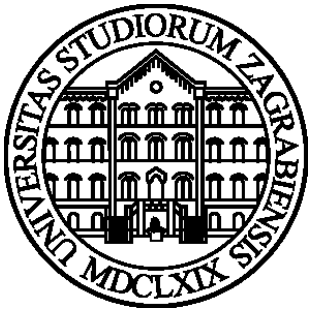
SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET

Krešimir Zauder

RAZVOJ SCIENOTOMETRIJE PRAĆEN KROZ ČASOPIS *SCIENTOMETRICS* OD POČETKA IZLAŽENJA 1978. DO 2010. GODINE

DOKTORSKI RAD

Zagreb, 2014.



UNIVERSITY OF ZAGREB

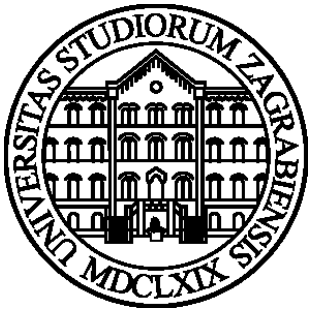
FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Krešimir Zauder

THE DEVELOPMENT OF SCIENTOMETRICS AS REPRESENTED BY THE JOURNAL *SCIENTOMETRICS* SINCE THE BEGINING OF ITS PUBLICATION IN 1978 TO 2010

DOKTORSKI RAD

Zagreb, 2014.



SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

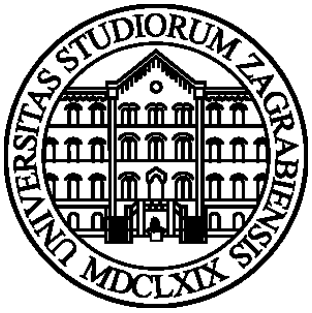
Krešimir Zauder

RAZVOJ SCIENOTOMETRIJE PRAĆEN KROZ ČASOPIS *SCIENTOMETRICS* OD POČETKA IZLAŽENJA 1978. DO 2010. GODINE

DOKTORSKI RAD

Mentor: dr. sc. Maja Jokić

Zagreb, 2014.



UNIVERSITY OF ZAGREB

FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Krešimir Zauder

THE DEVELOPMENT OF SCIENTOMETRICS AS REPRESENTED BY THE JOURNAL *SCIENTOMETRICS* SINCE THE BEGINING OF ITS PUBLICATION IN 1978 TO 2010

DOCTORAL THESIS

Supervisor: dr. sc. Maja Jokić

Zagreb, 2014.

SAŽETAK

Ciljevi ovog istraživanja su: 1. kvantitativno opisati glavno tijelo scientometrijske literature s posebnim fokusom na razvoj kroz vrijeme i 2. versatilno implementirati glavne scientometrijske postupke te opisati metodološku problematiku scientometrije.

Prvi doprinos ovog rada je dakle u razumijevanju i uopće validaciji scientometrije kao zasebne pod-discipline informacijskih znanosti kroz primarno kvantitativnu obradu svih radova objavljenih u časopisu *Scientometrics* od početka objavljivanja 1978. do 2010. godine s posebnim naglaskom na članke. Među provedenim analizama posebno se mogu istaknuti analize autorstva i suradnje, citatne i ko-citatne analize te pregled u odnosu na glavnu tematiku radova.

Drugi doprinos je u prikazu, implementaciji i povezivanju tradicionalnih metodoloških scientometrijskih postupaka s analizom mreža i tekstova radova. S obzirom na kompleksnost ulaznih podataka koji u većini slučajeva nisu stvoreni pod kontrolom istraživača, u sklopu prikaza metodologije posebno je detaljno prikazana i priprema podataka.

Rad pruža kratak uvod o znanosti i istraživanjima znanosti s fokusom na proučavanje znanosti kroz znanstvenu literaturu. Nakon toga slijedi poglavlje o metodologiji podijeljeno u dva dijela: priprema i analiza. Radi promatranja ovog tijela literature s više aspekata, rezultati i rasprava su prikazani zajedno i slijede nakon metodologije. Rad završava kratkim zaključkom o razvoju i relevantnosti scientometrije kao i o sve većoj zainteresiranosti za istu. U prilogu 1. prenesen je Python kôd kojim su implementirani postupci opisani u metodologiji, kao i proizvedene tablice i većina slika koji se koriste u tekstu.

Ključne riječi: scientometrija, scientometrijska metodologija, razvoj discipline, analiza suradnje, citatne analize, analiza mreža

SUMMARY

The goals of this research are: 1. to quantitatively describe the main corpus of scientometric literature with emphasis on its development through time, and 2. to implement main scientometric procedures in a versatile manner and to describe main scientometric methodological problems.

One contribution of the research is in understanding and validation of scientometrics as a separate sub-discipline of information sciences through mainly quantitative treatment of all papers published in journal *Scientometrics* since the beginning of its publication in 1978 to 2010 with special emphasis on articles. Among the used analytical procedures, authorship and cooperation analysis, citation and co-citation analysis as well as overview according to main paper focus are the most prominent.

The other contribution of the research is in description, implementation and merging of traditional scientometric procedures with network analysis and full texts of included articles. Given the complexity of input data which is not created under the control of the researcher, special attention has also been given to the process of data preparation.

The thesis gives a short introduction about science and scientific inquiry into science through scientific literature. Methodology is divided in two parts, preparation and analysis, and is detailed as appropriate for the stated goals of the research. Due to the multi-aspect nature of the research, results and discussion are kept together and are shown after the methodology. The thesis finishes with the conclusion which gives a data overview about the development and relevance of scientometrics as well as about growing interest for it. Appendix 1. gives the python code which implements procedures described in methodology and which produces all the analytical results (i.e. tables and images) used in the thesis.

Keywords: scientometrics, scientometric methods, field, development, cooperation analysis, citation analysis, network analysis

SADRŽAJ

1 Uvod 1

1.1 O znanosti i istraživanjima znanosti 1

1.1.1 O istraživanjima znanstvene literature 6

1.1.2 O scientometriji 9

1.1.3 O časopisu *Scientometrics* 27

1.2 O ovom istraživanju 28

2 Metodologija 30

2.1 Terminologija 31

2.2 Preuzimanje i priprema bibliografskih metapodataka 33

2.2.1 Izvori, ulazni skupovi podataka i njihovo preuzimanje 36

2.2.2 Osnovna priprema ulaznih podataka 39

2.2.3 Detaljna priprema podataka o radovima iz *Scientometricsa* 45

2.2.4 Priprema citirajućih radova i citata na radove u časopisu *Scientometrics* 60

2.3 Priprema tekstova radova za analizu 62

2.3.1 Ekstrakcija iz PDF formata 63

2.3.2 Priprema tekstova radova 67

2.4 Struktura istraživanja i korišteni alati 70

2.5 Analiza 71

2.5.1 Prebrojavanje 73

2.5.2 Grupiranje 76

2.5.3 Supojavnost i mapiranje znanosti 78

2.5.4 Analiza mreža 80

2.5.5 Bibliometrijski pregled radova u *Scientometrics* 86

2.5.6 Autorstvo radova, produktivnost autora i suradnja među autorima 86

2.5.7 Citatne i ko-citatne analize 93

2.5.8 Pokazatelji dobiveni iz tekstova radova 100

3 Rezultati i rasprava 103

3.1 Radovi u časopisu *Scientometrics* 1978-2012 103

3.2 Autorstvo, autori i suradnja u časopisu *Scientometrics* 1978-2010 109

3.2.1 Autorstvo članaka u časopisu *Scientometrics* 110

3.2.2 Produktivnost autora članaka u časopisu *Scientometrics* 115

3.2.3 Suradnja među znanstvenicima na člancima u časopisu *Scientometrics* 118

3.2.4 Zemlje ustanova autora i međunarodna suradnja 131

3.3 Citatne analize članaka u časopisu *Scientometrics* 134

3.3.1 Citiranost članaka u časopisu *Scientometrics* 134

3.3.2 Najcitiraniji članci objavljeni u časopisu *Scientometrics* 139

3.3.3 Visoko produktivni i citirani autori u časopisu *Scientometrics* 145

3.3.4 Radovi i časopisi koji su citirali časopis *Scientometrics* 148

3.3.5 Radovi i časopisi citirani u časopisu *Scientometrics* 153

3.3.6 Citirani časopisi u člancima u časopisu *Scientometrics* 154

3.3.7 Starost citata prema i iz časopisa *Scientometrics* 157

3.3.8 Samocitati 159

4 Zaključak 161

5 Literatura 168

6 Dodatak 1. Python kôd 176

POPIS TABLICA

Tablica 1. Prazne vrijednosti preuzetih atributa u ulaznim skupovima podataka 43

Tablica 2. Nepravilnosti u ulaznim skupovima podataka 47

Tablica 3. Pregled objavljivanja svih radova u časopisu *Scientometrics* 1978-2012..... 104

Tablica 4. Broj radova prema vrsti i godini objave 105

Tablica 5. Autorstvo članaka objavljenih u časopisu *Scientometrics* 1978-2010 u odnosu na razdoblje 112

Tablica 6. Autorstvo članaka objavljenih u časopisu *Scientometrics* 1978-2010 u odnosu na tematiku 115

Tablica 7. Pregled produktivnosti autora u časopisu *Scientometrics* 1978-2010..... 116

