa do teh informacij v elektronski obliki.

Četudi so metodološki pristopi v razvoju tako jezikovno-tehnoloških orodij kakor tudi orodij za luščenje informacij in besedilno rudarjenje različni, seveda velja za vse računalniške postopke, da so lahko bolj ali manj uspešni, odvisno od izpopolnjenosti algoritmov. Njihova dolgoročna uspešnost pa je odvisna tudi od tega, ali postopki sledijo načelom standardizacije, interoperabilnosti in posledične trajnosti podatkov (kolikor v elektronskem mediju o slednjem lahko govorimo). Zato je na področju razvoja orodij za obdelavo in dostop do podatkov moč opaziti prizadevanja za sodelovanje in koordinacijo raziskovalnih naporov.

Številčno veliki jeziki so pri razvoju orodij za digitalno knjižnico v prednosti, saj se že vsaj od pojava svetovnega spleta (WWW) pospešeno ukvarjajo s problemom, kako doseči čimbolj učinkovito avtomatsko ekstrakcijo informacij, neodvisno od operacijskega sistema. Kljub temu John Unsworth²⁴ v poročilu TAPoR²⁵ izpred nekaj let pravi: »Trenutna ponudba orodij za besedilno analizo je mrtva. Njihova uporabnost je omejena in že leta niso bila prenovljena.« Unsworth v tem poročilu poudarja potrebo po specifičnih orodjih, ki bodo izkoriščala označenost podatkov in omogočala, da »bomo z inteligentnimi teksti delali inteligentne stvari.« Poleg tega naj bi sodobna orodja upoštevala štiri bistvene poteze kvalitetne programske opreme za obdelavo besedil:²⁶

- modularnost: sistem mora biti zgrajen iz razmeroma neodvisnih programov, od katerih vsak izvaja nabor operacij, potrebnih za procesiranje besedilnih podatkov;
- profesionalnost: moduli morajo biti zmožni podpirati resno raziskovalno delo in ne zgolj pisarniških aplikacij;

²⁴ Predstojnik Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, eden vidnejših strokovnjakov na področju razvoja orodij za digitalno knjižnico v zadnjih letih.

²⁵ Text Analysis Portal for Research, University of Alberta (TAPoR) je nastal s sodelovanjem šestih kanadskih univerz z namenom, da bi ustvarili osrednji portal, ki bi predstavljal reprezentativne prispevke in najnovejša jezikovno-analitična orodja; obenem naj bi omogočal in vzpodbujal uporabo računalniške infrastrukture v raziskovalnih projektih jezikovnih tehnologij, tekstnega označevanja, pretvorb besedil in tehnologije XML. http://tapor.uvic.ca/resources/TAPoR_Report.pdf

²⁶ Prim. C. M. SPERBERG-MCQUEEN, *Text Analysis Software Planning Meeting*, Princeton, 17–19 May, Trip Report, 1996. <http://tiger.uic.edu/~cmsmcq/trips/ceth9505.html>

The Nameless Shakespeare XML
Project Homepage: <http://www.library.northwestern.edu/shakespeare/>

Search in Texts or Find Documents

Search in Texts for: (e.g., safe) ☐ Similar Word Search

Note: For pattern matching one may employ wildcard characters (e.g., widow* retrieves widow, widowe, widowhood, etc.). To search without considering diacritics "Caps Lock" and type in all uppercase (e.g., NAIVETE finds both naïvete and naïveté). Selecting similar word search applies to all words in the database.

§ Select a Search Option:
☒ Single Term and Phrase Search (default) ☐ Phrase separated by words
☐ Proximity Searching in the same Sentence or ☐ in the same Paragraph

§ Select a Results Format: ☒ Occurrences with Context (default) ☐ Occurrences Line by Line

§ Limit your search by the following fields: (Uppercase AND, OR, and NOT serve as Boolean operators.)

Title: (e.g., Henry)

Date: (e.g., 1535 or 1510-1590)

Title ID: (e.g., A18566)

Div Objects: Type: (castlist) Head: (prologue)

SubDiv Objects: Tag: (lg OR sp OR stage) Who: (JulCaes OR Hamlet)

Type: (quatrain OR stanza) ID: (s3.1)

One may also use the bibliographic fields on their own to find documents and sort the results.

