SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2017./2018.

Andrea Bosnić

Veliki podaci i knjižnice

Završni rad

Mentor: doc. dr. sc. Tomislav Ivanjko

Zagreb, 2018.

Veliki podaci i knjižnice

Bosnić, Andrea
Odsjek za informacijske i komunikacijske znanosti
andreabosnic@yahoo.com

Sažetak

Današnje je elektroničko doba donijelo mnoge promjene u načinu funkcioniranja svih sfera života, pa tako i u poslovnom svijetu. Sve se veće količine podataka stvaraju, pohranjuju, dijele i ažuriraju na dnevnoj bazi, te su korisnici i podatci koje oni dijele došli u prvi plan interesa kompanija kako bi poboljšali kvalitetu svog poslovanja. Veliki podaci su pojam za goleme količine tih podataka koji predstavljaju materijal za analizu u svrhu donošenja boljih odluka i strategija rada same institucije. Iako se ovaj pojam najčešće vezuje uz kompanije i profit, knjižnice su važna spona između drugih institucija i javnosti, pa stoga generiraju brojne važne podatke. Budući da ih je teško obrađivati s pomoću standardnih alata, bibliotekari bi ih u skladu s trendovima trebali znati analizirati i učiniti pristupačnijim i korisnijim samim korisnicima. Ovaj završni rad bavi se upravo pojmom velikih podataka i njegove teoretske i praktične primjene u bibliotekama. Cilj ovog rada je kroz analizu drugih znanstvenih radova iznijeti primjenjivost tehnologija velikih podataka na knjižnice, te prilike i nedostatke ovog trenda u okviru bibliotekarstva.

Ključne riječi: big data, veliki podaci, knjižnice, podatak, podatkovna znanost

Abstract

Today's electronic era has brought many changes to all aspects of human life, including both the private and the business sphere. The amounts of data created, stored, shared and updated on a daily basis is continually increasing, and both the users and the data they share have come to the forefront of the company's interests to improve the quality of their business. Big data is the notion of the huge amounts of this data that represent the material for analysis in order to make better decisions and strategies of the institution itself. Although this term is most often associated with companies and profits, libraries are an important link between other institutions and the public, and therefore generate many important data. Since it is difficult to process them using standard tools, librarians should be able to analyze them and make them more accessible and more user-friendly according to trends. This paper deals with the notion of big data and its theoretical and practical application in libraries. The aim of this paper is to outline the applicability of big data technologies in libraries, and the opportunities and disadvantages of this trend within the library industry through the analysis of other scientific papers.

Key words: big data, libraries, data, data science

Sadržaj

Popis slika

Uvod

Tijekom posljednjih nekoliko godina svjedočimo brojnim, sve bržim i sve većim promjenama u svijetu tehnologije koje ujedno utječu i na način funkcioniranja tvrtki i institucija koje ih koriste. Jedna od njih je i pojava fenomena *velikih podataka* (engl. *big data*), koji označava golemu količinu podataka koja se generira i kruži elektroničkim putevima, te koja se može upotrijebiti u svrhu poboljšanja kvalitete poslovanja, odnosno razvoja strategija upravljanja. Iako se najveća količina stvara na masovno korištenim platformama, te se najvažnije poslovne strategije razvijaju u kompanijama velikih uloga i profita, ovaj se pojam može primijeniti i na razini bibliotečnog poslovanja.

Ovaj završni rad bavit će se upravo pojmom velikih podataka i njegove teoretske i praktične primjene u knjižnicama. Budući da su knjižnice važne institucije koje se bave prikupljanjem, pohranjivanjem, ažuriranjem i dijeljenjem velike količine podataka, koje se u današnje tehnološko doba sve više odvija putem elektronike, količina podataka sve je veća, a s time su potrebni i sve recentniji i prilagođeniji alati za bavljenje navedenim aktivnostima. Veliki podaci su jedan od trenutno najpopularnijih trendova koji nudi suvremene metode za obradu, pohranu i diseminaciju ove vrste podataka. Cilj ovog rada je kroz analizu drugih znanstvenih radova iznijeti primjenjivost tehnologija velikih podataka na knjižnice, te prednosti, nedostatke i prilike ovog trenda u okviru bibliotekarstva, te ponuditi preporuke za budućnost. Literatura koja će biti korištena u radu bit će iz područja informacijskih znanosti, ali i iz drugih tehničkih i društvenih disciplina.

Rad je podijeljen u šest tematskih cjelina koje se nadopunjuju i daju cjelovitiji uvid u rad s velikim podacima iz perspektive knjižničarstva. U prvom poglavlju bit će objašnjeni razlozi sve veće i brže ekspanzije količine podataka koja se događa tijekom posljednjih nekoliko desetljeća, te područja rudarenja podataka i podatkovne znanosti kao novih disciplina rada s tim podacima. U drugom poglavlju bit će definiran pojam velikih podataka, njihove karakteristike, upotreba i tehnologije, ali i etička i zakonska problematika uz koju se vezuju. U trećem poglavlju bit će iznesena usporedba velikih podataka i knjižničnih podataka, dok će četvrto poglavlje govoriti o samim knjižničarima i njihovim vještinama korisnim u radu s velikim podacima te o podatkovnim knjižničarima kao novom tipu profesionalaca. Peto poglavlje će biti posvećeno mogućnostima velikih podataka u knjižnicama, a sedmo problematici u ovom području.

1. Porast količine podataka i nastanak velikih podataka

Tijekom posljednjih desetljeća tehnologija se razvija brzinom i načinima koje laici teško mogu pratiti i razumjeti, te je od strane stručnjaka potrebna stalna posvećenost kako bi bili u tijeku s trendovima i novìnama. Tehnološki razvoj kojem svjedočimo obuhvaća širok spektar grana, proizvoda, servisa i aplikacija, a jedan od tih aspekata je i sve ubrzaniji razvoj tehnologija za pohranu, obradu i upotrebu podataka.

Mike Loukides iz tima O'Reilly izdvaja dva velika razloga ubrzanog povećanja količine podataka s kojom se u posljednje vrijeme suočavamo (O'Reilly Media, 2011, str. 5), koji možemo nazvati svojevrsnom *podatkovnom revolucijom*. Prvi je **Mooreov zakon** koji predviđa da se svake godine udvostručava broj tranzistora na jednom čipu, ali primijenjen na podatke. Od 1980-ih godina svjedočimo nevjerojatnom povećanju brzine procesora, ali i još značajnijem povećanju kapaciteta pohrane podataka, te smanjenju veličine i cijene opreme.

Najmanja jedinica podatka na računalu je nula (0) ili jedinica (1) pod nazivom bit (b). Kombinacija osam bitova se naziva *bajt* (B) (engl. *byte*), te može predstavljati jedan alfanumerički znak. To je najmanja jedinica za količinu memorije. Takvih je kombinacija moguće 256 (2^8) pa njime možemo zapisati sva slova, brojeve i specijalne znakove na tipkovnici. Mnogi znakovi poput simbola i stranih slova zauzimaju više bajtova memorije. Stoga, dokument sačinjen od 100 jednostavnih znakova koristi 100 bajtova. Nadalje, jedan kilobajt (KB) označava 1024 bajtova (B), jedan megabajt (MB) 1024 kilobajtova (KB), jedan gigabajt (GB) 1024 megabajtova (MB) i tako dalje. Odnos tih mjernih jedinica možemo vidjeti na slici 1:

SIMBOL	IME	VRIJEDNOST (u bajtima)
KB	kilobyte	1.024
MB	megabyte	1.048.576
GB	gigabyte	1.073.741.824
TB	terabyte	1.099.511.627.776
PB	petabyte	1.125.899.906.842.624
EB	exabyte	1.152.921.504.606.846.976
ZB	zettabyte	1.180.591.620.717.411.303.424
YB	yottabyte	1.208.925.819.614.629.174.706.176

Slika 1 Skala mjernih jedinica za računanje memorije
Preuzeto sa: http://ss-mbalote-porec.skole.hr/upload/ss-mbalote-porec/images/static3/1425/attachment/08.Mjerne_jedinice_za_memoriju.pdf

Jedan od prvih komercijalnih tvrdih diskova proizveden od američke tvrtke IBM[1] iz 1956. godine imao je kapacitet pohrane 5 MB podataka te je bio smješten u ormaru veličine luksuznog hladnjaka. S druge strane, današnja standardna microSD kartica koja može pohraniti 8, 16, 32, 64, 128, 200, 256 ili najrecentnije čak 512 GB (Pavlić, 2018) podataka zauzima površinu od samo 11x15 milimetara i teži pola grama (O'Reilly Media, 2011, str. 6). Odnos veličina možemo vidjeti na slici 2:

Slika 2 Usporedba veličine IBM Disk Drive-a od 5MB iz 1956. i SanDisk microSD kartice od 64 GB iz 2013. godine
Preuzeto sa: https://www.pinterest.co.uk/pin/192177109068905397/ i https://www.balkangadgets.com/sandisk-ultra-64-gb-microsd-kartica-klasa-10-100-mbs-samo-16-eur-na-tomtop-webshopu/

Drugi je razlog ubrzanog povećanja količine podataka pojava druge generacije mrežnih tehnologija, točnije koncepta i pokreta poznatog pod nazivom **Web 2.0**, kojim su se javili novi oblici i tehnike upravljanja podacima. Web 2.0 obilježen je značajkom socijalizacije i decentralizacije, to jest promjene kontrole informacija, što znači da je doveo do sve većeg umrežavanja korisnika, nastanka blogova, *wikija*, društvenih mreža, razvoja internetske prodaje i oglašavanja, te tehnologija i algoritama koji nastoje korisnicima prilagoditi sadržaj, ali i omogućiti njihovo aktivno sudjelovanje u kreiranju tog sadržaja.