Tablica 8. Produktivnost autora s obzirom na broj članaka objavljenih u časopisu *Scientometrics* 1978-2010 116

Tablica 9. Pregled mreže koautorstva na člancima objavljenim u časopisu *Scientometrics* 1978-2010..... 120

Tablica 10. Značajke mreže koautorstva na člancima objavljenim u časopisu ....*Scientometrics* 1978-2010..... 122

Tablica 11. Suradnja na člancima objavljenim u časopisu *Scientometrics* 1978-2010..... 123

Tablica 12. Značajke glavne komponente mreže koautorstva na člancima objavljenim u časopisu *Scientometrics* 1978-2010 125

Tablica 13. Promjene u broju autora po razdoblju 131

Tablica 14. Pregled zemalja koje su zastupljene u časopisu *Scientometrics* s više od 19 članaka..... 132

Tablica 15. Pregled citata svih radova objavljenih u časopisu *Scientometrics* 1978-2010 .. koje su dobili iz časopisa indeksiranih u WoS-u 1978-2012 135

Tablica 16. Pregled citata prema vrstama radova i priloga u časopisu *Scienotmetrics*..... 136

Tablica 17. Pregled citata članaka objavljenih u časopisu *Scientometrics* 1978-2010 i citiranih 1978-2012 u odnosu na razdoblja 137

Tablica 18. Pregled citata članaka objavljenih u časopisu *Scientometrics* 1978-2010 i citiranih 1978-2012 u odnosu na njihovu tematiku 138

Tablica 19. Pregled autora koji su u časopisu *Scientometrics* 1978-2010 objavili više od ..... 19 članaka 146

Tablica 20. Pregled radova objavljenih u časopisima u WoS citatnim indeksima 1978-2012 koji su citirali članke objavljene u časopisu *Scientometrics* 1978-2010 149

Tablica 21. Top 20 časopisa indeksiranih u WoS-u 1978-2012 koji su objavljivali radove koji citiraju časopis *Scientometrics* 150

Tablica 22. Pregled podataka o citiranim publikacijama u člancima objavljenim u časopisu *Scientometrics* 1978-2010. u odnosu na razdoblja 153

Tablica 23. Pregled podataka o citiranim publikacijama u člancima objavljenim u ..... časopisu *Scientometrics* 1978-2010. u odnosu na razdoblja 154

Tablica 24. Top 20 najčešće citiranih časopisa u člancima objavljenim u časopisu *Scientometrics* 155

Tablica 25. Starost citata članaka u časopisu Scientometrics u odnosu na razdoblja 158

Tablica 26. Starost citata članaka u časopisu Scientometrics u odnosu na tematiku 159

Tablica 27. Medijan starosti citiranih publikaciju u časopisu *Scientometrics* prema godinama objave 159

Tablica 28. Samocitati autora i časopisa 160

Tablica 29. Pregled pokazatelja o člancima objavljenim u časopisu *Scientometrics* 1978- 2010 kao indikatora razvoja scientometrije 162

Tablica 30. Pregled pokazatelja o autorima članaka objavljenih u časopisu *Scientometrics* 1978-2010 kao indikatora razvoja scientometrije 165

POPIS SLIKA

Slika 1. Scientometrija, bibliometrija, informetrija ... 8

Slika 2. Pojednostavljen prikaz citatnog indeksa ... 12

Slika 3. Citirajući rad, citat i citirani rad ... 33

Slika 4. Preuzimanje i priprema ulaznih podataka za analizu ... 35

Slika 5. Preliminarna priprema svih skupova podataka ... 40

Slika 6. Provjera vrijednosti za različite skupove podataka o radovima iz časopisa *Scientometrics* ... 46

Slika 7. Spajanje zapisa o radovima iz časopisa *Scientometrics* ... 49

Slika 8. Ujednačavanje imena autora ... 52

Slika 9. Prikaz sučelja razvijenog za efikasnu klasifikaciju radova i priloga u časopisu *Scientometrics* ... 57

Slika 10. Stilovi navođenja literature korišteni u člancima u časopisu *Scientometrics* ... 65

Slika 11. Primjer slučajno generiranog neusmjerenog grafa ... 81

Slika 12. Primjer slučajno generiranog usmjerenog grafa ... 82

Slika 13. Broj svih radova i članaka po godinama objave ... 104

Slika 14. Udijeli radova po tematici za sva vremenska razdoblja ... 107

Slika 15. Tematika članaka po godinama objave ... 108

Slika 16. Broj članaka i autora članaka po godinama objave ... 110

Slika 17. Distribucija radova po broju autora ... 112

Slika 18. Proporcija višeautorskih članaka po godinama objave ... 114

Slika 19. Mreža suradnje autora na člancima u časopisu *Scientometrics* 1978-2010 ... 119

Slika 20. Glavna komponenta u mreži suradnje autora na člancima u časopisu *Scientometrics* 1978-2010 ... 124

Slika 21. Distribucija stupnja centralnosti po autoru ... 126

Slika 22. Mreža suradnje autora na člancima u časopisu *Scientometrics* 1978-1988 ... 127

Slika 23. Mreža suradnje autora na člancima u časopisu *Scientometrics* 1989-1999 ... 128

Slika 24. Mreža suradnje autora na člancima u časopisu *Scientometrics* 2000-2010 129

Slika 25. Mreža međunarodne suradnje na radovima u časopisu *Scientometrics* 133

Slika 26. Distribucija citata po citiranim člancima i visiko citirani radovi 140

Slika 27. Udjeli teorijskih, metodoloških i primijenjenih visokocitiranih članaka 141

Slika 28. Mreža kocitata visoko citiranih članaka objavljenih u časopisu *Scientometrics* koji su kocitirani barem 10 puta 143

Slika 29. Mreža suradnje autora koji su objavili više od 20 članaka u časopisu *Scientometrics* 147

Slika 30. Distribucija citata na članke u časopisu *Scientometrics* po citirajućim časopisima i visoko citirajući časopisi 152

1 UVOD

1.1 O znanosti i istraživanjima znanosti

O temama usko vezanim za znanost raspravlja se od antike. Često navođena "prva definicija znanosti" je Aristotelova: "Pouzdano znanje koje se može logički i racionalno objasniti". Međutim, u ovom smislu u kojem se riječ katkad i danas koristi, "znanost" se tretira kao vrsta znanja što je i originalno značenje latinske riječi *scientia*, a Aristotel ne naglašava razliku između grčkih pojmova *philosophia* i *episteme* tj. znanstvenog znanja (Waugh i Ariew, 2008).

Iako su problemi znanja temeljni u filozofiji od početaka, povijesno gledano, o proučavanju "znanosti" možemo pričati tek nakon formalizacije pojma. Današnja znanost se često smatra proizvodom tzv. znanstvene revolucije. Točan period u kojem se znanstvena revolucija odvijala varira među povjesničarima, ali kulminacija se svakako smješta u 17. stoljeće, konsolidacija u 18., a formalizacija pojma znanost u današnjem smislu tek u 19. stoljeće (Henry, 2008). Drugim riječima, moderna upotreba termina "znanost" je morala pričekati još dva stoljeća nakon perioda koji se smatra ključnim u znanstvenoj revoluciji da se razvije iz termina "prirodna filozofija" i "prirodna povijest" (Godfrey-Smith, 2003). Filozofija znanosti kao prepoznatljiva pod-disciplina filozofije tako nastaje tek u 20. stoljeću (Sarkar i Pfeifer, 2006). Usko gledano, povijest znanosti od antičke Grčke do relativno nedavno je uvelike povijest odvajanja znanstvenih disciplina od filozofije. (Rosenberg, 2005).

Ipak, isprepleteni pojmovi poput "filozofija", "filozofija znanosti", "povijest filozofije" i "povijest znanosti" trebaju se tretirati različito prema fokusu i vlastitom razvoju teorija i metoda (Waugh i Ariew, 2008) iako u različitim kontekstima promatraju sličnu ili istu problematiku. Za povijest znanosti može se reći da je institucionalizirana početkom 20. stoljeća, sociologija znanosti sredinom 20. stoljeća, a u posljednje vrijeme govori se i o psihologiji znanosti koja je prema nekim autorima zrela za izdavanje prvih časopisa (Feist, 2006). Ovakve discipline često se nazivaju "metaznanostima", a multidisciplinarni znanstveni pristup istraživanju znanosti se u novije vrijeme naziva i "znanost o znanosti".

Kao početak filozofije znanosti kao zasebne discipline s distinktivnom metodologijom često se spominje logički pozitivizam (tj. logički empiricizam ili neo-pozitivizam) (Uebel, 2008) s kojim često i počinje uvod u problematiku znanosti 20-og stoljeća. Centralne ideje logičkog pozitivizma su analitičko-sintetička distinkcija i verifikacijska teorija značenja (Godfrey-

Smith, 2003). Pojednostavljeno rečeno, analitičko-sintetička distinkcija je razlika između analitičkih tvrdnji koje su istinite ili ne prema značenju inherentnom njima od sintetičkih tvrdnji koje zahtijevaju dodatno znanje o svijetu za procjenu istinitosti. Verifikacijska teorija značenja je aplikativna na sintetičke tvrdnje i bazira se na ideji da značenje tvrdnje potiče iz znanja kako verificirati tvrdnju. Uz navedeno logički pozitivizam ima svojstven pristup empiricizmu koji se temelji na jasnom odvajanju npr. matematike kao striktno analitičkog izvora znanja.