Slika 6. *PhiloLogicov uporabniški vmesnik v primeru zbirke*
The Nameless Shakespeare

- integracijo: celota modulov mora biti zmožna v celoti pokriti vse faze obdelave podatkov: zbiranje, analizo, procesiranje in prikaz;
- prenosljivost: programi in podatki morajo biti neodvisni od operacijskih sistemov, torej odporni na hitre razvojne spremembe računalniške opreme in zmožni omogočati medsebojno sodelovanje.

Kar nekaj raziskovalnih centrov po svetu se ukvarja z iskanjem novih možnosti in razvojem naprednih orodij za obdelavo in dostop do informacij. Naj tukaj omenimo zgolj dva: zanimiv je denimo projekt prav Unsworthove raziskovalne skupine, imenovan D2K (Data to Knowledge),²⁷ katerega financiranje je podprla University of Illinois in predstavlja poskus izgradnje sistema orodij za podatkovno rudarjenje in vizualizacijo za okolje digitalne knjižnice, načrtovan po načelu

²⁷ Namen projekta je avtomatizirati obdelavo nepreglednih količin podatkov, ki jih dandanes generirajo razna področja (ekonomija, vladne organizacije, znanstveno-raziskovalna področja) in tako prihraniti dragoceno človeško delo. <http://alg.ncsa.uiuc.edu/do/tools/d2k>

modularnosti; v okviru D2K je skupina razvila tudi nabor orodij za besedilno rudarjenje in vizualizacijo podatkov T2K (Text to Knowledge); prav tako zanimiv projekt za obdelavo besedilnih zbirk je PhiloLogic,²⁸ ki so ga razvili na University of Chicago v okviru projekta ARTFL (American and French Research on the Treasury of the French Language)²⁹ in razvojnega centra Digital Library Development Center (DLDC).³⁰ PhiloLogic je prav tako modularno načrtovan, novejši modul v tem sistemu pa je PhiloMine, ki že z imenom nakazuje, da je namenjen izvajanju postopkov besedilnega rudarjenja.³¹

4 Digitalna knjižnica – čemu?

»Knjižnice so po svoji naravi konzervativne organizacije,« pravi avtor knjige *Digital Libraries* William Y. Arms, toda v isti knjigi tudi beremo: »Zgodba o digitalnih knjižnicah je zgodba o spremembah.«³² Prvemu stavku današnji bibliotekarji verjetno neradi pritrdijo, in vendar: knjižnične zbirke in katalogi so pogosto nastajali ne samo desetletja, ampak celo stoletja, in spremembe so si le počasi utirale pot v njihove prostore. Hkrati pa napredni pojav digitalne knjižnice kot živahnega raziskovalnega področja v zadnjem času vzpostavlja zanimivo kontradiktornost na področju bibliotekarstva, ki obenem predstavlja tudi nevarnost, da bi se informacijski in bibliotekarski raziskovalci posvečali sicer zanimivim teoretičnim problemom, ki se porajajo v simbiozi računalniško-informacijske znanosti in uporabniškega vidika, vendar pa ob tem zanemarili tehnološki vidik in uporabne rezultate, brez katerih izgradnja in razvoj digitalne knjižnice nista mogoča.

Da bi bila digitalna knjižnica v pomoč znanstveno-raziskovalnemu delu, mora uporabniku ponuditi čim več funkcij, ki jih omogočajo razvoj in zadnji dosežki računalništva in informacijskih tehnologij. Področje, ki ta orodja razvija, je interdisciplinarno, in razumljivo je, da poteka v okoljih, kjer je na voljo velika količina besedil, hkrati pa interes znanstveno-raziskovalne sfere po naprednih raziskovalnih pripomočkih. V svetu se je uveljavil izraz digitalna humanistika (angl. *digital humanities* ali *humanities computing*) in pokriva tisto večdiscipli-

²⁸ PhiloLogic je implementiralo kar nekaj projektov, mdr. The Nameless Shakespeare (slika 6), British Women Romantic Poets, Documenting the American South, Balzac: La Comédie Humaine, Victorian Women Writers, Sanger Archive idr. <http://philologic.uchicago.edu/>

²⁹ <http://humanities.uchicago.edu/orgs/ARTFL/>

³⁰ <http://dlcdc.lib.uchicago.edu/>

³¹ CHARLES M. COONEY et al., *Extending PhiloLogic*. April 2007. <http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=175>