Dakle, transformacija Web-a od standardne "nakladničke/medijske" djelatnosti u interaktivnu platformu u kojoj korisnici mrežnih usluga djeluju i kao sudionici i stvaratelji sadržaja, omogućila je eksponencijalni rast količine podataka koja se generira i kruži

[1] Skraćeno od *International Business Machines*

elektroničkim putevima. U to su uključeni i podaci o pretraživanju korisnika, njegovim preferencijama i slično, koje određene kompanije mogu iskoristiti u svrhu razvoja svoje poslovne strategije. No, za ovaj nov tip podataka su potrebne nove, suvremene tehnike za analizu i upravljanje.

Priča se dalje razvila pojavom pametnih telefona i mobilnih aplikacija koje za sobom ostavljaju izrazito bogat trag podataka, uključujući lokaciju, video i audio sadržaje, osobne podatke i kontakte. Skidajući aplikaciju na mobilni telefon, dopuštamo vlasnicima aplikacije da rudare našim podacima te ih koriste u vlastite svrhe. Taj proces prikupljanja informacija o svemu što nas okružuje te transformacija u format za lakše prebrojavanje i analizu naziva se *datafikacija* (engl. *datafication*) (Kocijan, 2014, str. 6). Upravo ta velika količina podataka s kojom se u posljednje vrijeme susreću stručnjaci i analitičari naziva se veliki podaci.

1.1 Rudarenje podataka

Rudarenje podataka ili *podatkovno rudarenje* (engl. *data mining*) pojavilo se 1980-ih godina, no procvat je doživjelo tek tijekom 1990-ih godina pa sve u 21. stoljeće. To je multidisciplinarno područje koje obuhvaća različite grane poput umjetne inteligencije, tehnologija baza podataka, *strojnog učenja* (engl. *machine learning*), statistike, dohvata informacija, vizualizacije podataka, i tako dalje (Han i Kamber, 2000), a odnosi se na rad s digitalnim podacima. Ono označava tehnike pronalaska obrazaca u velikim skupovima podataka, odnosno „traženje vrijednih informacija u velikim količinama podataka“ ili, preciznije, „istraživanje i analiza velikih količina podataka pomoću automatskih ili poluautomatskih metoda s ciljem otkrivanja smislenih pravilnosti“ (Pejić Bach, 2005, str. 183).

1.2 Podatkovna znanost

Razvojem računalne tehnologije i interneta došlo je do nastanka nove discipline zvane *podatkovna znanost* (engl. *data science*). Ova znanost znači primjenu određenih principa kako bi se iz podataka izvuklo znanje i pretvorilo podatke u *podatkovne proizvode* (engl. *data products*). Cilj *podatkovnih znanstvenika* (engl. *data scientists*) je poboljšati donošenje odluka i rad organizacija te cijelog društva (Klapwijk, 2016). Mike Loukides (O'Reilly Media, 2011, str. 6) to opisuje na sljedeći način:

„Podaci se povećavaju kako bi zauzeli prostor koji imamo za pohranu. Što je više prostora dostupno, nalazimo više podataka kojima ga možemo ispuniti. *Ispušni plinovi podataka* (engl. *data exhaust*) koji za sobom ostavljamo svaki put kada pretražujemo internet, postanemo prijatelji s nekim na Facebook-u ili obavimo kupovinu u svojem lokalnom supermarketu, pažljivo se prikuplja i analizira. Povećan kapacitet pohrane zahtjeva povećanje sofisticirane analize i upotrebe tih podataka. To je temelj *podatkovne znanosti*.“

Podatkovna znanost je srodna *informacijskoj znanosti* koja uključuje „prikupljanje, klasifikaciju, skladištenje, pronalaženje i širenje snimljenog znanja koje se tretira i kao čista i primijenjena znanost“ (Information Science, 2018), no označava samo rad s digitalnim podacima, za razliku od informacijske. Podatkovna znanost je interdisciplinarna znanost srodna rudarenju podataka, koja obuhvaća različite aktivnosti i pojmove vezane uz ekstrakciju znanja poput strojnog učenja, velikih podatka i vizualizacije podataka, statistike, računalne znanosti, i tako dalje.

2. Veliki podaci

U prethodnom su poglavlju izneseni kratki uvodi u srodna područja rudarenja podataka i podatkovne znanosti, no u čemu se ona razlikuju međusobno, ali i od velikih podataka? Podatkovna znanost je područje koje se bavi različitim metodama i alatima rada s podacima, dok rudarenje podataka predstavlja specifične analitičke i statističke metode analize podataka.

S druge strane, veliki podaci su novi oblik podataka koji se razvio kao posljedica promjena u svijetu tehnologije i elektronike. Veliki podaci predstavljaju oblik podataka koji dosad nije bio poznat čovječanstvu budući da nikada ranije nije bilo moguće stvoriti i zabilježiti tako veliku količinu podataka kao što je danas moguće digitalnim putevima. Od 2012. godine 90 posto svjetskih podataka je nastalo u posljednje dvije godine, a u roku od 15 minuta ljudi stvaraju oko 20 petabajta podataka pri čemu jedan petabajt iznosi količinu teksta od 20 milijuna ormara za spise (Wittmann i Reinhalter, 2014, str. 4). Razvoj digitalne tehnologije omogućio je nastanak alata i metoda kojima se sa znatno manjim vremenskim i financijskim troškovima može učinkovito pohraniti i obraditi golema količina podataka. Stoga, veliki podaci označavaju tip podataka koji se ne obrađuju i ne koriste zasebno već im je potreban potpuno nov i drugačiji pristup analize i upotrebe.

Za razliku od tradicionalnog shvaćanja podataka kao zasebnih jedinica koje zasebno bilježimo i analiziramo, ali i koristimo, veliki podaci označavaju veliku promjenu u samom načinu kako podatke percipiramo. Kao što Schönberger i Cukier (2013, str. 9) navode:

> „Došlo je do promjene u načinu razmišljanja o tome kako se podaci mogu koristiti. Podaci više nisu bili statični i ustajali, čija je korisnost bila gotova kada bi se svrha za koju su prikupljeni bila izvršena (...) Štoviše, podaci su postali sirovi materijal biznisa, vitalni ekonomski ulog, korišten za kreiranje novog oblika ekonomske vrijednosti. Ustvari, s pravim načinom razmišljanja, podaci se mogu pametno ponovo koristiti kako bi postali fontana inovacija i novih usluga. Podaci mogu otkriti tajne onima s poniznošću, željom i alatima za slušanje.“

U današnje suvremeno vrijeme, gotovo svaka osoba posjeduje barem jedan elektronički uređaj, od mobilnih telefona do računala, kojima čak i nenamjerno stvaraju

podatke koji mogu biti od koristi kompanijama i institucijama. Podaci poput lokacije uređaja i osobe ili povijesti pretraživanja mogu se iskoristiti za razvijanje pretpostavki korisnih za tvrtke, ali i same korisnike. Autori (ibid. 14) stoga u daljnjem tekstu navode da veliki podaci ustvari označavaju vjerojatnosti bazirane na primjeni matematike na ogromnoj količini podataka:

> „U svojoj suštini, veliki podaci znače pretpostavke. Iako je opisan kao dio grane računalne znanosti zvane umjetna inteligencija, točnije, područja zvanog strojno učenje, ova karakterizacija krivo upućuje. Veliki podaci ne znače pokušaj „učenja" računala da „razmišlja" kao čovjek. Štoviše, to znači primjenjivanje matematike na ogromne količine podataka kako bi se izvele vjerojatnosti: vjerojatnost da je e-mail poruka *spam*; da je napisana riječ „teh" trebala biti „the"; da putanja i brzina osobe koja neobazrivo prelazi ulicu znači da će ju uspjeti prijeći na vrijeme – samovozeći automobil treba samo lagano usporiti. Ključ je da ovi sistemi funkcioniraju dobro zato što su ispunjeni velikom količinom podataka na kojima mogu bazirati svoje pretpostavke."

Dakle, podaci koji se nama kao pojedincima mogu izvan konteksta činiti nevažnim i nekorisnim, kao dio veće cjeline mogu poslužiti kao vitalna komponenta u razvijanju poslovnih strategija, ali i usluga koje nam mogu biti veoma korisne. No, kakvi su to ustvari podaci?

2.1 Definicija i tri karakteristike

Iako ne postoji jedna općeprihvaćena i potpuno cjelovita definicija, Wang, Chen, Xu i Chen navode da „Veliki podaci opisuje inovativne tehnike i tehnologije za skupljanje, pohranu, distribuciju, upravljanje i analiziranje skupova podataka kojima tradicionalne metode upravljanja podacima nisu sposobne rukovati (2016, str. 1)". Iz citata je vidljivo da pojam veliki podaci ne predstavlja samo gomilu podataka, već i specifične metode i tehnologije za njihovu obradu.