Na temelju ovih ideja logički pozitivisti su predlagali ne samo teoriju znanosti već i okvir u kojima bi se znanost i filozofija trebali odvijati kako bi bili svrsishodni tj. predlagali su općenitu teoriju jezika, znanja i značenja. Logički pozitivizam i zatim kritika istoga, čest su početak uvoda u filozofiju znanosti iako je sam logički pozitivizam interpretabilniji kao filozofski pokret više nego skup doktrina (Creath, 2013). Kasnija analiza velikog broja doktrina koji su različiti članovi pokreta zastupali, nedavno vraća nešto poštovanja prema pokretu nakon perioda potpunog odbacivanja (Uebel, 2008). Ipak, logički pozitivizam dobar je početak za teorije koje su uslijedile, ako ništa drugo onda radi toga što je u mnogima prisutna reakcija upravo na ovaj holistički pristup propisivanju određene vrste racionalnog ponašanja.

Sredina dvadesetog stoljeća, u ovom kontekstu, teško može proći bez spomena Karla Poppera. Ključan problem za Poppera je problem demarkacije, problem razdvajanja znanosti od ne-znanosti, a njegov odgovor na problem se temelji na konceptu oborivosti (eng. *falsifiability*) tj. mogućnosti dokazivanja netočnosti. Oborivost je, sama po sebi, relativno jednostavna ideja da je hipoteza znanstvena samo ako se može odbaciti na temelju neke moguće opservacije. Ova ideja obilježila je razlikovanje znanstvenog od ne-znanstvenog pristupa postavljanju tvrdnji. Korištena je, na primjer, i na Američkom sudu pri odluci o tome da li se kreacionizam može smatrati znanošću i kao takav biti prisutan u edukaciji (Bird, 1998). Popper kasnije, kao i gotovo sve ostalo u istraživanjima znanosti u 20-om stoljeću, doživljava kritiku. Najčešća kritika je vjerojatno njegov preskriptivan pristup "kakva bi znanost trebala biti" bez previše upita u kako se znanost u praksi odvija, a i sam je u svojoj autobiografiji zapisao (prema Feist, 2006) da je u ponovnom čitanju Kantove *Kritike* kao centralnu ideju interpretirao: "Znanstvene teorije su ljudsko djelo, a mi ih pokušavamo nametnuti svijetu".

U ovom smislu se vremenom razvija i posljednji prepoznatljivi pokret u filozofiji znanosti, prirodna epistemologija, među čijim osnovnim postavkama je da se filozofija znanosti mora više bazirati na onome što znanstvenici rade nego na onome što bi trebali raditi (Feist, 2006).

Par godina nakon objavljivanja Popperovog djela *The logic of scientific discovery* na engleskom jeziku 1959. godine, pojavljuje se još jedno djelo koje je obilježilo današnje razumijevanje znanosti. Riječ je o *The structure of scientific revolutions* Thomasa Kuhna. Centralni koncepti koje je Kuhn unio su jamačno "normalna znanost" i "znanstvene revolucije" uz bogatu upotrebu riječi "paradigma". Normalna znanost jest onaj period znanosti u kojoj postoji standardno prihvaćen teoretski okvir i metodologija unutar kojih znanstvenici djeluju i razmišljaju. Ovakav okvir Kuhn naziva paradigmom. Nova otkrića i razmišljanja unutar i izvan neke paradigme s vremenom će u nju unijeti nelogičnosti tj. istinite tvrdnje koje se kose s postavkama te paradigme. Akumulacija ovakvih "iznimaka" otvara put novoj paradigmi koja zamijenjuje prijašnju. Promjena paradigme je "znanstvena revolucija" koja mijenja osnovne postavke znanosti u kojoj se dešava.

Veliki doprinos Kuhna, pored samih teorija koje kasnije doživljavaju kritiku, je u razbijanju filozofskih mitova. Kuhn je pokazao da praksa znanstvenih aktivnosti nema puno zajedničkog s tradicionalnim teorijama znanja i racionalnosti pri čemu je poslužio i kao podloga protiv logičkog pozitivizma. Među kritičarima Kuhna koji su koristili slične koncepte nalazimo Imrea Lakatosa s "istraživačkim programima" i Larrya Laudana s "istraživačkim tradicijama" što su koncepti slični Kuhnovim paradigmama, ali različiti u internoj organizaciji i međusobnoj interakciji. Njihova glavna zamjerka Kuhnu je iracionalnost prikazanih procesa u znanosti (Godfrey-Smith, 2003).

Četrdesetih godina 20. stoljeća u istraživanja znanosti se uključuje i sociologija, primarno s Mertonom i njegovim radovima kasnije zbirno objavljenim u *Sociology of science*. Mertonova sociologija znanosti primarno je ekstenzija standardnih sociooloških koncepata i metoda na znanstvene aktivnosti (Godfrey-Smith, 2003). Sam Mertonov značaj je radi širine teško sažeti, a proteže se i do same scientometrije. U tom kontekstu Fox (2004) tvrdi kako je Merton etablirao istraživanje znanosti kao društvene institucije u kojoj su pravila, sustavi procjene i procesi nagrađivanja društveno locirani i determinirani. Među njegovim doprinosima relevantnim za scientometriju nalazimo istraživanje stratifikacije u znanosti kao i razumijevanje fenomena kumulativne prednosti znanstvenika što je poznato i kao Matejev

efekt. Uz to, Merton se često spominje i uz teoriju citatnih analiza, što je opisano kasnije kao zasebna problematika.

U drugoj polovici 20. stoljeća razvija se sociologija znanosti dramatično drugačija od Mertonove (Feist, 2006), prema kojoj znanost i znanstvena djelatnost (u širem smislu) nisu posebne kao takve već njihove specifične značajke proizlaze iz specifičnog uređenja znanstvenih zajednica. Zanimljiva posljedica ovakvog viđenja znanosti je što uloga "stvarnog svijeta" i promatranja istog nema toliko važnu ulogu u objašnjenju znanstvenih aktivnosti. Spojeno s Wittgensteinovom idejom "igre jezika" proces kulminira u ideji "konstrukcije stvarnosti" za razliku od "proučavanja stvarnosti" (Godfrey-Smith, 2003).

Ukratko rečeno, dok o problemima znanosti možemo govoriti od antičkog svijeta, oznanosti u današnjem smislu možemo govoriti tek u posljednja dva stoljeća. Koncept i problemi znanosti od tada proživljavaju daljnje važne promjene u kojima je posebno turbulentna filozofija znanosti 20-og stoljeća. Ove "važnepromjene", uključuju promjene shvaćanja i opsega koncepta. Na primjer, početkom 20-og stoljeća status ekonomije i psihologije kao znanosti bio je upitan (Godfrey-Smith, 2003). Također, razvoj razmišljanja o znanosti stavlja upitnike i u ideje da je znanost izvor pouzdanog znanja o stvarnom svijetu koja se koristi znanstvenom metodom i koja je nastala u današnjem smislu u znanstvenoj revoluciji. U drugoj polovici 20. stoljeća, brane se i tvrdnje da znanost konstruira stvarnost kakvu poznajemo, ne samo da znanstvena metoda nije nimalo precizan pojam (Godfrey-Smith, 2003) već da takve normativne koncepte treba izbjegavati (Feyerabend, 1975) te da znanstvena revolucija koje je današnja znanost proizvod nije postojala na način na koji ju se često prikazuje i zamišlja (Shapin, 1996).

U svakom slučaju, prema Godfrey-Smithu (2003), možemo razlikovati tri pristupa koja su se koristila za objašnjavanje znanosti: empirijski, matematički i društveno-organizacijski. Empirijski, znanost ima podlogu u objektivnom promatranju svijeta. U matematičkom smislu ono što čini znanost posebnom je korištenje matematičkih principa u svrhu razumijevanja svijeta. U društvenom smislu, ono što znanost čini posebnom jesu znanstvene zajednice s jedinstvenim odnosima validacije, suradnje i povjerenja.

Dok se znanost manifestira kao kombinacija ovih objašnjenja, za predmet ovog doktorata društveno objašnjenje je posebno zanimljivo. Shvaćena kao proces i kao kompleksna ljudska djelatnost, znanost se temelji na kreativnom i inovativnom radu znanstvenika i kontinuiranom stvaranju novih spoznaja što ju čini jednim od ključnih čimbenika razvoja ljudskog znanja općenito pa tako i civilizacije, kulture i tehnologije. Obzirom na njezine glavne ciljeve, znanost je djelatnost u konstantnim promjenama, ali s čvrstom jezgrom koju primarno tvori formalno empirijski pristup istraživanjima te komunikacija kroz znanstvene publikacije. Razni čimbenici, poput razvoja teorije, primjene novih istraživačkih metoda i instrumenta, otvorenosti znanstvenih zajednica i adekvatnosti infrastrukture, bitno utječu na brzinu razvoja pojedinih znanstvenih discipline i znanosti općenito.