³² WILLIAM Y. ARMS, *Digital Libraries*, London 2000, str. 23 in 77. <http://www.cs.cornell.edu/wya/DigLib/>

narno polje, kjer se zadnjih nekaj let intenzivno srečujejo in delujejo z roko v roki humanistične vede in računalništvo, katerega začetki pa vendar sovpadajo s časom že omenjenega Bushevega memexa: namreč kmalu po koncu 2. svetovne vojne je italijanski jezuit p. Roberto Busa (1913–) dobil zamisel o orodju za iskanje po obsežnem korpusu besedil sv. Tomaža Akvinskega; projekt *Index Thomisticus* je trajal več kot 30 let, l. 1970 je delo izšlo v 56 knjigah, l. 1989 je sledila verzija na CD-ROMu, danes pa je prosto dostopna tudi spletna izdaja.³³ S pojavom svetovnega spleta je digitalna humanistika doživela pravi razmah, in sicer, kakor pišejo nekateri najvidnejši na tem področju, »zavoljo stalnega prizadevanja za uporabo informacijske tehnologije, ki naj osvetluje človekove zapise, obenem pa naj to razumevanje človekovih zapisov vpliva na razvoj in uporabo informacijske tehnologije.«³⁴ Razmah digitalne humanistike obenem sovpada prav s povečano aktivnostjo na področju digitalne knjižnice, tako pri gradnji obsežnih digitalnih zbirk kakor tudi v smislu razvoja naprednih informacijsko-tehnoloških orodij.

Pregled obstoječih digitalnih knjižnic z besedilnimi dokumenti, ki so relevantne zlasti za raziskovalce humanističnih, pa tudi nekaterih drugih ved, pokaže, da te nastajajo večidel na obzorju štirih poglavitnih tipov publikacij in iz naslednjih razlogov:

1. *Primarni viri*: ker so mnogi primarni viri stari, redki in dragoceni ter je njihova dosegljivost omejena. S spletno objavo postanejo ti viri prosto dostopni tako strokovni kakor najširši javnosti in omogočajo neomejeno širjenje znanja; obenem v svoji primarni obliki ostanejo nedotaknjeni, digitalni medij pa – morda paradokсно – skrbi za ohranjanje virov za prihodnje rodove. Viri so lahko zgolj poskenirani, tako da se uporabniku na zaslonu prikazujejo slike strani primarnega vira in uporabnik lahko »lista« po njih. Kadar so viri opremljeni s prepisi in znanstvenim aparatom, imajo tovrstne znanstvenokritične izdaje poseben status, saj je vanje vloženo večidel dolgotrajno humanistično, deloma tudi tehnološko raziskovanje ter služijo poglobljenemu vpogledu v primarno gradivo.³⁵

³³ <http://www.corpusthomisticum.org/it/index.age>

³⁴ SUSAN SCHREIBMAN, RAY SIEMENS, JOHN UNSWORTH, *The Digital Humanities and Humanities Computing: An introduction*, v: SCHREIBMANN, S., SIEMENS, R., UNSWORTH, J. (ed): *A Companion to Digital Humanities*, Malden 2004, str. XXIII.

³⁵ Prim. diplomatični, kritični, fonetični prepis v znanstvenokritični izdaji Brižinskih spomenikov: <http://nl.ijs.si/e-zrc/bs/> ter MATIJA OGRIN (ur.), *Znanstvene izdaje in elektronski medij: razprave*, Ljubljana 2005.

2. *Periodične (zlasti znanstvene) publikacije*: ker na spletu objavljene periodične publikacije služijo širjenju znanja ter hitrejšemu, udobnejšemu (od doma, z delovnega mesta) in učinkovitejšemu dostopanju do informacij: orodja za avtomatsko opremljanje dokumentov s ključnimi besedami, za kategorizacijo dokumentov, za gručenje dokumentov glede na tematsko-problemsko podobnost, za raznotere vizualizacijske postopke ipd. omogočajo izpolnjevanje izrecno definiranih uporabnikovih potreb po informacijah in pospešujejo znanstveno-raziskovalno delo.
3. *Referenčne zbirke*: ker uporaba spletnih referenčnih zbirk, kakor so slovarji, leksikoni ali bibliografije, olajša in pospeši iskanje potrebnih informacij, hkrati pa omogoča analitične postopke, ki bi bili z uporabo tiskane izdaje nemogoči: zahtevno iskanje v primeru biografskega leksikona denimo omogoča iskanje tipa »Poišči mi vse zdravnike, ki so se ukvarjali tudi s pisateljevanjem v obdobju 1830–1860.«³⁶
4. *Literarna besedila*: ker literarna besedila v elektronski obliki omogočajo lingvistično-analitične in literarno-stilistične postopke ter gradnjo referenčnih zbirk: slovarjev, pravopisa ipd.; uporabnik se običajno ne loti branja literarnega dela v elektronski obliki zavoljo literarnega doživetja in zadovoljstva. Vendar pa so e-izdaje literarnih del izjemno primerne za računalniško podprto lingvistično ter literarno-stilistično analizo: računalnik lahko obdela veliko količino podatkov v kratkem času, kar omogoča zanimive izsledke. Hkrati računalniški postopki omogočajo (pol)avtomatsko gradnjo slovarjev, tezavrov in drugih leksikalnih virov ali so podpora pri gradnji pravopisne zbirke.