Prema nekim izvorima, ovaj je koncept osmislio Doug Laney 2001. godine kao 3-D model upravljanja podacima kojima tradicionalni principi neće biti u mogućnosti upravljati.

Ovaj model ima tri glavne karakteristike poznate kao 3V-a: *volumen*, *velocitet* i *varijetet* (engl. *volume*, *velocity*, *variety*).

2.1.1 Volumen

Prema Wang i sur., prva karakteristika označava **volumen** podataka, budući da je, u usporedbi sa uobičajenim podacima, veličina skupova podataka velikih podataka golema. No, važno je napomenuti da ne postoji stroga definicija za tu veličinu, odnosno koliko veliki podaci se mogu klasificirati kao veliki podaci, stoga veličina može varirati ovisno o disciplini. Tradicionalni *software* se koristi za rukovanje kilobajtima ili megabajtima podataka, no alati za rukovanje velikim podacima moraju biti sposobni upravljati terabajtima i petabajtima skupova podataka.

2.1.2 Velocitet

Druga karakteristika, **velocitet** ili brzina, se odnosi na činjenicu da se podaci stvaraju i protječu izrazito dinamično i brzo. Iako ne postoji točna odrednica o kojoj se brzini radi, smatra se da se u ovom slučaju radi o generiranju podataka gotovo svake sekunde.

2.1.3 Varijetet

Treća karakteristika je **varijetet** ili raznovrsnost, koja čini skupove velikih podataka kompliciranijima za organizaciju i analizu. Uobičajeni tipovi podataka uglavnom su strukturirani i mogu se unijeti u tablice sa redovima i stupcima, no skupovi velikih podataka mogu obuhvaćati i nestrukturirane podatke i različite tipove podataka, poput e-mail poruka ili bilješki.

2.1.4 Ostale karakteristike

Wang, Chen, Xu i Chen (2016, str. 2) navode da se s vremenom ova karakterizacija proširila s tri na pet značajki ili 5V-a. Pridodane su karakteristike **vjerodostojnosti** (integritet podataka, engl. *veracity*), **vrijednosti** (korisnost, engl. *value*) i **složenosti** (stupanj povezanosti između struktura podataka, engl. *complexity*). Kocijan (2014, str. 3) navodi i značajku vizije (nove ideje sa starim podacima), **verifikacije** (zadovoljavaju li podaci

određeni skup specifikacija) i **validacije** (provjera je li svrha podataka zadovoljena i konzistentna). No, osnovne i najvažnije ostale su prve tri. Schönberger i Cukier (2013) pak smatraju da su ovi podaci okarakterizirani, osim svojim volumenom, i činjenicom da su često **nestrukturirani** i u neredu, te da se među njima može pronaći **korelacija**, odnosno uzročnost i uzorkovanost.

2.2 Kakvi su to podaci i kako se koriste

Kao što je ranije navedeno, veliki podaci obuhvaćaju širok spektar oblika podataka. To uključuje i standardne, strukturirane dokumente, ali i e-mail poruke, bilješke, fotografije, video i audio zapise i slično. Alati i algoritmi za analizu ovih dokumenata napretkom su tehnologije omogućili stjecanje uvida i iz dokumenata iz kojih to ranije nije bilo moguće: na primjer prepoznavanja lica iz slika ili videa, ili neobičnog ponašanja u trgovini (Gojčeta, 2013).

Podaci koje ubrajamo u velike podatke stoga obuhvaćaju sve informacije i djeliće informacija koje možemo sakupiti i analizirati u digitalnom obliku. Oni mogu potjecati sa svih elektroničkih uređaja: telefona, tableta ili računala, ali i sa mjernih stanica, elektroničkih igračaka, perilica rublja s elektronskim sustavom, automobilskog računala, alarmnih sustava ili bilo kojeg drugog uređaja sa digitalnim zapisom. Podaci stoga mogu obuhvaćati i pozive, navigaciju, stanje prometa, meteorološke podatke, i tako dalje, koji se potom mogu iskoristiti na različite načine. Upravo je raznovrsnost ovih podataka i mogućnost izvlačenja generalizacija i predikcija iz njih ono što ih obilježava.

Primjeri iskoristivosti velikih podataka su također izrazito širokog spektra. Podaci o prijašnjim letovima, cijenama i interesu mogu se iskoristiti za pretpostavljanje buduće cijene karata za pojedine letove, a s time i biti korisni kupcima u odabiru jeftinijeg leta (Schönberger i Cukier 2013, str. 9). Nadalje, ovu vrstu podataka, na primjer, upotrebljava Google u funkcioniranju svog pretraživača pomoću algoritma PageRank koji koristi podatke izvan same stranice, poput poveznica na nju, ali i algoritme koji čine pretraživanje personaliziranim tako što koriste podatke iz našeg zapisnika *kolačića* (engl. *cookies*) pretraživanja, dakle naših interesa i preferencija. Drugi primjer je Amazon, internetska trgovina, koja čuva naše pretrage te ih uspoređuje s pretragama drugih korisnika i na taj način kreira iznenađujuće prikladne preporuke za daljnju kupovinu (O'Reilly Media, 2011, str. 3). Facebook koristi podatke

korisnika u algoritmima za pretpostavku prijatelja koje potencijalno poznajemo, te za kreiranje personaliziranih reklama i objava na našem NewsFeed-u. GoogleMaps koristi podatke sa različitih izvora kako bi stvorio što točniju i kvalitetniju navigaciju, poput liste objekata, podataka o stanju u prometu i vremenskim prilikama. Stoga iz navedenog možemo da „to uključuje procjenu vrijednosti samog podatka, te stvaranje novih podataka na temelju njih – podatkovni proizvod“ (O'Reilly Media, 2011, str. 2).

Neki od primjera koristi velikih podataka koje navode Sawant i Shah (2013, str. 3) su:

- Energetske kompanije prate i kombiniraju podatke o upotrebi za poboljšanje usluge,
- Mrežne stranice i televizijski kanali prilagođavaju strategije oglašavanja na temelju demografije kućanstva i njihovih obrazaca gledanja,
- Sistemi za otkrivanje prijevara analiziraju ponašanje i aktivnosti,
- Kompanije visoke tehnologije koriste infrastrukturu velikih podataka za poboljšanje rješavanje problema, smanjenje kršenja sigurnosti i izvođenje predviđanja održavanja aplikacija,
- Analiza sadržaja društvenih mreža omogućava pojačavanje sentimenta potrošača i poboljšanje proizvoda, usluga i interakcija.

No, takvi se podaci mogu koristiti i u nekomercijalne svrhe, na primjer u domeni zdravstva, urbanog planiranja i okolišne znanosti. Jedan primjer je pronalaženje uzorka u dokumentima bolesnika kako bi se pokušalo utvrditi što je zajedničko onima koji su se uspješno izliječili ili onima koji nisu (Schönberger i Cukier, 2013, str. 16). Druga mogućnost su koncepti velikih podataka Smart Farming ili Precision Agriculture koji bi analizom raznih podataka mogli pronaći rješenje za buduće nestašice hrane uslijed sve većeg broja ljudi (Schönfeld, Heil i Bittner, 2018, str. 110).

Ovdje je važno naglasiti da je kod velikih podataka važno „što“ a ne „zašto“, to jest pronalazak uzorka i generalizacija, to jest korelacija, a ne potpuno precizne i objašnjive kauzalnosti. U mnogim je slučajevima uzrok fenomena nepoznat, pa dopuštamo da podaci govore za sebe (Schönberger i Cukier, 2013, str. 16). Ova činjenica može imati i negativne implikacije, no o tome nešto kasnije.

2.3 Razlika između standardnih i velikih podataka

Razliku između velikih podataka (VP) i standardnih podataka (SP) opisao je Berman (Kocijan 2017) na sljedeći način:

1. Ciljevi – SP daju odgovor na specifično pitanje s unaprijed određenim ciljem, a VP na raznovrsna pitanja s prilagodljivim ciljem,
2. Lokacija – SP su uglavnom unutar jedne organizacije, a VP mogu biti na različitim lokacijama,
3. Struktura i sadržaj podataka – SP su strukturirani podaci, a VP nestrukturirani,
4. Priprema podataka – SP najčešće priprema korisnik, a VP različiti stručnjaci,
5. Životni vijek – SP imaju ograničen vijek postojanja (prosječno 7 godina od završetka projekta), a VP neograničen zbog svoje višestruke namjene,
6. Mjerenja – SP se uglavnom mjere pomoću jednog protokola, a VP različitim,
7. Reproduciranje – projekti sa SP se lako mogu reproducirati, a s VP teško,
8. Financijsko ulaganje – financije uložene u projekte sa SP su relativno male, dok su ulozi u projekti s VP golemi i mogu dovesti do bankrota,
9. Introspekcija – identifikacija pojedinačnih SP olakšana je i vrši se pomoću lokacije unutar tablice, a identifikacija VP mnogo je složenija i vrši se pomoću tehnike introspekcije,
10. Analiza – kod SP moguće je vršiti analizu nad svim podacima istovremeno, dok se kod VP analiza odvija u koracima.