Iz ove su perspektive znanstvena istraživanja znanosti posebno zanimljiva. Znanost se može promatrati kao istraživački proces u kojem znanstvenici koordiniraju svoje aktivnosti kako bi različitim metodama proizveli novo znanje. Kako bi komunicirali i očuvali novostečeno znanje, objavljuju ga u formi različitih publikacija koje predstavljaju srž znanstvene komunikacije. To su primarno članci u znanstvenim časopisima ili zbornicima skupova, knjige, patenti i razna izvješća (Whitley, 2000). U novije vrijeme možemo govoriti i o publikaciji podataka ili alata u obliku baza podataka i softverskih rješenja.

Znanstvenici komuniciraju i kodificiraju rezultate istraživačkog rada na prepoznatljiv način od pojave prvih znanstvenih časopisa u 17-om stoljeću što pak omogućuje sustavno promatranje znanstvene literature. Fenomen objavljivanja u međunarodnim znanstvenim časopisima je posebno značajan u ovom smislu i to pogotovo u kontekstu scientometrijskih istraživanja (Van Raan, 2005).

Znanstvena komunikacija kroz publikacije se temelji na povjerenju u pouzdanost objavljenih informacija. Pouzdanost objavljenih informacija postiže se recenzentskim postupkom. Povjerenje u pouzdanost objavljenog znanja kolega s jedne strane i zadaća propitivanja znanstvenih informacija s druge strane, tvori kompleksne odnose povjerenja i validacije u znanosti. Navedeni odnosi se u literaturi formalno manifestiraju u citatima, što je detaljnije opisano kasnije u tekstu.

Ako se vratimo do Aristotela i poimanja znanosti kao pouzdanog znanja u kombinaciji s društvenim objašnjenjem posebnih struktura povjerenja i validacije u znanosti, tada promatranjem znanstvenih publikacija kao i aktera i procesa koji su ih proizveli možemo operacionalizirati promatranje znanosti kao kontinuiranog procesa.

1.1.1 O istraživanjima znanstvene literature

Prebrojavanje publikacija prisutno je od samih početaka njihovog prikupljanja i organizacije. Na primjer, već u trećem stoljeću prije Krista nailazimo na primjere ukupnih brojeva svitaka u Aleksandrijskoj knjižnici (Broadus, 1987). Prema većini autora, prva istraživanja literature matematičkim i statističkim metodama, koje danas nazivamo bibliometrijskim istraživanjima, pojavljuju se na prijelazu između 19-og. i 20-og stoljeća, gdje različiti autori često prijavljuju različita "prva bibliometrijska istraživanja" (Lawani, 1981; Broadus, 1987; Shapiro, 1992; Osareh, 1996; Godin, 2006). U svakom slučaju, ako ostavimo po strani raspravu o tome koje se istraživanje može smatrati prvim primjerom, dostupni su nam primjeri iz "prapovijesti" bibliometrije i scientometrije tj. prije no što su etablirana kao prepoznata područja informacijskih znanosti i prije no što uopće postoje ti nazivi.

Ako krenemo kronološkim redom rane bibliometrije, Shapiro (1992) navodi često zanemarivanu činjenicu tradicije "bibliometrije" u području prava. Brojevi publikacija se nalaze u pravnim spisima od 1817. godine, a rane preteče citatnih indeksa postoje od 1743. Iako se ova upotreba bibliometrije ne odnosi na znanstvenu literaturu, kao što ćemo kasnije u tekstu vidjeti, ona ima važan doprinos u razvoju citatnih indeksa znanstvene literature koji omogućuju citatnu analizu kao vrlo važnu komponentu scientometrije.

Prema Osareh (1996) jedno od najstarijih bibliometrijskih istraživanja datira još iz 1890. godine i vezano je uz proučavanje raspršenja tema publikacija statističkim metodama. Prema Godin (2006) sustavna istraživanja provodio je James McKeen Cattell, dugogodišnji urednik časopisa *Science*. Od 1906. godine Cattel sustavno prikuplja i objavljuje statistike vezane za produktivnost znanstvenika i usporedbu zemalja. Među njegovim statistikama značajno mjesto imaju brojevi publikacija, ali i druge informacije, poput broja znanstvenika i informacija o ustanovama, što ga čini i pretečom scientometrije. Često navođen (Lawani, 1981; Shapiro, 1992; Osareh, 1996) primjer bibliometrijskog istraživanja znanstvene literature koji je popularizirao pristup je istraživanje koje su 1917. proveli Cole i Eales. Oni su primjenom statističkih analiza istraživali literaturu komparativne anatomije od 1550. do 1860.

Nešto kasnije Alfred James Lotka (1926) objavljuje vrlo utjecajan rad o produktivnosti istraživača u području kemije i fizike. Za rad kemičara koristio se i danas dostupan sekundarni izvor *Chemical abstracts*. Lotka pronalazi pravilnosti u produktivnosti autora tj. da je za većinu radova odgovorna manjina autora i postavlja važan zakon koji se naziva Lotkinim zakonom koji je pobliže opisan u poglavlju o produktivnosti kao vrlo važan koncept u scientometriji.

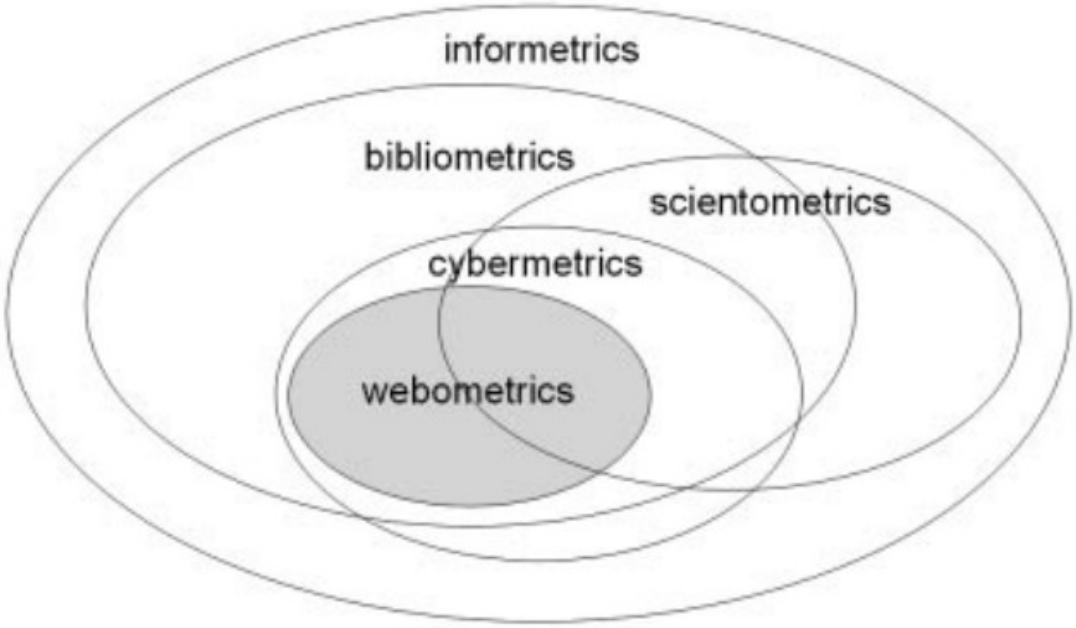
Prvi naziv za ovu vrstu istraživanja bio je "statistička bibliografija" koji se koristi od 1923. godine nakon istoimenih predavanja Edwarda Wyndhama Hulmea (Lawani, 1981). Pojam "bibliometrija" se koristi kao precizniji nakon preporuke Alana Pritcharda 1969. godine u članku "Statistička bibliografija ili bibliometrija?" (Pritchard, 1969). Pritchardov problem s terminom "statistička bibliografija" bio je što je teško reći da li se odnosi na statističku analizu bibliografija ili bibliografiju publikacija o statistici. Upotreba pojma "bibliometrija", pak, može se pronaći već 1934. kod Paula Otleta u obliku *bibliometrie* (Garfield, 2004). Ipak, brojni autori se slažu da je Pritchard originalni autor termina (Osareh, 1996).

Pritchard definira bibliometriju kao upotrebu matematičkih i statističkih metoda u svrhu proučavanja knjiga i ostalih komunikacijskih medija (Pritchard, 1969). Možda svrsishodnija, iako potencijalno preširoka, je Fairthornova definicija (prema Cronin, 1984) koji po uzoru na Pritcharda definira bibliometriju kao kvantitativno proučavanje svojstava zapisanog diskursa i vezanog ponašanja.

Temeljna ideja bibliometrije znanstvenih tekstova je korištenje elemenata koje pružaju bibliografski metapodaci za aproksimaciju važnih aspekata znanosti (Van Raan, 2005). Polja koja se koriste su, na primjer, imena autora, adrese ustanova, časopis (koji indicira područje i status), citirani radovi, ključne riječi i slično. Dok operacionalizacija proučavanja znanosti samo kroz njene publikacija nije savršena, publikacije se prihvaćaju kao osnovna jedinica znanosti i izvor podataka (Van Raan, 2005).

Kao što je opisano u sljedećem poglavlju, uz pojam bibliometrije pojavljuje se i pojam scientometrije. Kroz razvoj elektroničkih medija i informacijskih znanosti, ovakvim "metrijama" se dodaju i informetrija, kibermetrija i webometrija. Obzirom na fokus ovog rada od interesa su nam scientometrija i bibliometrija, ali u ilustrativne svrhe preklapanja ovih različitih pristupa tj. predmeta promatranja mogu se vizualizirati kao što je prikazano na slici 1, koja je preuzeta iz (Björneborn i Ingwersen, 2004).