Kljub raznovrstni produkciji digitalnih zbirk se vedno več pozornosti namenja koordiniranemu razvoju tehnologije za gradnjo, obdelavo in ohranjanje digitalnih virov. Digitalna knjižnica kot raziskovalna dejavnost, ki razvija napredno tehnologijo za svoje delovanje, lahko omogoča:

- ohranjanje in neomejeno razširjanje kulturne dediščine;
- inovativne pristope v znanstveno-raziskovalnem delu in izobraževalnem procesu;
- hitro diseminacijo znanja in pospešeno porajanje novega znanja;
- spodbujanje medsebojnega sodelovanja raziskovalcev v različnih oblikah iz raznih ved in raznih delov sveta.

³⁶ Prim. Oxford Dictionary of National Biography (ODNB, dostopen proti plačilu): <http://www.oxforddnb.com/>

»Tehnologija je, kolikor je uporabna, lahko v pomoč človeku v njegovi izjemni lastnosti, ki je ustvarjalni vpogled in ustvarjanje novega znanja,«³⁷ so zapisali v svojem poročilu udeleženci srečanja o razvoju digitalnih orodij za humanistiko (*Summit on Digital Tools for the Humanities*, University of Virginia, 2005), povečini strokovnjaki s področij humanistike, družboslovja, informacijske znanosti in računalništva, ki so se zbrali z namenom, da bi povzeli dotlejšnja skupna prizadevanja v razvoju orodij za obdelavo besedil in načrtali pot naprej. Še nedavno so si strokovnjaki s področja računalništva in digitalne humanistike prizadevali za uveljavitev jezika XML kot standardnega zapisa za shranjevanje podatkov. Pristojne konference in drugi sestanki v tej raziskovalni sferi ter zlasti praksa zadnjih nekaj let kažejo, da je uporaba jezika XML in pridruženih standardov postala dejstvo in izhodišče za nadaljnji razvoj. Pobud za koordiniran razvoj in nujnost izmenljivosti podatkov, podobnih prej omenjenemu *Summitu*, je kar nekaj, med njimi tudi pobuda konzorcija *Cultural Heritage Language Technologies*,³⁸ v katerem so sodelovali predstavniki osmih ustanov iz štirih držav in ga je podprla mdr. tudi Evropska komisija in ki si je postavil bolj konkretno nalogo, namreč preučiti učinkovite načine uporabe tehnologij in metod s področij računalniškega jezikoslovja, procesiranja naravnega jezika in poizvedovalnih sistemov v okolju, kjer se študentje in raziskovalci ukvarjajo z grškimi, latinskimi in staronordijskimi besedili. Četudi je bil projekt specifično omejen, je v svojem poročilu poudaril pomen upoštevanja mednarodno sprejetih standardov, npr. EAGLES za standardizacijo korpusov, protokol OAI (*Open Archives Protocol*) za izmenjavo podatkov,³⁹ protokol SOAP (*Simple Object Access Protocol*) za izmenjavo sporočil, FRBR (*Functional Requirements for Bibliographic Records*),⁴⁰ EAD (*Encoded Archival*

³⁷ <http://www.iath.virginia.edu/dtsummit/SummitText.pdf>; Institute for Advanced Technology in the Humanities Univerze v Virginiji, kjer je sestanek potekal, je eden večjih centrov, ki se ukvarjajo z raziskovanjem in razvojem naprednih IT orodij za področje humanistike, in pod svojo streho združuje vrsto zanimivih projektov: The World of Dante, Dickinson Electronic Archives, Homer's Trojan Theatre, The Complete Writings and Pictures of Dante Gabriel Rossetti idr. <http://jefferson.village.virginia.edu/>