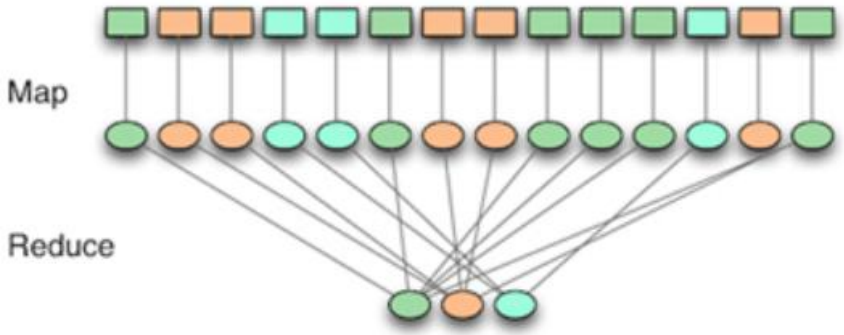
2.4 Tehnologije velikih podataka

Tehnologije velikih podataka obuhvaćaju novije alate koji dosad nisu korišteni za obradu standardnih, strukturiranih podataka manjeg volumena. Ovi se alati uglavnom baziraju na sofisticiranim algoritmima[2] kojima se nastoji analizirati na masovnoj razini, za razliku od standardnih SQL (Structured Query Languge) alata za obradu strukturiranih podataka. U ovom poglavlju iznesene su dvije tehnologije obrade velikih podataka, *MapReduce* i *Apache Hadoop*, koji nisu jedini alati, ali su najčešće spominjani i korišteni.

[2] Skupovi koraka koji se prate kako bi se riješio matematički problem ili izvršio računalni proces (Affelt, 2015, str. 51)

2.4.1 MapReduce

Google je razvio program *MapReduce* kako bi velike podatke razlomio na manje, upravljive dijelove za analizu, povezivanje i uvid. Ovaj algoritam funkcionira tako što procesuira sirove podatke pohranjene na nekoliko sistema („mapa“) u iskoristivu formu („redukcija“) kako bi mogli biti pregledani od strane programera i analitičara (Affelt, 2015, str. 45). Ključna inovacija je mogućnost poduzimanja upita preko skupa podataka, dijeljenje i pokretanje paralelno preko mnogih čvorova, to jest servera, budući da se radi o prevelikim podacima za jedan stroj (O'Reilly Media, 2011, str. 17).



Slika 3 Shema algoritma MapReduce
Preuzeto iz: O'REILLY MEDIA, 2011, *s*tr. 18

Na slici 3 možemo vidjeti dvije faze koje se nalaze u samom imenu algoritma:

1. **Mapiranje**: u prvoj fazi se ulazni podaci zasebno procesuiraju i preoblikuju u posredničke skupove podataka.
2. **Reduciranje**: u drugoj fazi se ti rezultati reduciraju u sažeti skup podataka koji predstavlja krajnji rezultat (ibid.18).

2.4.2 Hadoop

Apache Hadoop je „okvir koji dopušta distribuciju procesiranja velikih skupova podataka na klastere računala koristeći jednostavne programske modele“ (Affelt, 2015, str. 43). Zasnovan je na algoritmu MapReduce i implementiran na programskom jeziku Java, a služi za ekstrakciju podataka primjenom filtera i logike kako bi ih transformirao u upotrebljivi format. U svom radu koristi više računala i mreža, pa ima gotovo neograničen kapacitet pohrane.

2.5 Etičke i zakonske implikacije

Kao što je ranije navedeno, veliki podaci označavaju pretpostavke i vjerojatnosti koje mogu imati i negativne posljedice. Smatra se da je moguće da će u budućnosti algoritmi predviđati vjerojatnost da će, na primjer, osoba imati srčani udar i zato više platiti osiguranje, da će imati hipoteku pa će joj biti odbijen kredit ili da će počiniti zločin i čak biti unaprijed uhićena (Schönberger i Cukier, 2013, str. 16). Doba tehnologije i algoritama dovodi do pitanja etike, te odnosa slobodne volje i diktature podataka. Budući da algoritmi izvlače zaključke koji su u mnogim slučajevima potpuno validni i bazirani na golemoj količini podataka što im daje širinu, treba uzeti u obzir i činjenicu da oni iznose samo pretpostavke, temeljene na algoritmu koji je netko sastavio.

Ovo bismo etičko pitanje mogli postaviti kao pitanje privatnosti. Mnogi podaci koji se prikupljaju privatne su prirode, što predstavlja zadiranje u privatnu ili čak intimnu sferu osobe. Ti se podaci mogu koristiti u interesu neke tvrtke, biti neosigurani od tuđeg pristupa ili čak postati dostupni javnosti protiv volje pojedinca. U 2018. godini svjedočimo implementaciji europske zakonske regulative GDPR[3] za privatnost i zaštitu podataka, kojom se po prvi put na razini Europe nastoji omogućiti osobi da se njeni podaci zaborave, odnosno izbrišu (Culik i Döpke, 2018, str. 23). Princip ograničenja svrhe govori da se podaci mogu prikupljati samo ako je svrha jasno određena i zakonita, a osoba informirana i na to pristaje, te se podaci ne mogu koristiti u druge svrhe. No, kritika GDPR-a nalaže da u uredbi piše da se podaci mogu kasnije obrađivati u znanstvene, statističke i povijesne svrhe, pri čemu nije jasno o kojim se statističkim svrhama radi budući da je većina analiza velikih podataka upravo bazirana na njima, koristeći ranije prikupljene podatke (ibid. 32).

Važno je spomenuti i pitanje kvalitete podataka budući da se mnoge analize velikih podataka vrše na temelju nekvalitetnih podataka ili podataka iz nepovjerljivih izvora, ili mogu biti nerelevantni za svrhu za koju se koriste (Hoeren i Kolany-Raiser, 2018). Izuzetno je bitno i pitanje odgovornosti i vlasništva, koje se odnosi na činjenicu da pri upitnim situacijama nije posve jasno je li odgovorna sama osoba, tvrtka koja prikuplja i analizira podatke ili arhitekt algoritma, a u mnogim situacijama nepravedno štetu podnosi samo pojedinac, to jest korisnik.

[3] General Data Protection Regulation ili Opća uredba o zaštiti podataka

3. Usporedba knjižničnih podataka i velikih podataka

Posljednjih desetljeća svjedočimo mnogobrojnim ubrzanim promjenama u svijetu, pa tako i na planu knjižnica. Prema Zhan i Widen (2017, str. 134), knjižnice 2.0 su naglašavale sudjelovanje korisnika, knjižnice 3.0 upravljanje korisničkog sadržaja, dok generacija 4.0 označava ne samo dostupnost zaključaka i istraživanja, već i sistem analizira informacije te tumačenje rezultata s korisnicima. Volumen i usluge budućih knjižnica bit će masivni pa će ih se stoga moći nazivati *knjižnicama masivnih podataka*, a veliki podaci će u njima igrati važnu ulogu. Trenutni model za knjižnice se transformira u knjižnice 4.0 ili inteligentnu knjižnicu koja može analizirati informacije i prezentirati rezultate korisnicima. Značajka ove knjižnice također ima masivnost podataka kojima se bavi. Dakle, veoma važan koncept za razvoj budućih knjižnica svakako su i veliki podaci.

Koncept velikih podataka nije primjenjiv samo na kompanije kojima je cilj iskoristiti gomilu podataka u svrhu poboljšanja poslovanja i povećanja profita. Veliki podaci se sve više koriste i u bolnicama, manufakturama, sveučilištima, bankama, ali i od strane vlade kako bi ponudili nove korisne usluge ili poboljšali učinkovitost (Wang i sur., 2016, str. 2). Isti princip primjenjuju i knjižnice, čiji se podaci, način upravljanja i ponuda znatno razlikuju od koncepta velikih podataka, ali se i u mnogim aspektima podudaraju. Stoga je potrebno proučiti jesu li knjižnični podaci usporedivi sa tri glavne karakteristike velikih podataka, prema Wang i sur. (ibid. 2-3).

3.1 Volumen

Kao što je i ranije navedeno, veliki podaci su okarakterizirani velikim volumenom, odnosno veličinom koja može značajno varirati, no nerijetko iznosi desetke terabajta ili petabajta memorije. Podaci u knjižnicama ugrubo se mogu podijeliti na kataložne ili istraživačke podatke, te procesne/transakcijske podatke. Iako knjižnice uglavnom sakupljaju manje istraživačke podatke stvorene od strane pojedinačnih istraživača, kada ih se nakupi na jedno mjesto, oni mogu zauzimati čak i jednako memorije kao i veliki podaci. Na primjer, *Kongresna knjižnica* (engl. *Library of Congress*) posjeduje više od 200 terabajta mrežnih arhiva, stotine terabajta digitaliziranih novina, i petabajte podataka iz drugih izvora poput filmova i sličnih materijala (O'Reilly Media, 2011, str. 76).

Knjižnice također mogu imati i vanjske poveznice na veće mreže velikih podataka, ali i ponuditi vlastitu podatkovnu shemu baziranu na svojim zbirkama. Odnosi između ko-autora, citati, geo-lokacije, datumi, imena, klasifikacije, nakladnici, institucije, i slično, mogu se izdvojiti iz knjiga ili časopisa te se povezati s drugim djelima, ljudima, patentima, događajima, itd.