Slika 1. Scientometrija, bibliometrija, informetrija ...



Zanimljivo rano djelo u scientometriji koje se nije baziralo na bibliometrijskim pokazateljima je "Histoire des sciences et des savants depuis deux siècles" Alphonsea de Candollea (1873, (prema Van Raan, 1997)). Knjiga opisuje promjene u znanstvenom uspjehu nacija koristeći se informacijama o članstvu u znanstvenim društvima te povezuje informacije s vanjskim faktorima različitih vrsta (uključujući, na primjer i ulogu celibata).

Ovaj kratak pregled ranih istraživanja literature i znanosti temelji se, u skladu s temom rada, na kvantitativnom pristupu. U slučaju kvalitativnog pristupa, u načelu se radi o fokusiranim sadržajnim analizama manjeg broja publikacija. Kvantitativan pristup, pak, primarno uzima u obzir bibliografske značajke znanstvenih radova poput podataka o autorstvu, časopisu, godini i citatima koje zatim obrađuje raznim kvantitativnim metodama mnoge od kojih su provedene u ovom istraživanju i detaljnije opisane u metodologiji.

U praksi se, naravno, kvantitativan i kvalitativan pristup mogu kombinirati. Na primjer kvantitativnim opisom većeg tijela literature i kvalitativnom obradom manjeg uzorka koji može biti odabran na temelju kvantitativnog istraživanja ili pak kvalitativnom validacijom kvantitativno dobivenih pokazatelja. Također, istraživanja literature se često nadopunjuju informacijama poteklim iz drugih izvora poput informacija o dobi i spolu autora ili pak informacijama dobivenim kroz upitnike, intervjue i slično. Kombinacija podataka dobivenih

kroz bibliometrijske analize s drugim relevantnim podacima u svrhu proučavanja znanosti, otvara prostor scientometrijskim istraživanjima.

1.1.2 O scientometriji

Sustavna mjerenja znanosti bazirana na znanstvenim publikacijama i razvoj kvantitativnih metoda u tu svrhu šezdesetih godina dvadesetoga stoljeća predstavljaju početak razvoja scientometrije kao discipline. Ranija sporadična istraživanja obično se tretiraju kao "prapovijest" scientometrije (Godin, 2006).

Od početaka ovu granu karakterizira interdisciplinaran pristup koji su razvijali znanstvenici iz područja prirodnih znanosti, filozofi, sociolozi i knjižničari. Kao i u razvoju drugih znanstvenih disciplina, sustavnom razvoju scientometrije prethodila su povremena istraživanja najčešće vezana uz karakteristike znanstvenog objavljivanja.

Kao začetnik scientometrije često se spominje Derek John de Solla Price i njegova knjiga dobro poznata i izvan granica informacijskih znanosti *Little science, Big science* (de Solla Price, 1963) u kojoj Price tvrdi:

> *"Science is a measurable substance, consequently, the manpower engaged in science, the scientific literature, talent and expenses afforded to science can be measured by properly selected statistical methods."*

U uredničkom uvodniku u prvi broj časopisa *Scientometrics* de Solla Price sumira "bili bismo loši znanstvenici kada ne bismo mogli koristiti naše profesionalne analitičke alate na vlastitim aktivnostima" (de Solla Price, 1978). Srž scientometrije, dakle, proizlazi iz promatranja znanosti kao mjerive supstance odnosno promatranja aktera te ulaza i izlaza znanstvenih procesa. O važnosti de Solle Pricea za scientometriju gotovo da je bespredmetno raspravljati. Nakon njegove smrti 1983. časopis *Scientometrics* počinje sponzorirati nagradu i medalju nazvanu po njemu (Bonitz, 1994). Od 1993. godine, nagrada se dodjeljuje svake dvije godine. Zaključno s 2013-om godinom dodijeljeno je 17 nagrada.

Pojam scientometrija pripisuje se istoimenoj knjizi na ruskom jeziku "Naukometrija" (Nalimov i Mulchenko, 1969) objavljenoj 1969. godine (Vinkler, 1994). Vassily Vassilievich Nalimov, ruski matematičar, svoji prvi rad koji se može interpretirati kao scientometrijski

objavljuje 1959. godine (Granovsky, 2001) i smatra se autorom s prepoznatljivim doprinosom informacijskim znanostima (Cherny i Gilyarevsky, 2001).

Što se definicija scientometrije tiče, Nalimov i Mulchenko, autori "Naukometrije", definiraju scientometrijska istraživanja kao ona koja promatraju znanost kao informacijski proces aplikacijom kvantitativnih (statističkih) metoda (Vinkler, 1994). Braun, Glänzel i Schubert (1985) definiraju scientometriju kao disciplinu koja analizira kvantitativne aspekte stvaranja, prijenosa i korištenja znanstvenih informacija u svrhu doprinosa razumijevanju mehanizama znanstvenih istraživanja kao društvenih aktivnosti.

Sam razvoj scientometrije vezan je uz razvoj izvora podataka koje koristi. S obzirom da je bibliometrija centralan pristup u scientometriji, jasno je da su kvalitetni izvori bibliografskih metapodataka od ključne važnosti. Uz to, važna tema u scientometriji je praćenje odjeka koji su publikacije tj. vezani akteri imali. U znanstvenoj literaturi ovaj odjek, a i ranije spomenuti odnosi povjerenja i validacije, formalno su prisutni kroz citate koji su važna značajka znanstvenih tekstova. U scientometriji, dakle, uz standardne bibliografske zapise o znanstvenim publikacija potrebna je i formalno zapisana informacija o mreži citata među tim publikacijama. Baza podataka koja uz bibliografske informacije sadrži i informacije o citatima naziva se citatni indeks što je među centralnim konceptima koji su omogućili status scientometrije kao prepoznate zasebno discipline.

Autor prvog citatnog indeksa radova objavljenih u znanstvenim časopisima je Eugene Garfield, koji se uz de Sollu Pricea često veže uz osnivače scientometrije. Garfield predlaže prvi citatni indeks znanosti 1953. godine po uzoru na *Shepardove citate* (Garfield, 1998), citatni indeks pravne dokumentacije čije je prve verzije Frank Shepard počeo izrađivati već 1873. godine (Shapiro, 1992).

1.1.2.1 Citatni indeksi

Razvoj Science Citation Indexa (SCI) 1964. godine među ključnim je trenucima u razvoju scientometrije. SCI je prvi citatni indeks znanstvene literature (koji postoji još i danas) i sadrži bibliografske i citatne podatke o radovima objavljenim u odabranim znanstvenim časopisima iz područja prirodnih znanosti, tehnologije i medicine. Nakon razvoja SCI, a posebno nakon dostupnosti ovih podataka u računalnom obliku i razvojem računalne obrade podataka, pojavljuje se sve veći broj scientometrijskih istraživanja navedenih područja.

SCI je originalno zamišljen kao pouzdan izvor za pronalaženje znanstvene literature (Garfield, 1978), a zbog svoje strukture omogućio je statističke analize znanstvene literature na makro razinama. Njegova pojava označava početak bibliometrije kao vrlo snažnog područja u istraživanjima znanosti. Poznati znanstvenici poput Dereka de Solle Pricea i Roberta Mertona vrlo brzo uviđaju važnost ovog izvora podataka. Price iz perspektive moderne povijesti znanosti, a Merton iz perspektive normativne sociologije (Van Raan, 2005). S vremenom ovaj citatni indeks dobiva prestižan status i počinje se često koristiti za vrednovanje znanstvenog rada.

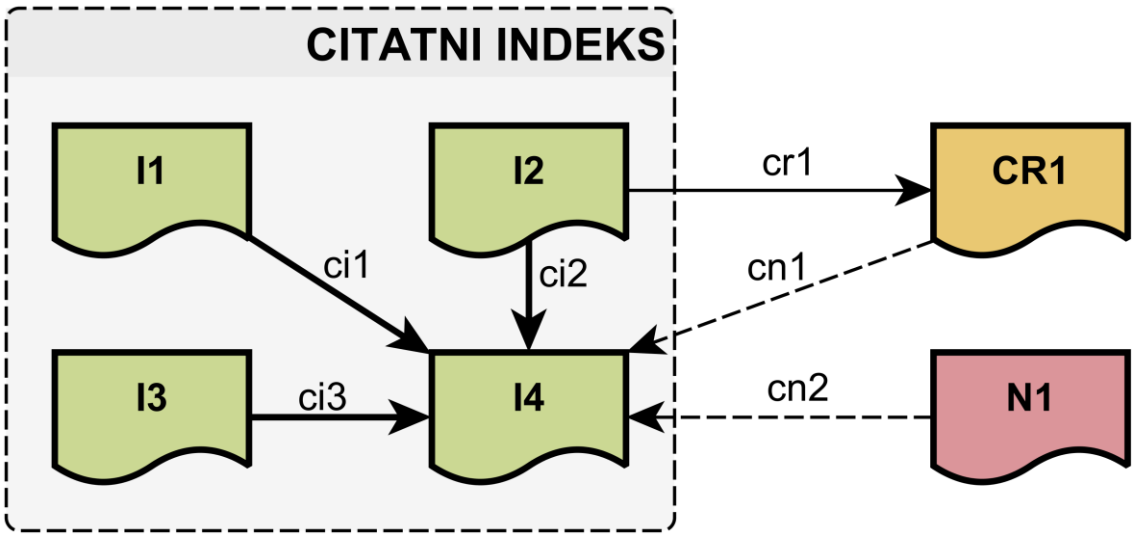
Nastanak citatnog indeksa može na prvi pogled djelovati kao puka tehnološka ekstenzija već poznatog. Kao i kod drugih tehnologija, stvarnost je kompleksnija te u ovom smislu stoji McLuhanova "medij je poruka". Citatni indeksi i konkretno SCI nisu samo omogućili pristup tim informacijama. Samo korištenje takvih alata pridodalo je velik značaj problematici citiranja i značenja citata. Prije korištenja ovakvih alata, citati se implicitno uzimaju zdravo za gotovo i, kao i kod mnogih društvenih konvencija, dolaze u prvi plan tek rijetko i to u negativnom smislu (Cronin, 1984). Sve češće pragmatično korištenje ovakvih izvora postavlja zaokruženu teoriju citiranja gotovo kao sveti gral scientometrije. U skladu s metaforom, preciziranje značenja citata i formalno definiranje teorije citiranja nisu jednostavni problemi, kao što je u uvodnom tekstu o citatnim analizama detaljnije obrazloženo.