³⁸ <http://www.chlt.org/>

³⁹ Open Archives Initiative razvija in vzpodbuja standardizacijo z namenom interoperabilnosti, ki omogoča učinkovito diseminacijo digitalnih vsebin; izhodišče pobude sta prost dostop do vsebin in urejen sistem repozitorijev, kar je sčasoma preraslo v spodbujanje čim širšega prostega dostopa do digitalnih vsebin na področjih e-raziskovanja, e-izobraževanja in e-znanosti. <http://www.openarchives.org/>

⁴⁰ <http://www.ifla.org/VII/s13/frbr/frbr.pdf>; Prevod v slov. se glasi »Funkcionalne zahteve za bibliografske zapise«; gre za holistično pojmovanje dostopanja do informacij, za entitetno-relacijski model bibliografskih zapisov in v bistvu za konceptualni načrt hierarhije entitet in relacij med

Description),⁴¹ TEI (*Text Encoded Initiative*)⁴² ter CIDOC CRM (*Conceptual Reference Model*),⁴³ s pomočjo katerih si je potrebno prizadevati za usklajena načela uporabe metapodatkovnih modelov v okviru večjih ontoloških okolij muzejev, knjižnic in arhivov. Nenazadnje tudi dejavnost pristojnih teles Evropske zveze zadnjih nekaj let na področju informacijske družbe kaže, kako pomembno vlogo posveča razvoju digitalnih vsebin in orodij z namenom graditve enotnega evropskega digitalnega prostora in s tem krepitve skupne kulturne zavesti, civilizacijske pripadnosti in znanstvenega razvoja; ena od pobud je *i2010: Digital Library Initiative* z naslovom »S klikom miške do evropske kulturne in znanstvene dediščine« (Europe's Cultural and Scientific Heritage at a Click of a Mouse), katere poglavitni rezultat naj bi bila Evropska digitalna knjižnica, ki bi omogočala enoten dostop do raznovrstnega gradiva muzejev, arhivov in knjižnic iz vseh članic Evropske zveze.⁴⁴

Sklep

Za knjižničarja Aleksandrijske knjižnice je veljalo, da si je želel zbrati, če bi le mogel, vse knjige obljudenega sveta, in da je, če je le videl ali slišal za knjigo, vredno študija, želel to knjigo tudi dobiti. In anekdota pravi, da če bi dandanašnji

njimi, ki bi zajemal načela bibliografskega opisa ne glede na posamezne katalogizacijske standarde. (prim. MAJA ŽUMER, Uporaba konceptualnega modela funkcionalne zahteve za bibliografske zapise (FZBZ) v praksi. V: STOPAR, K. (ur.): *60 let strokovnega združenja slovenskih knjižničarjev: spominski zbornik*. Ljubljana 2007, str. 105–112.

⁴¹ Encoded Archival Description je standard v zapisu XML za označevanje ali opis lastnosti in zaloge zbirk v arhivih, muzejih, specialnih knjižnicah in rokopisnih repozitorijih na način, ki bi bil strojno berljiv in torej primeren za iskanje, vzdrževanje in izmenjavo podatkov. Razvija ga Library of Congress v sodelovanju s Society of American Archivists. <http://www.loc.gov/ead/>

⁴² TEI Consortium (eds.), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Version 1.0. Last modified 28th October 2007. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/>; Text Encoding Initiative (TEI) je konzorcij, ki razvija in vzdržuje priporočila za metode označevanja strojno berljivih besedil zlasti na področju humanistike, družboslovja in jezikoslovja. Oznake TEI predstavljajo danes najpomembnejši način označevanja za spletno prezentacijo in ohranjanje besedil v raziskovalne in izobraževalne namene, na kar kaže tudi uporaba številnih projektov knjižnic, muzejev, založnikov ter posameznih raziskovalcev ali raziskovalnih skupin. <http://www.tei-c.org/index.xml>

⁴³ CIDOC CRM ali CIDOC Conceptual Reference Model razvija definicije in formalno strukturo za opisovanje implicitnih in eksplicitnih konceptov in odnosov, ki nastopajo v dokumentaciji kulturne dediščine; s tem si prizadeva za razvoj enotnega in razširljivega semantičnega okvirja, v katerega bo mogoče umestiti katero koli informacijo s področja kulturne dediščine, torej za skupni jezik strokovnjakov iz različnih ustanov s področja kulturne dediščine, tako muzejev, knjižnic ali arhivov, ki popisujejo svoje gradivo. <http://cidoc.ics.forth.gr/>