Još je jedna mogućnost da knjižnice sakupljaju procesne, odnosno transakcijske podatke o pretraživanjima korisnika ili njihovoj upotrebi knjižničnih podataka, čime bi se okupila impresivna količina podataka. Njenim povećanjem, mogli bi se analizirati uzorci, a time i poboljšati upotrebljivost, to jest omogućiti korisnicima da nalaze uzorke koji su im potrebni.

Sve u svemu, podaci u knjižnicama obuhvaćaju stotinama godina stara djela i zbirke, mnogobrojne istraživačke podatke te podatke generirane od strane korisnika pri korištenju usluga knjižnice, stoga ih se svakako može poistovjetiti s velikim podacima zbog volumena.

3.2 Velocitet

Karakteristika velociteta također je primjenjiva na knjižnice, budući da one održavaju višestruke kopije dokumenata na serverima i vrpcama, te na različitim geografskim lokacijama. Također je prisutna razmjena dokumenata između različitih organizacija, a i istraživački radovi se danas stvaraju i sakupljaju većom brzinom nego ikada prije, dinamično i svakodnevno. Stoga je potrebna i brza i učinkovita analiza kako bi ti podaci što prije bili dostupni običnim korisnicima, ali i istraživačima za vrijednu uporabu.

3.3 Varijetet

Treća karakteristika varijeteta svakako je primjenjiva na knjižnice, budući da one obuhvaćaju vrlo širok spektar oblika dokumenata: knjige, časopise, izvješća, bilješke, karte, filmove, fotografije, audio zapise, i tako dalje. Neki podaci su strukturirani, a neki nestrukturirani poput bilješki, poruka i knjiga, ili slika, audio i video zapisa. Čak i digitalni podaci zaista obuhvaćaju sve moguće oblike: od skeniranih povijesnih negativa fotografija do mikroskopskih slika jednostaničnih organizama.

Podaci koji se mogu sakupiti u knjižnici također mogu biti stvoreni i od strane korisnika, na primjer podaci o upotrebi, pretraživanjima i transakcijama u interakciji sa sustavom i uslugama. Ovi su podaci uglavnom nestrukturirani i mogu se analizirati i koristiti kao i svi drugi veliki podaci.

3.4 Ostale karakteristike

Mnogi istraživački podaci u knjižnicama veoma su neorganizirani, lošijeg opisa i u formatima koji nisu pogodni za dugoročnu ponovnu upotrebu. Također, često im nedostaje standard i format, koji ovisi o disciplinama i pojedinačnim knjižnicama. Iako neke discipline imaju standarde, u većini slučajeva oni ne postoje ili ih se autori ne pridržavaju. Još je jedan problem onaj u vezi formata. Istraživači koriste različite formate za podatke koje prikupe, čak i u slučaju više radova samo jednog autora. Ovaj problem onemogućava ili otežava integraciju takvih podataka.

4. Knjižničari i veliki podaci

Primjenjivost tehnologija velikih podataka i metoda nudi mnoge mogućnosti pojedincima u području bibliotekarstva i informacijskih znanosti. Iako u velikom dijelu ove aktivnosti uključuju programiranje i visoko poznavanje računalnih tehnologija, velik dio se odnosi i na komunikacijske i analitičke vještine koje su u suštini rada bibliotekara i informacijskih stručnjaka.

Affelt (2015, str. 126) predlaže tri moguća zanimanja pogodna za knjižničare i informacijske stručnjake u radu s velikim podacima:

- *Kustos* (engl. *curator*) – odlučivanje o tome odakle i kako će se preuzeti podaci, te upotreba tradicionalnih vještina stjecanja, selekcije i očuvanja za odabir najrelevantnijih podataka s kredibilitetom, uz upozorenja i kontekst, ali i izlaganje rezultata u čistom i iskoristivom obliku podatkovnim analitičarima ili menadžerima,

- *Podatkovni čistač* (engl. *data cleanser*) – uklanjanje krivih i dupliciranih podataka pomoću mapiranja, taksonomije, kontroliranog vokabulara i spajanja strukturiranih i nestrukturiranih podataka,

- *Menadžer podatkovnih arhiva* (engl. *data archive manager*) – zbog velikih troškova projekata, podaci se nakon završetka često ne odbacuju, stoga knjižničari mogu pomoći pri izgradnji i održavanju skladišta podataka, što uključuje pohranu, prenamjenu, indeksiranje, očuvanje i distribuciju podataka.

4.1 Usporedba podatkovnih znanstvenika i knjižničara

Stoljećima su knjižničari glavni dežurni stručnjaci za sakupljanje, pohranu, zaštitu, organizaciju i dohvat podataka i znanja sakupljenih u knjigama, dokumentima i drugim oblicima zapisa. Knjižničarstvo pripada širem znanstvenom području informacijskih znanosti stoga njihov posao obuhvaća vještine organizacije, zaštite, analize i istraživanja informacija i informacija i podataka. Upravo je značajka upravljanja podacima i informacijama u suštini njihove profesije stoga ne čudi povezanost sa suvremenijim znanstvenicima zaduženim za rad s podacima i znanjem koje se iz njih može izlučiti i upotrijebiti.

Iako s velikim podacima rade i drugi stručnjaci iz područja *poslovne inteligencije*[4] i računalnih znanosti poput programera, statističara, poslovnih analitičara i podatkovnih arhitekata (Inside Big Data Editorial Team, 2018), *podatkovni znanstvenici* predstavljaju važnu okosnicu u razumijevanju rada s ovim tipom podataka.

Podatkovni znanstvenici imaju sljedeće karakteristike (Klapwijk, 2016):

1. pitaju pitanja - istražuju i znatiželjni su,
2. pokušavaju riješiti probleme - analitičko razmišljanje i otkrića,
3. kultiviraju nove vještine - komunikacija i vizualizacija podataka.

Kao što možemo vidjeti, sve navedene karakteristike primjenjive su i na knjižničare i informacijske stručnjake. Dosad je već postalo jasno da podatkovni znanstvenici nastoje iz velike količine podataka izlučiti znanje i prezentirati ga kao podatkovni proizvod koji se može koristiti u različite svrhe. Knjižničari također rade sa velikim količinama podataka iz kojih nastoje generirati znanje za daljnju upotrebu. S obzirom na značajno podudaranje ovih područja, ali i na njihovu razliku u perspektivi, pristupu i metodama, mogućnosti za suradnju su velike. No, koje su to posebne vještine koje knjižničari imaju, a koje mogu biti korisne za rad sa velikim podacima?

4.2 Vještine knjižničara korisne u radu s velikim podacima

Prema Amy Affelt (2015, str. 22-23), osnovne vještine potrebne za rad s velikim podacima su referentne usluge, indeksiranje i sažimanje, umijeća kojima se knjižničari moraju baviti svakodnevno. Bez preciznog određivanja pitanja s obzirom na koje okupljamo i analiziramo podatke, ali i kvalitetnog indeksiranja i sažimanja pomoću kojih podatke kasnije možemo lakše locirati i povratiti, skupovi podataka gube vrijednost. Upravo su bibliotekari najveći stručnjaci u ovom području budući da su vješti i dobro upućeni u rad s metapodacima, taksonomijom i upotrebom specijalne i kontrolirane terminologije i vokabulara za ovu vrstu organizacije. No, autorica naglašava da je najvažniji razlog zbog kojeg su knjižničari i informacijski stručnjaci izuzetno korisni u radu s velikim podacima činjenica da su izvrsni u

[4] Tehnologijski upravljani proces za analizu podataka i predstavljanje djelotvornih informacija kako bi rukovoditelji, menadžeri i drugi krajnji korisnici tvrtki donosili informirane poslovne odluke (Rouse, 2017).

pričanju priča, što smatra jednim od najvažnijih načina za razumijevanje uvida unutar projekta velikih podataka. Podaci ovakvog volumena mogu biti veoma kompleksni i konfuzni, a bibliotekari ključna komponenta u njihovom interpretiranju i objašnjavanju. Oni mogu naglasiti obrasce u podacima i upozoriti da ih se ne koristi u stvaranju predviđanja. Također mogu upozoriti na činjenicu da korelacija nije uvijek isto što i kauzalnost, te smjestiti svoje opažanja u kontekst drugih ne-podatkovnih faktora koji mogu utjecati na ove uvjete. To mogu činiti pišući narativ koji će potaknuti donositelje odluka i pomoći im da razumiju kontekst podataka u smislu pravih rješenja za prave probleme, a upravo je to fundamentalno u misiji knjižničara.

Sirovi podaci su uvijek bili integralni dio posla knjižničara i informacijskih stručnjaka. U svim tipovima knjižnica, od javnih, školskih i akademskih do specijalnih, podaci se koriste u gotovo svakoj ulozi institucije. Podaci se koriste kao odgovor na većinu pitanja, te za planiranje budućih narudžbi, interesa i mjerenje korištenja knjižnice. Katalozi, indeksi i sažeci mijenjaju pisanu riječ u podatkovne točke kako bi se brže locirali i povratili potrebni materijali. Knjižničari su stručnjaci koji sadržaj materijala trebaju pravilno interpretirati, označiti i spremiti.