Izbor uključene literature za SCI se temeljio na Bradfordovom zakonu (Bradford, 1934), o raspršenosti ($1:n:n^2$) literature (Shenton i Hay-Gibson, 2009). Pojednostavljeno, ovaj zakon kaže da je najveći broj relevantnih radova objavljen u relativno malom broju časopisa što ih čini "jezgrom" časopisa za određeno znanstveno polje odnosno disciplinu. Na primjer, pri pretraživanju radova u časopisima, znanstvenik će pronaći velik broj relevantnih članaka u malom broju časopisa koji tvore jezgru. Izvan jezgre, broj časopisa potreban da se pronađe taj broj članaka dramatično raste što, pogotovo kod tiskanih publikacija, umanjuje efikasnost pretraživanja i nabave literature. Ovaj zakon je stoga logična početna točka u kriterijima selekcije časopisa za indeksiranje ili nabavu. Slične ideje poznate su izvan granica informacijskih znanosti kroz Paretovu distribuciju tj. poznato "80/20" pravilo koje potiče iz tvrdnje da 20% pučanstva kontrolira 80% bogatstva. Kao što je opisano kasnije i vidljivo u rezultatima, ovakva raspodjela je sveprisutna u scientometriji.

Citatni indeks znanstvenih tekstova se može shvatiti kao bibliografska baza podataka s dodanom razinom usmjerene mreže citata među zastupljenim publikacijama. Pojednostavljena

vizualizacija citatnog indeksa koji indeksira četiri publikacije u ukupnoj populaciji od šest je prikazana na slici 2.

Slika 2. Pojednostavljen prikaz citatnog indeksa



Na slici je prikazan citatni indeks koji uključuje (tj. indeksira) četiri rada I1, I2, I3 i I4. Svaki od ovih radova opisan je bibliografskim metapodacima kao i u bibliografskoj bazi. Selekcija ovih radova se temelji na nekom vanjskom kriteriju poput selekcije časopisa, discipline, nakladnika, zemalja i sličnog.

Prikazan citatni indeks sadrži i informaciju o postojanju petog rada, CR1, jer postoji takav rad, I2, koji je uključen u indeks i citira ga, ali za rad CR1 ne postoji detaljan bibliografski zapis, a i sama jednoznačna identifikacija tog rada je upitna kao što je podrobnije opisano u metodologiji. Drugim riječima CR1 nije indeksiran, ali je citiran u indeksiranim radovima. U prikazanom indeksu ne postoji informacija o postojanju rada N1. Rad I3 jedini je citirani rad u ovom indeksu i prema njemu je dobio tri citata (ci1, ci2, ci3), budući da citatni indeks prati citate samo između indeksiranih radova.

Društvene znanosti dobivaju svoju inačicu ovog instrumentarija 1973. godine razvojem Social Science Citation Indexa (SSCI). 1978. godine se pojavljuje Arts & Humanities Citation Index (A&HCI) koji pokriva humanističke znanosti i umjetničko područje. Sva tri indeksa uključuju radove iz odabranih znanstvenih časopisa, a metrika časopisa koja proizlazi iz broja i citiranosti radova pomaže u kontinuiranom vrednovanju uključenih časopisa. Iako postoji manje publikacija koje se bave bibliometrijskim istraživanjima društvenih znanosti, humanistike i umjetnosti, mnoga navode nedostatke spomenutih citatnih indeksa za ovu

svrhu, tj. objašnjavaju nužnost nadopune ovih izvora za obuhvatnija istraživanja npr. (Van Leeuwen et al., 2003; Narvaez-Berthelemot i Russell, 2001; Nederhof et al., 1989; Sivertsen i Larsen, 2012).

Glavni navedeni nedostaci su:

- dominacija radova u časopisima, nedostatak ostalih publikacija
- orijentiranost na literaturu na engleskom jeziku
- različita primjerenost opsega navedenih baza za različita područja, npr. humanističke znanosti i neka polja društvenih znanosti

Kao danas aktualni citatni indeksi koji pokrivaju sva znanstvena područja često se spominju i uspoređuju Web of Science (WoS), Scopus i Google Scholar.

WoS je direktna kontinuacija originalnih ISI (Institute for Scientific Information, Philadelphia) citatnih indeksa koje je 1992. godine kupila Thomson korporacija, a danas su dio Thomson Reuters korporacije nakon spajanja tih dvaju tvrtki 2008. godine. SCI, SSCI i A&HCI danas su, dakle, dio WoS usluge[1]. Prema izdavačima, Web of Science je "najkompletnija i najversatilnija platforma za istraživanja". U trenutku pisanja ovog doktorata, WoS pruža pristup 20 različitih baza različitog opsega i sadržaja. Većina sadrži bibliografske informacije o znanstvenim publikacijama u nekom okviru (poput SCI ili Chinese Science Citation Database). Neke su vanjske baze koje su uključene u više usluga za pristup bazama poput CAB Abstracts ili MEDLINE baza. Neke, pak, nisu baze znanstvenih tekstova već drugih informacija poput taksonomija (Zoological Record) ili podataka (Data Citation Index). 10 od ukupno 20 baza mogu se interpretirati kao citatni indeksi.

Scopus je citatni indeks u vlasništvu tvrtke Elsevier koji je kao i WoS dostupan *online* uz pretplatu. Za razliku od usluge WoS, Scopus se može interpretirati kao jedna baza koja sadrži[2] 21 000 naslova od preko 5 000 međunarodnih nakladnika, od čega 20 000 recenziranih časopisa, 390 profesionalnih publikacija i 370 serijala knjiga. Uz navedeno, sadrži i 5.5 milijuna radova s konferencija. Prema istom izvoru, Scopus se kao citatni indeks može smatrati u potpunosti relevantnim za uključene radove počevši od 1996. godine iako selektivno uključuje i informacije do 19-og stoljeća.

[1] http://thomsonreuters.com/thomson-reuters-web-of-science , 10.12.2013.
[2] http://www.elsevier.com/online-tools/scopus/content-overview 10.12.2013.

Google Scholar citatni je indeks u vlasništvu tvrtke Google, koji je drugačiji u pristupu od WoS-a i Scopusa. Google Scholar se temelji na općenitom indeksiranju weba, a ne na strogo definiranom opsegu pa mu je opseg puno širi od WoS-a i Scopusa. Navedeno ga čini podobnijim za istraživanja u kojima je potrebno uključiti publikacije koje se ne nalaze u WoS-u i Scopusu tj. zahvatiti inače nevidljiv dio produkcije. Unatoč tome, Scholar je predmet čestih kritika. Podaci iz bibliografskih baza, uključujući tu i prestižne izvore poput WoS-a i Scopusa, uglavnom ne zadovoljavaju potrebe istraživača (Glänzel i Schoepflin, 1994) bez detaljnog postupka pripreme. Kod Google Scholara koji informacije povlači iz izvora neprovjerene kvalitete, situacija je višestruko gora (Jacsó, 2010; Zauder et al., 2011). Navedeno, uz tehničke probleme poput nemogućnosti masovnog izvoza podataka, čini opseg Google Scholara ograničeno korisnim za scientometrijska istraživanja. S pozitivne strane, Scholar zahvaća inače nevidljiv dio produkcije i citiranosti, ali, s negativne strane, u praksi može biti iskorišten samo u istraživanjima koja uključuju manje skupove podataka koje je moguće u cijelosti ručno preuzeti, pregledati i ispraviti.

Također, u bibliometrijska istraživanja se mogu uključiti i razni bibliografski izvori poput bibliografskih baza nakladnika znanstvene literature ili pak kataloga nacionalnih knjižnica (Torres-Salinas i Moed, 2009) iako se ta mogućnost još uvijek rijetko koristi. Kako su računalni dohvat, pohrana, priprema i analiza podataka u znatnom razvoju, navedeni postupci s bibliografskim i citatnim metapodacima iz različitih izvora su često kompleksni te za njih gotovo da ne postoji preporučena metodologija za pripremu podataka ili gotova softverska rješenja koja bi ovaj proces standardizirala.