⁴⁴ http://ec.europa.eu/information_society/activities/digital_libraries/index_en.htm

vprašali bibliotekarja, kako velika naj bi bila njegova knjižnica, bi ta odgovoril: »Večja. In z možnostjo nadaljnje širitve.« Vprašanje je v sodobnem času nekoliko dvoumno, kajti z razvojem svetovnega spleta in čedalje širše zagotovljene infrastrukturne podpore je v resnici potreben razmislek, ali je navedeni odgovor našega bibliotekarja v smislu polnjenja knjižničnih polic upravičen oz. kako se temu odgovoru po drugi strani smiselno približuje pravzaprav neomejeni prostor digitalne knjižnice. Tradicionalna funkcija knjižnice je, da svojim uporabnikom posreduje informacije, ki omogočajo nova spoznanja, novo znanje. Digitalna knjižnica posreduje informacije v digitalni obliki, in sicer na način, ki omogoča hitrejši dostop do informacij, nove vpogleda v obstoječa dejstva ter obenem ohranjanje kulturne dediščine.⁴⁵

Da bi mogla digitalna knjižnica kar najbolje odgovoriti na potrebe svojih uporabnikov, se mora razvijati tudi kot samostojno raziskovalno področje, ki bo razvijalo potrebna orodja za svoje učinkovito delovanje. Področji jezikovnih tehnologij in besedilnega rudarjenja sta pomembni za razvoj in izpopolnjevanje teh možnosti, saj širita možnosti podpore poizvedovalnemu sistemu kot srčki digitalne knjižnice, ki more uporabnika privedi do želene informacije, včasih pa tudi do povsem nepričakovanih odkritij. Čeravno natančnost orodij za avtomatsko obdelavo in luščenje podatkov ni vedno popolna in je pogosto potrebno dodatno človeško delo za natančen končni rezultat, razvoj teh postopkov v primerih nekaterih številčno velikih jezikov kaže, da se izpopolnjujejo in s svojimi vedno boljšimi rezultati vabijo v nadaljnje raziskave. Kljub temu, da določena orodja za računalniško obdelavo slovenskih besedil sicer obstajajo, imamo vendar v primeri z nekaterimi drugimi jeziki na tem področju še precej razvojnih možnosti. Digitalna knjižnica predstavlja ustrezno okolje za razvoj, ki poteka v zanimivem sožitju

⁴⁵ Pri čemer ni zanemarljiv tudi ekonomski vidik, za katerega temeljit razmislek na tem mestu sicer ni dovolj prostora, vendar velja omeniti zgolj pomenljive podatke evalvacijske študije stroškov v primeru projekta digitalne knjižnice Virginia Historical Inventory (VHI), ki obsega zbirko fotografij, zemljevidov in poročil, ki dokumentirajo arhitekturno, kulturno in rodbinsko zgodovino več tisoč zgradb iz 18. in 19. st. v skupnosti s področja Virginije (http://ajax.lva.lib.va.us/F/?func=file&file_name=find-b-clas15&local_base=clas15). Študija je pokazala, da so bili začetni stroški za postavitev digitalne knjižnice (digitalizacija slikovnega in besedilnega gradiva ter ostali potrebni postopki) sicer zelo visoki, vendar pa je primerjava cene uporabe digitalne enote in fizične enote gradiva v trimesečnem obdobju pokazala, da je ustanova z odločitvijo za digitalno knjižnico celo precej prihranila, pri čemer se prihranek dolgoročno še povečuje. (Za natančne podatke prim.: Sam BYRD et al., *Cost/Benefit Analysis for Digital Library Projects: the Virginia Historical Inventory Project (VHI)*. *The Bottom Line: Managing Library Finances*, 2001, vol. 14, issue 2, 65–75. <http://www.emeraldinsight.com/Insight/viewContentItem.do;jsessionid=BE83578F8E34913F15CC89DC265FA22F?contentType=Article&hdAction=lnkpdf&contentId=860775>)

tehnike in humanistike in v katerem more tehnološki napredek prispevati k humanistični akribiji; kjer tehnika človeku pomaga odkrivati sledi, ki mu jih je zapustila njegova ustvarjalna dediščina, v katere bistvu je ujeta življenjska raznolikost in duhovno bogastvo, ali, z besedami vélikega pesnika, »vse to, kar se v vesolju razlistuje«.⁴⁶

⁴⁶ ALIGHIERI DANTE, prevedel Andrej Capuder, *Božanska komedija*, Trst 1991, str. 644.