Osim prve tri karakteristike, koje uključuju volumen, brzinu i raznovrsnost, razmotrit ćemo mogućnosti knjižničara u domeni druge dvije karakteristike. Prva je vjerodostojnost podatka. Autorica (ibid. 35) nadalje naglašava da je verifikacija velikih podataka izuzetno važan dio procesa za knjižničare i informacijske stručnjake, u kojem oni analiziraju izvore podataka i sustave povrata informacija kako bi utvrdili kvalitetu podataka. Pritom se mogu koristiti iste vještine kao kod provjere integriteta bilo koje vrste informacija i izvora. To uključuje provjeru vjerodostojnosti izvora i je li informacija nepromijenjena od svog originalnog izvora. Ovaj korak se ne može previše naglasiti budući da o njemu ovise zaključci analize podataka. Bibliotekari se posebno razumiju u ovu aktivnost te mogu poslužiti kao kritički glasovi koji upućuju na oprez i skepticizam unutar tima koji radi na podatkovnom projektu.

Autorica navodi da se za verifikaciju velikih podataka može koristiti sljedeći kontrolni popis za provjeru podataka:

1. Koji je izvor podataka? Potrebno je istražiti reputaciju i kredibilitet izvora. Jesu li drugi koristili podatke iz istog izvora i utvrdili njihove greške?
2. Koriste li se podaci u originalnom formatu? Jesu li programirani ili transportirani? Potrebno istražiti proces programiranja kako bi se osiguralo da se podaci nisu slučajno izmijenili.
3. Je li moguće da drugi podaci koji se koriste mogu utjecati na podatke koje proučavamo?
4. Postoje li drugi izvori istih podataka? Potrebno je utvrditi jesu li potpuno isti ili različiti, te koliko.

Ovaj popis služi kao predložak, odnosno kao početna točka za svaki projekt. Temeljit pregled mogućih mana podataka može biti jedna od glavnih odgovornosti knjižničara koji rade na podatkovnom projektu. Iako statističari i voditelji projekta donose finalne odluke u vezi korištenja podataka, knjižničari mogu odigrati važnu ulogu u analizi kvalitete podataka.

Druga karakteristika, odnosno peta, je vrijednost podataka, te na njoj možemo pokazati jedinstvene vještine bibliotekarstva. Izdvajanje prave vrijednosti podataka je vrlo teško iz tri razloga: izazovno je, skupo i riskantno.

4.3 Podatkovni knjižničari

Novina u svijetu knjižničarstva su svakako i podatkovni knjižničari, odnosno „profesionalno knjižničarsko osoblje koje se bavi upravljanjem istraživačkim podacima, korištenjem istraživačkih podataka kao resursa ili potporom istraživača u tim aktivnostima“ (Information specialists and Data librarian skills, n. d.). Njihova je djelatnost u bliskom srodstvu s podatkovnim znanstvenicima budući da rade s digitalnim podacima, ali svoj rad baziraju na tradicionalnim aktivnostima knjižničara.

5. Knjižnice i veliki podaci

Tradicionalno su se knjižničari bavili podacima sakupljenim u knjigama i drugim dokumentima, a danas sa promjenama s kojima smo svi suočeni, knjižničari se susreću s podacima pohranjenim u digitalnim bazama podataka i skupovima podataka pa čak i sa sirovim podacima koji predstavljaju novost u njihovom području rada. Iz tog su razloga knjižničari primorani naći nove načine, alate i metode za upravljanje ovim informacijama.

Prema Wang, Chen, Xu i Chen (2016, str. 4) knjižnice bi trebale pronaći alate za pohranu podataka, za indeksiranje, ali i za pokretanje upita. No, knjižničari mogu savjetovati lokalne vlasnike biznisa koji traže informacije o tržištu ili pomoći studentima ili istraživačima. Neke funkcije mogu biti vrlo slične rudarenju financijskih podataka ili podataka o transakcijama, pri čemu korisnici knjižnice ovo izvode kako bi tražili reference, pa rudarenje ponašanja korisnika može dati koristan uvid u pružanje bolje usluge. Stoga treba postići dva aspekta rudarenja podataka: korištenje podataka pohranjenih u knjižnici i korištenje podataka sakupljenih tijekom procesa u kojem korisnici koriste usluge knjižnice.

5.1 Dvije vrste velikih podataka u knjižnicama

Veliki podaci omogućavaju knjižnici da bude pametna i prilagođena korisnicima nudeći personalizirane inteligentne usluge. Kao što je ranije navedeno, knjižnični veliki podaci se mogu razvrstati u dvije grupe: kataložne i procesne/transakcijske podatke (Liu i Shen, 2018). **Kataložni podaci** su podaci i informacije iz knjižničnih dokumenata, dok **procesni podaci** nastaju tijekom procesa upravljanja knjižnicom i korištenja njenih usluga ili su stvoreni od strane korisnika knjižnice.

Prva grupa podataka obuhvaća dokumentne, bibliografske i slične podatke, a druga zapisnike, korisnike i bilješke. Inovacije uključuju personalizirane usluge, preporuke i analizu ponašanja i navika korisnika, koji imaju veliku vrijednost i saznanja za knjižničare, korisnike i usluge. Knjižničarima to omogućava ponudu kvalitetnih proizvoda i usluga sa minimalnim troškovima, a korisnicima poboljšanje iskustva korištenja i zadovoljstvo. Uslužna vrijednost uključuje poboljšanje usluga i procesa kvalitete i učinkovitosti analizom knjižničnih velikih podataka u svojim raznim oblicima.

5.2 Vrste knjižnica i prilike za rad s velikim podacima

Mogućnosti rada s velikim podacima u knjižnicama svakako nisu jednake za sve institucije. Knjižnice se prema tipu organizacije i uslugama koje nude, mogu podijeliti na akademske i školske, zatim javne te specijalne knjižnice (Chowdhury, Burton, McMenemy i Poulter, 2008, str. 25-33). Glavna funkcija akademske, odnosno sveučilišne, te školske knjižnice je podržavati istraživanje i rad organizacije kojoj su posvećene, no njihove mogućnosti i financije nerijetko su ograničene. Javne knjižnice dijele sličan problem, čak i u značajnijem obujmu. Specijalne knjižnice pak označavaju rad sa specijaliziranim zbirkama pojedine organizacije, dakle nisu namijenjene ni javnosti ni akademiji. Ove knjižnice mogu biti nacionalne, vladine, bolničke, umjetničke, zatvorske, digitalne, pravne, poslovne i slično, stoga su im financijske, pa s time i sve druge mogućnosti u znatno boljoj poziciji. Mogućnosti rada knjižničara s velikim podacima sukladne su s navedenim financijskim i srodnim kapacitetima organizacije. Javne i obrazovne knjižnice uvelike se oslanjanju na tradicionalnim djelatnostima, a mnogim specijaliziranim knjižnicama u interesu je poboljšati kvalitetu poslovanja i time povećati profit ili usluge i proizvode koje nude korisnicima. Budući da implementacija tehnologija i osoblja za rad sa velikim podacima, te dodatna edukacija knjižničara zahtijeva veliku količinu novca, ovo mnogim institucijama može predstavljati značaj problem.

5.3 Mogućnosti rada s velikim podacima u knjižnicama

Wang, Chen, Xu i Chen (2016, str. 4-6) su naveli sljedeće mogućnosti velikih podataka u knjižnicama

Pristup iz perspektive podataka za donošenje odluka

Pristup temeljen na podacima (engl. *data-driven approach*) je ključan pristup knjižničnim velikim podacima. Na primjer, bazirano na transakcijama posudbe kupaca ili pretraživanjima, knjižnice mogu koristiti razne tehnike rudarenja podataka i analitike teksta kako bi optimizirali zbirke knjiga ili časopisa i tako generirali bolje rezultate pretraživanja i ponudile preporuke za knjige. Ovo bi povećalo zadovoljstvo kupaca nudeći bolju uslugu i učinkovito korištenje knjižničnih resursa.

Novi format podataka

Najvažniji ciljevi knjižnice su dugoročno dijeljenje podataka i činjenje podataka dostupnim. No, velika količina podataka se treba promijeniti, pogotovo oni sakupljeni prije mnogo vremena. Prvi korak je digitalizacija, skeniranje ili mikrofilmiranje, ali i izgradnja novih alata infrastruktura. Primjer je **ScienceBase** platforma za upravljanje podacima koja omogućuje prijenos i katalogizaciju podataka. Preoblikovanje knjižničnih podataka kako bi bolje radili sa drugim mrežnim resursima je također važno. Knjižnični podaci mogu postati povezani s drugim podacima kako bi postigli interoperabilnost na Web-u.

Standardizacija podataka i modeliranje podataka

Ključna karakteristika podataka u bazama podataka su metapodaci no oni ni standardi u slučaju istraživačkih podataka još nisu definirani. Izgradnja podataka pomoću metapodataka svakako bi potaknula dijeljenje i mijenjanje knjižničnih podataka, odnosno istraživačkih podataka. Shema podataka je jako korisna u unificiranju podataka sa različitih resursa. Na primjer, iz jednog djela poput istraživačkog rada ili knjige, mogu se izvući odnosi između ko-autora, citati, geo-lokacije, datumi i imena, klasifikacije, institucije, izdavači i povijesne cirkulacije, a kasnije i povezati s drugim djelima. Primjer je **WorldCat** platforma za razmjenu, to jest svjetski knjižnični katalog koji je izgradio OCLC (Online Computer Library Center). Radi se o najvećoj bibliografskoj bazi podataka koja obuhvaća građu 72 tisuće svjetskih biblioteka, muzeja i arhiva iz 172 države.