Baze i pristup bazama važna su tema u dizajnu scientometrijskih istraživanja. Prema van Leeuwenu (van Leeuwen, 2005), možemo razlikovati dva osnovna pristupa. Van Leeuwen ih naziva deksriptivan i evaluativan što nazivno više odgovara percipiranoj namjeni nego strategiji pretraživanja. Prvi, deskriptivan, pristup se odnosi na sveobuhvatne upite, na primjer definirane kroz područje časopisa ili zemlju podrijetla. Drugi, evaluativan, se odnosi na više specifičnih upita koji odražavaju glavnu jedinicu analize. Koji će se pristup koristiti ovisi o razini obrade, ali podaci dobiveni iz sveobuhvatnih pretraživanja gube snagu na nižim razinama. Drugim riječima, dok sveobuhvatan upit kvalitetno opisuje predmet na koji se referira, primjerice znanstvenu produkciju zemlje i obrasce objavljivanja te zemlje na razini različitih područja, isti skup podataka je rizično koristiti za vrednovanje rada individualnog

znanstvenika. Za potonje najprimjerenije je pretraživati formiranjem upita za svakog autora zasebno i zatim rigorozno validirati pronađene zapise.

1.1.2.2 O scientometrijskim istraživanjima

Srž scientometrije tvori razvoj metoda i tehnika za dizajn, konstrukciju i aplikaciju kvantitativnih indikatora o važnim aspektima znanosti (Van Raan, 1997). Tradicionalno, ovi indikatori se uvelike baziraju na bibliometrijskim mjerama poput broja publikacija, citata ili određenog derivata. Svrha korištenja ovakvih indikatora može biti višestruka. Prema ciljanoj publici scientometrijskih istraživanja možemo ih razdvojiti na scientometriju za scientometričare, scientometriju za pojedine znanstvene discipline i scientometriju za znanstvenu politiku (Glänzel i Schoepflin, 1994). U prvom slučaju radi se o bazičnim scientometrijskim istraživanjima s namjenom u napretku teorije ili metode. U drugom slučaju, radi se o istraživanjima znanstvenih disciplina tj. uključenih časopisa, autora, ustanova, zemalja i sličnog s ciljem razumijevanja razvoja ili prikaza stanja u nekoj disciplini. Scientometrija za zadnju skupinu u službi je pružanja indikatora za upravljanje znanošću i upravo je ovakva upotreba najčešće primijećena i kritizirana (što je podrobnije opisano kroz opis scientometrijskih analiza i pokazatelja u metodologiji uz osvrt na podatke).

Upotreba scientometrije s temeljem na bibliometriji za potrebe vrednovanja rezultata (i.e. "izlaza") znanstvenog rada često se naziva evaluativnom bibliometrijom (eng. *evaluative bibliometrics*) (Narin, 1976; Narin, 2012). Francis Narin, koji je prvi počeo koristiti koncept "evaluativne bibliometrije", razvija bibliometrijske pokazatelje znanstvenog uspjeha i to primarno na makro razinama poput znanstvenog uspjeha zemalja. Njegov rad je značajno doprinio sustavnim mjerenjima znanstvenih aktivnosti (Van Raan, 2005).

Od tada je ovakva upotreba scientometrije sve češća jer potreba upravljanja znanstvenim sustavom na objektivnim i transparentnim meritokratskim principima čini scientometrijske postupke korisnim u ove svrhe. Društvene okolnosti, poput sve manjih sustavnih potpora znanstvenim istraživanjima, smanjenog dotoka financijskih sredstava iz državnih izvora, sve veći broj znanstvenika i sličnih rezultiraju potrebom za dodatne indikatore u procesima planiranja i odlučivanja.

Drugim riječima, radi se o pomoći pri pronalasku odgovora na organizacijska pitanja poput: Kako uspostaviti sustav zapošljavanja, napredovanja i nagrađivanja znanstvenika? Kako dodjeljivati sredstva ustanovama? U koja znanstvena područja više ulagati? Koji projekt

financirati? Kako pratiti izvrsnost znanstvenih i znanstveno-nastavnih ustanova na nacionalnoj, regionalnoj i svjetskoj razini? Transparentni odgovori na postavljena pitanja su od velike važnosti za upravljanje sustavom znanosti, a kvantitativna proučavanja literature pružaju korisne informacije prilikom traženja istih. U ovom smislu, najčešće upućena kritika ovim indikatorima je u njihovu korištenju više nego konstrukciji tj. u preširokoj interpretaciji informacija koju pružaju ili zanemarivanju nedostataka izvora na temelju kojih su izračunati. Jednostavni primjeri su korištenje broja radova koji se u scientometriji često uzima kao mjera produktivnosti kao mjeru svih aspekata produktivnosti nekog znanstvenika i korištenje citata kao mjere kvalitete na mikro razinama prije nego užeg koncepta poput odjeka (Phelan, 1999).

Uz vrednovanje, scientometrija se često koristi u svrhu opisa ili "mapiranja" znanosti radi boljeg razumijevanja ili radi omogućavanja bogatijih sustava za pronalaženje i upravljanje znanstvenom literaturom. U prvom slučaju se često radi o ranije spomenutoj "scientometriji za znanstvene discipline" ili o sociološkim i povijesnim istraživanjima znanosti koja uključuju kvantitativne metode u značajnoj mjeri, a u drugom o korištenju koncepata i vizualizacija korištenih u scientometrijskoj metodologiji, npr. vezanih uz mapiranje znanosti, u svrhu izrade bogatijih informacijskih sustava za organizaciju znanja ili prezentaciju podataka. U ovom posljednjem slučaju scientometrija ima najviše dodira s knjižničarstvom i vezanim dijelom informacijskih znanosti (Van Raan, 1997).

U svrhu proučavanja suradnje i mapiranja znanstvene literature koriste se informacije o supojavnosti aktera, citata ili tema. Koautorstvo je jednostavan primjer ovakvog pristupa dok kompleksniji uključuju ko-citatne analize ili supojavnost riječi.

U scientometrijskoj literaturi najčešće zastupljeni fenomeni koji se proučavaju jesu produktivnost, citiranost i suradnja i to manifestirana kroz koautorstva. Neka individualna istraživanja će ovisno o cilju i ulaznim podacima koristiti drugačije pokazatelje za isti fenomen od drugih istraživanja koja proučavaju sličnu ili čak istu tematiku. Razlozi ovome su višestruki: različite razine promatranja (e.g. autor, ustanova, časopis, područje ...) svojstva ulaznih podataka tj. izvora iz kojeg su dobiveni, vremenska dinamika, namijenjena svrha rezultata (e.g. rangiranje znanstvenika ili ustanova s jedne strane i opis područja s druge). U nastavku su ukratko opisana ova istraživanja, a problematika njihove operacionalizacije, u ovom istraživanju, ali i općenito, je opisana u metodologiji.

Za operacionalizaciju scientometrijskih istraživanja od posebnog značaja je i priprema ulaznih podataka. Radi značajki ulaznih podataka, njihova priprema je jedan od osnovnih postupaka u scientometriji. Naime, radi se o tekstualnim podacima sabranim kroz više procesa (e.g. ručnim prikupljanjem, automatskim pobiranjem, ručnom ili automatskom pripremom za pohranu), a zatim preuzetim iz agregatnih baza. U ovim procesima greške i varijante mogu nastati na različitim mjestima (e.g. kod autora, nakladnika, oblikovanja za objavu, preuzimanju u agregatne baze, itd.), a i sama priroda podataka donosi mnoge probleme prirodnog jezika poput homonimije i sinonimije. Ovaj proces je podrobnije opisan u metodologiji s obzirom da se radi o problematici koja je preduvjet za bavljenje scientometrijom, više nego predmet promatranja poput tema opisanih u nastavku.

Istraživanja produktivnosti tj. broja objavljenih publikacija u nekom okviru i odjeka tj. broja citata u nekom okviru, svakako su među najvidljivijim dijelovima scientometrije, bilo kroz korištenje za potrebe napredovanja bilo kroz popularne derivirane pokazatelje poput faktora odjeka časopisa. Među ovim temama nailazimo i najviše kritike scientometrijskim pokazateljima. Kritika se najčešće odnosi na nekritičku aplikaciju rezultata, ali i na korištenje metodoloških postupaka i izvora podataka u operacionalizaciji istraživanja koji ne odgovaraju namjeni rezultata. Ipak, radi se o korisnim konceptima i istraživanjima za proučavanje znanosti. U nastavku slijede općeniti opisi istraživanja produktivnosti i citiranosti u scientometriji kako bi se prikazali kao važan dio scientometrije i pojasnila problematika. Dodatan opis ovih koncepata vidljiv je kroz metodologiju u kojoj se detaljnije opisuje problematika preciznog dobivanja indikatora o produktivnosti i citiranosti.

Uz produktivnost i citiranost, značajno mjesto u scientometriji ima proučavanje znanstvene suradnje. Znanstvena suradnja je tema koja se često proučava kao zasebna problematika i to metodama drugačijim od onih koji se koriste za uobičajena istraživanja produktivnosti i odjeka. Također, istraživanja znanstvene suradnje primaju manje kritike izvan scientometrije, najvjerojatnije zato jer se ne koriste za dodjelu sredstava ili napredovanje u znanosti tj. pružaju rezultate kojima namjena uglavnom nije vezana uz temu vrednovanja u znanosti. Istraživanja suradnje su stoga prikazana kao zasebna problematika nakon kratkih prikaza istraživanja tj. pokazatelja produktivnosti i odjeka.