Vizualizacija knjižničnih podataka

Jedan od najvažnijih koraka u radu s podacima svakako je i vizualizacija podataka koja omogućava njihovu ilustraciju i prikaz na sažet, privlačan i učinkovit način. Upotrebom grafova, dijagrama, slika, tablica i karti, te različitih boja lakše možemo uočiti elemente i njihove veze, a time i pojednostaviti razumijevanje i učenje. Knjižnični podaci se mogu selektirati i vizualizirati pomoću alata poput **Tableau** *dashboard*-a kako bi se prezentirali korisnicima po njihovim potrebama. Knjižničari u sveučilišnim knjižnicama mogu koristiti vizualizaciju podataka kako bi usporedili sekcije knjižničnih zbirki s brojem predmeta. Disbalansi u zbirkama ili budžetiranju u određenim područjima mogu poslužiti za određivanje i ponudu savjeta za planiranje.

Studije ponašanja korisnika

Informacije o knjižničnim zbirkama mogu biti rudarene pomoću tehnologije velikih podataka. Moguće je zabilježiti, pratiti i pohraniti ponašanje korisnika, to jest procesne ili transakcijske podatke, te iz njih provesti podatkovnu analizu. Rezultat se može koristiti za potencijalno poboljšanje sveukupnog iskustva korisnika i zadovoljstva uslugom knjižnice. Ova bi se usluga mogla temeljiti na sličnim, već postojećim projektima kakve je, na primjer, proveo i implementirao **Amazon**, nudeći svojim korisnicima kvalitetne i personalizirane preporuke bazirane na aktivnostima i preferencijama samog korisnika.

6. Problematika velikih podataka u knjižnici

U prethodnim smo poglavljima zaključili da knjižnice sadrže vrijedne velike podatke, ali se oni razlikuju od podataka iz drugih područja te je njihovo istraživanje u domeni knjižnica relativno novo. Stoga postoje različiti problemi u procesu preoblikovanja, zaštite, analize i prezentacije podataka u bibliotekama. Wang, Chen, Xu i Chen (2016, str. 3) smatraju sljedeće promjene u knjižnicama ključnim, ali ne i jedinima, kako bi se implementirali pokušaji podatkovne znanosti u knjižnične zbirke:

- središnji podatkovni repozitorij gdje će se podaci sačuvati održavati i katalogizirati,
- podatkovni standardi koje će sakupljeni podaci slijediti,
- podatkovne zajednice koje sakupljaju održavaju i štite podatke,
- analitički alati.

Autori (ibid. 3-4) izdvajaju sljedeće temeljne probleme s kojima se susreću knjižnice u radu s velikim podacima:

Nedostatak podatkovnih znanstvenika

Ključni problem je da podatkovni analitičari trebaju znanja o statistici i računalnoj znanosti, ali i vještine iz domene znanja i sposobnosti suradnje, stoga je izazov s kojim se susreću knjižničari sposobnost da upravljaju informacijama velikih podataka, a kratki tečajevi za to nisu dovoljni.

Sposobnost preuzimanja velikih podataka

Nedostatak osoblja i platforma može biti veliki problem, a istraživanje velikih podataka u knjižnicama još i sporije nego u drugim disciplinama. Ključni razlog je da digitalne knjižnice nastoje biti samostalne organizacijske jedinice i ograditi se od novih tehnologija.

Budžet

Iako su koristi korištenja analiza velikih podataka velike, potrebna su ulaganja u IT[5] analitičke servere, te računalne poslužitelje visokih performansi, što može biti izrazito skupo, pogotovo za knjižnice koje najčešće nemaju veliki budžet. Drugi problem je i digitalizacija koja zahtijeva mnogo vremena i osoblja, a nedostatak ljudskih resursa je također prisutan. Na području jugoistočne Europe cijena implementacije hardware-a za pohranu, software-a za analizu, te osoblja može iznositi više od 200 000 američkih dolara, dok je u SAD-u taj iznos i gotovo tri puta veći (Big Data solutions: Example of the development cost, 2018). Stoga je navedeno zasad ograničeno samo na profitne organizacije i specijalizirane knjižnice.

Tehnički izazovi

Tehnike velikih podataka uključuju bilježenje, pohranjivanje, obradu i prezentiranje podataka, a podaci u knjižnicama obuhvaćaju različite tipove i oblike. Neki podaci možda tek čekaju digitalizaciju, a neki sadrže oštećene ili krive podatke što zahtjeva mnogo rada. Zbog heterogenosti tipova i formata podataka, integracija može biti veoma težak posao. Također, mnogi tipovi podataka su slabije iskoristivi u sirovom obliku nego nakon primjene filtera i algoritama ili obrade, što zahtjeva novac kako bi se izgradili alati i ponudile druge potpore.

Privatnost

Veliki podaci označavaju rudarenje podataka i otkrivanje znanja, što uključuje i problem privatnosti. Za knjižnice i knjižničare koncept privatnosti je iznimno važan jer predstavlja „slobodu pristupa svim materijalima koje pojedinac želi, bez znanja ili uplitanja drugih“ (Cooke, 2016, str. 167), to jest odnos povjerenja između knjižničara i klijenata. Ponekad se to pravo na slobodu pristupa nalazi u opoziciji s interesima drugih pojedinaca ili grupa u društvu, pa može doći do problema. Profesionalci u ovom području naglašavaju važnost kodeksa prakse i etičkog kodeksa kojim se štite korisnici i njihovi podaci, osim u slučaju vlade koja mora imati transparentan pristup.

[5] Skraćeno od informacijska tehnologija

Problem privatnosti uključuje i rizike upadanja u sistem koji se pojavljuju zbog pristupačnosti velike količine podataka, a sigurnost podataka u knjižnicama još nije razrađena. Ovo je posebno rizično kod digitalnih knjižnica, koje zahtijevaju snažniju normativnu zaštitu prava na privatnost. Pritom je važna i javna diskusija na temu *Prava na zaborav* (engl. *Right To Be Forgotten* ili RTBF), koji korisnicima omogućava da se njihovi podaci i mrežni sadržaj uklone. Maceli (ibid.) pak naglašava da knjižničari i knjižnice mogu imati važnu ulogu u edukaciji ljudi o važnosti njihove privatnosti, prijetnjama koje ih okružuju te alatima i tehnikama za zaštitu. Ovaj se problem može predstaviti i kao „smrt privatnosti", budući da se čak i anonimni podaci često prikupljaju i koriste pri analitici velikih podataka, te mogu naštetiti našim interesima, a sredstva zaštite gotovo da i nemaju učinka (ibid.).

Neprimjenjivost velikih podataka na sve organizacije

Jasno je da organizacije koje nastoje koristiti velike podatke trebaju velike investicije u IT infrastrukturu i osoblje pa male knjižnice bez dovoljno budžeta trebaju dijeliti resurse s drugim organizacijama. Stoga, budući da su veliki podaci relativno novi, tradicionalni analitički pristupi još uvijek dominiraju većinom organizacija.

Drugi uobičajeni problemi s kojima se susreću istraživači velikih podataka uključuju nerazumijevanje vremena unutar podataka, razinu velikih podataka, nemogućnost analize velikih podataka da se nosi sa ne-linearnim dinamikama, te pitanje kauzalnosti i izazove interdisciplinarnosti. S druge strane, sve više kompanija, organizacija i institucija ulaže u razvoj i inkorporaciju velikih podataka u svoj rad kako bi poboljšali vlastito poslovanje i usluge, stoga knjižnice ne bi trebale odudarati od tog trenda. Iako postoje mnogi izazovi za učinkovito upravljanje i zaštitu podataka koji su slični ili različiti od upravljanja dokumentima i arhivima kao što su knjige i časopisi, nužno je da knjižnice budu u toku sa suvremenim alatima i tehnologijama te razvijaju svoj rad i znanstveno područje sukladno s njima.

Zaključak

Knjižničarstvo je područje djelatnosti koje je poznato već tisućljećima, budući da otkad je pisanog traga postoji i potreba da se ti zapisi pohrane i sačuvaju za buduću upotrebu. Naše je čitavo društvo izgrađeno na znanju koje ne bi bilo moguće prenositi bez knjiga i drugih dokumenata koji nam ga generacijama otkrivaju tako precizno i u toliko velikim količinama, a knjižnice su svakako najvažnije institucije za rad s podacima i informacijama zapisanim u tim materijalima širokog spektra oblika.

Tradicionalno su knjižničari doživljavani kao profesionalni djelatnici zaduženi samo za čuvanje zbirki knjiga i omogućavanje pristupa podacima pohranjenim u tim zbirkama, no s vremenom su se njihov obujam rada, ali i metode koje pri njemu koriste, znatno izmijenili i razvili. Tijekom posljednjih nekoliko desetljeća, posebice od kraja 20. stoljeća, odnosno od početka 21. stoljeća, elektronička tehnologija se razvija tempom i načinima koji velikom brzinom i silinom utječu na sve sfere života, privatnog i poslovnog, pa samim time i na rad knjižničara. Jedna od najznačajnijih posljedica ubrzanog razvoja tehnologije je i porast količine podataka koji se elektronički generiraju od strane znanstvenika, organizacija, ali i korisnika mrežnih usluga, te koji su doveli do nastanka novog pojma velikih podataka, podataka koji svojim volumenom, brzinom nastajanja i raznovrsnošću, zahtijevaju potpuno nov, suvremen pristup i alate obrade na masovnijoj razini kako bi se iz tih podataka izvuklo neko novo znanje, korist ili pretpostavka u obliku podatkovnog proizvoda.

Kao što je u radu navedeno, podaci u knjižnicama mogu dolaziti iz dva izvora, iz samih materijala pohranjenih u zbirci, ali i iz aktivnosti korisnika i upravljanja knjižnicom, dakle mogu biti kataložni ili procesni/transakcijski. Budući da ove obje vrste podataka imaju velik volumen, relativno veliku brzinu nastajanja te su izrazito raznovrsni, od strukturiranih do nestrukturiranih podataka, podatke koji nastaju u knjižnicama te koji se mogu u njima i koristiti, svakako možemo svrstati u kategoriju velikih podataka.

U radu s ovim podacima dodatna je prednost i činjenica da knjižničarstvo pripada području informacijskih znanosti koje se uvelike podudara s radom podatkovnih znanstvenika, jednim od najvažnijih stručnjaka u radu s velikim podacima. Djelatnost knjižničara obuhvaća mnogobrojne istovjetne aktivnosti i usluge koje su od velike koristi te nude mogućnosti u radu s velikim podacima. Neke od njih su indeksiranje sa specijalnim

vokabularom, sažimanje i klasifikacija, ali i sposobnost knjižničara da kvalitetno interpretiraju rezultate, odnosno razumijevaju sadržaj i uviđaju obrasce u podacima. Kako bi se navedeno učinkovito izvršilo, korisna je i kompetentnost knjižničara da dobro procijene vrijednost i vjerodostojnost podataka, te njihovu relevantnost, ali i ponude preporuke korisnicima. No, rad s velikim podacima u knjižnicama uključuje i nekoliko problema ili izazova, poput pitanja privatnosti, financijskih troškova, tehničkih izazova, nedovoljne upućenosti knjižničara u rad s tehnologijom što zahtijeva dodatnu edukaciju, te nedostatak podatkovnih znanstvenika i neprimjenjivost velikih podataka u svim institucijama.

Zaključno, smatram da su mogućnosti knjižničara za rad s velikim podacima, ali i primjena alata velikih podataka u knjižnicama od velike koristi. Kompatibilnost rada knjižničara s velikim podacima vidljiva je i u činjenici da se u posljednje vrijeme razvija i djelatnost podatkovnih knjižničara kao specijalne grane knjižničarstva, koji rade s digitalnim podacima na vrlo sličan način kao i s velikim podacima, ali i razvoj digitalnih knjižnica i globalnih platformi za razmjenu znanja. Uz konkretniju specijalizaciju i educiranje za rad s novijim elektroničkim alatima, knjižničari mogu u budućnosti postati važna stavka u ovom području, te svoju djelatnost proširiti i izvan tradicionalnog prostora knjižnice. Također, primjenom tehnologija velikih podataka u knjižnicama, knjižničari bi u budućnosti mogli znatno poboljšati svoje poslovanje i usluge, te korisnicima ponuditi kvalitetnije i personalizirane preporuke za daljnje čitanje ili radove slične tematike, kao što, na primjer, nudi Amazon, što bi svakako bilo veoma korisno, te povećalo zadovoljstvo korisnika i otvorilo vrata za daljnji tehnološki napredak, kako knjižničarstva, tako i cjelokupnog društva.

Drago mi je da sam se u završnom radu bavila temom velikih podataka i knjižnica, budući da ovo područje ima velik potencijal te su mi se otvorili neki novi vidici i ideje za budući rad. O velikim podacima nisam mnogo znala prije početka pisanja rada, no sada sam se bliže upoznala sa ovim vrlo aktualnim pitanjem, koje zaista predstavlja novu eru funkcioniranja našeg društva. U budućnosti bih voljela imati priliku za praktičan rad s velikim podacima jer bih time bila bliže stvarnom razumijevanju ovog kompleksnog fenomena i njegove praktične primjene.

Literatura

1. Affelt, A. (2015). *The Accidental Data Scientist*. Medford, New Jersey: Information Today, Inc.

2. Big Data solutions: Example of the development cost. (25. travnja 2018.) Preuzeto s: https://existek.com/blog/big-data-solutions-development-cost-example-software/ (26.8.2018.)

3. Chowdhury, G. G., Burton, P. F., McMenemy, D. I Poulter, A. (2008). *Librarianship: An Introduction.* London: Facet Publishing.

4. Cooke, L. (2018). Privacy, Libraries and the Era of Big Data. U *IFLA Journal*, *44* (3), 167-169.

5. Culik, N. i Döpke, C. (2018). About Forgetting and Being Forgotten. U Hoeren, T. i Kolany-Reiser, B. (ur.), *Big Data in Context:Legal, Social and Technological Insights.* Cham: Springer Open.

6. Gojčeta, A. (2013, 12. lipnja). Big Data zaista veliki ili samo uobičajeni podaci? *Poslovni.hr*. Preuzeto s: http://www.poslovni.hr/komentari/big-data-zaista-veliki-ili-samo-uobicajeni-podaci-244123 (26.8.2018.)

7. Han, J. i Kamer, M. (2000). *Data Mining: Concepts and Techniques*. Simon Fraser University.

8. Information Science (2018). U *Merriam Webster Dictionary*. Preuzeto s: https://www.merriam-webster.com/dictionary/information%20science?utm_campaign=sd&utm_medium=serp&utm_source=jsonld (26.8.2018.)

9. Information specialists and Data librarian skills. (n. d.) Preuzeto s: http://www.ands.org.au/working-with-data/data-management/overview/data-management-skills/information-specialists-and-data-librarian-skills (26.8.2018.)

10. Inside Big Data Editorial Team. (2018, 16. veljače). Key members of every Big Data team. *Inside Big Data*. Preuzeto s: https://insidebigdata.com/2018/02/16/7-key-members-every-big-data-team/ (26.8.2018.)

11. Klapwijk, W. (2016). *The Library (Big Data) Scientist*. IFLA/ALA Webinar "Big Data: New Roles and Opportunities for New Librarians". Preuzeto s: https://npsig.files.wordpress.com/2016/04/bd-sig-wouter-klapwijk.pdf [26.8.2018.]

12. Kocijan, K. (2014). Big Data: kako smo došli do Velikih podataka i kamo nas oni vode. U Komunikacijski obrasci i informacijska znanost. Zagreb: Zavod za informacijske studije.

13. Liu, S. i Shen, X. L. (2018). Library Management and Innovation in the Big Data Era. U *Library Hi Tech, 36* (3), 374-377.

14. Mayer-Schönberger, V. i Cukier, K. (2013). *Big Data. A revolution that will transform how we live, work and think.* Houghton Mifflin Harcourt: Boston.

15. O'Reilly Media (2011). *Big Data now – current perspectives from O'Reilly media.* Beijing-Cambridge-Farnham-Köln-Sebastopol-Tokyo: O'Reilly.

16. Pavlić, M. (2018, 23. siječnja). Stiže microSD kartica od 512 GB. *Bug*. Preuzeto s: https://www.bug.hr/dogadjaji/stize-microsd-kartica-od-512-gb-2517 (26.8.2018.)

17. Pejić Bach, M. (2005). Rudarenje podataka u bankarstvu. U *Zbornik Ekonomskog fakulteta u Zagrebu, 3* (1), 181-193.

18. Rouse, M. (2017, kolovoz). Business Intelligence. *Search Business Analytics*. Preuzeto s: https://searchbusinessanalytics.techtarget.com/definition/business-intelligence-BI (26.8.2018.)

19. Sawant, N. i Shah, H. (2013). *Big data application architecture Q & A: a problem-solution approach.* Apress.

20. v. Schönfeld, M., Heil, R. i Bittner, L. (2018). Big data on a Farm – Smart Farming. U Hoeren, T. i Kolany-Reiser, B. (ur.), *Big Data in Context:Legal, Social and Technological Insights.* Cham: Springer Open.

21. Wang, C., Chen, L., Xu, S. i Chen, X. (2016). Exposing Library Data with Big Data Technology: A Review. U *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*.

22. Wittmann, R. J. i Reinhalter, L. (2014). The Library: Big Data's Boomtown. *The Serials Librarian: From the Printed Page to the Digital Age*, *67* (4), 363-372. Preuzeto sa: https://tigerprints.clemson.edu/cgi/viewcontent.cgi?article=1089&context=lib_pubs [10.8.2018.]

23. Zhan, M. i Wilden, G. (2017). Public Libraries: Roles in Big Data. U *The Electronic Library*, *36* (1), 133-145.

24. Zhan, M. i Wilden, G. (2017). Understanding Big Data in Librarianship. U *Journal of Librarianship and Information Science*, *00* (0), 1-16.