1.1.2.2.1 Produktivnost

Ako promatramo znanost kao proces s izlazima i ulazima, broj radova mjeri produktivnost objavljivanja, a ta mjera uzima se kao aproksimacija produktivnosti aktera tj. znanstvenog procesa za potrebu istraživanja znanosti. Također, kao što je prikazano u metodologiji, ta mjera nije toliko jednostavna i jednoznačna koliko se na prvi pogled može činiti. Istraživanja znanstvene produktivnosti većinom pokazuju konzistentne rezultate koji se mogu opisati tipičnom krivuljom produktivnosti koja odražava slične ideje već prikazane kroz Bradfordov zakon tj. Paretovu distribuciju. U smislu produktivnosti autora, sličan princip prvi opisuje demograf i statističar Lotka (1926) koji postavlja već spomenuti Lotkin zakon.

Lotkin zakon opisuje distribuciju objavljenih radova neke skupine znanstvenika koja je izrazito asimetrična, gotovo eksponencijalna. Ta asimetričnost upućuje na činjenicu da je za većinu radova odgovoran relativno mali broj znanstvenika.

Konkretno, Lotka je sugerirao da je broj autora s *n* doprinosa otprilike oko $1/n^2$. Drugim riječima, oko 60% svih autora u nekom polju objavi jedan rad, 15% objavi dva rada, oko 7% tri rada, itd. *Petek2008*. Mali broj autora je visoko produktivan i odgovorni su za relativno velik udio objavljenih radova, a većina objavi samo jedan ili dva rada.

Prema tome, za većinu radova odgovorna je manjina znanstvenika. Konkretno, Lotka je sugerirao da je broj autora s *n* doprinosa otprilike oko $1/n^2$. Oko 60% svih autora u nekom polju objavi jedan rad, 15% objavi dva rada, oko 7% tri rada, itd. (Petek, 2008). Mali broj autora su vrlo plodni i odgovorni su za relativno velik udio objavljenih radova, a većina objavi samo jedan ili dva rada.

Promatrajući povijesne obrasce objavljivanja, De Sola Price (1963) je formulirao princip (tzv. Priceov zakon) prema kojem znanstvena produktivnost podliježe zakonu "inverznog kvadrata" prema kojem korijen od ukupnog broja svih aktivnih znanstvenika produciraju polovinu svih objavljenih radova. Prema tom principu, 25% autora je odgovorno za 75% radova. Slično, dakle, već spomenutoj "80/20" ideji. Simonton (2004) procjenjuje da kad bi se u nekom slučaju publikacije znanstvenika iz donje polovice distribucije po produktivnosti zauvijek izgubile, svaka disciplina bi i dalje zadržala 82% svojih publikacija.

U većini znanstvenih disciplina dobiva se dakle tipična asimetrična krivulja (u obliku slova L) koju karakterizira visoka varijanca i relativno niska prosječna produktivnost i to bez obzira na duljinu zahvaćenog razdoblja odnosno radnog iskustva autora/znanstvenika. Asimetričnost ove krivulje čini parametrijsku statistiku gotovo neupotrebljivom. Korištenje aritmetičke sredine, na primjer, u ovom slučaju može navesti na krivu interpretaciju podataka jer se potencijalno velika razlika u aritmetičkim sredinama često može pripisati nekolicini slučajeva tj. ekstremima (eng. *outlier*). Za razliku od podataka dobivenih, na primjer, kroz očitanja na mjernom instrumentu, ovi ekstremi nisu aberantni već su značajka predmeta proučavanja, a u mnogim slučajevima prikazuju najistaknutije autore tj. časopise, citirane radove i slično. Navedeno utječe na razvoj specifične scientometrijske metodologije u smislu koje pokazatelje koristi, među kojima su i pristup istraživanju produktivnosti i citatnim analizama.

U većini istraživanja koja se bave produktivnošću pojedinog znanstvenika, ustanove ili zemlje, kao mjera produktivnosti se koristi ukupan broj objavljenih radova (ili neka derivirana mjera, poput prosječne stope objavljenih radova pojedinca tj. ukupan broj radova podijeljen s brojem aktivnih godina znanstvenika u promatranom razdoblju). Upotreba pojma "produktivnost", dakle, u scientometriji je precizna, ali uska. Kada se ovakva aproksimacija (što je slučaj, kao što ćemo vidjeti, i kod citata), koristi na makro razinama, ona omogućuje operacionalizaciju proučavanja različitih aspekata znanosti kao društvene institucije. Na primjer, ovakva definicija produktivnosti je svrsishodna za proučavanje odnosa između društvenih okolnosti ili informacija o autorima (e.g. spol) i produktivnosti znanstvenog rada ili pak za promatranje različitih razdoblja u razvoju nekog područja ili u znanstvenoj djelatnosti neke zemlje ili regije.

Problem nastaje kada se ovi pokazatelji koriste na mikro razinama bez potrebnih prilagodbi tj. bez posebnih strategija prikupljanja podataka ili konceptualno kao ukupna mjera produktivnosti pojedinca. Drugim riječima, problem nije samo u potrebnosti uključivanja ostalih aspekata znanstvene produktivnosti u koncept već i u samim podacima. Na razini pojedinaca, pogotovo među različitim područjima, problemi inherentni u izračunu iz podataka preuzetih iz baza podataka su izraženiji nego kod makro studija. Greške kod pojedinaca su često lokalizirane kod problematičnih autora, poput onih s više varijanti pisanja imena. To što je rad nekih autora krivo interpretiran (na primjer, kao rad više različitih autora) neće bitno utjecati na rezultate na makro razinama ukoliko je postotak takvih pogrešaka malen, ali kod procjene rada pojedinca hoće, jer se podaci promatraju za svakog pojedinog autora.

Pored toga, važnost i učestalost objavljivanja se razlikuje s obzirom na konkretno radno mjesto i status znanstvenika, polje (i tematiku) kojom se bavi, a ovisi i o širem povijesnom i društvenom kontekstu. Situacijski faktori snažno utječu na takve mjere radnog učinka, pa se korištenje *isključivo* takvih indikatora za usporedbu produktivnosti različitih pojedinaca i njihovu evaluaciju ne može smatrati valjanim (Beck et al., 2013). Naime, takva operacionalizacija samo djelomično zahvaća konstrukt radnog učinka određenog znanstvenog djelatnika, budući da ne uključuje druge aktivnosti u kojima znanstvenik može biti produktivan (poput nastavne aktivnosti, aktivnosti vezane uz diseminaciju rezultata i popularizaciju znanosti, itd.), te druge komponente radnog ponašanja (poput drugih organizacijski važnih ponašanja, ali i kontraproduktivnih oblika radnog ponašanja). Iz toga slijedi da je znanstvena djelatnost, poput drugih složenih radnih aktivnosti, multidimenzionalna, i nije uputno donositi procjene o produktivnosti individualnog znanstvenika isključivo na temelju jedne dimenzije.

Navedeno, naravno, ne znači da se podaci o broju radova ne bi trebali ili ne mogu koristiti i na mikro razinama. Znači, međutim, da takva istraživanja moraju biti posebno operacionalizirana kako bi umanjila mogućnost pogrešaka na razini na kojoj promatraju podatke te da se dobiven scientometrijski pokazatelj produktivnosti treba koristiti kao jedan od aspekata multidimenzionalnog koncepta produktivnosti znanstvenika.

1.1.2.2.2 Citatne analize

U znanstvenim radovima, ranije spomenuti odnosi validacije i povjerenja se manifestiraju kroz citate tj. kroz direktno povezivanje s drugim recenziranim publikacijama. Citat se često smatra društvenim priznanjem intelektualnog utjecaja. Ipak, bez obzira na sveprisutnu upotrebu u znanstvenoj literature, preciznija interpretacija značenja citata je sporna. Intuitivno, temeljna uloga citata je omogućavanje nastavka razvoja znanja referiranjem na ranije objavljene informacije što omogućuje korištenje informacija ili problematizaciju istih bez ponovnog objašnjavanja ili dokazivanja. Točni razlozi citiranja, međutim, daleko su kompleksniji i odražavaju spomenute kompleksne društvene odnose u znanosti.

Utjecajan (Sugimoto, 2014) pregled različitih načina na koji se može pristupiti analizi razloga citiranja može se pronaći u (Cronin, 1984). U pokušaju sažetka korisno je prikazati dio podjele razloga citiranja od Chubin and Moitra (1975) koji dijele citate na:

- afirmativne citate
    - centralne citate
    - dodatne citate
- negirajuće citate

Drugim riječima, citati ne moraju značiti pozitivan odnos prema citiranome. Također, pitanje uloge citata se ne tiče samo vrste nego i stupnja upotrebe. Na primjer, citati pruženi u pozitivnom smislu mogu biti centralni tematici, ali mogu se prenositi i kao dodatna literatura. Dodatno pitanje je sam kontekst citiranja. U bilo kakvom pokušaju popisivanja razloga citiranja kompleksnost problematike postaje vidljivija. U nastavku je prikazan popis razloga citiranja koji je sabran iz različitih predočenih kategorizacija u (Cronin, 1984) u svrhu problematizacije pojma.

Razlozi citiranja:

- odavanje priznanja seminalnim autorima i pionirima
- povijesni pregled i identifikacija publikacija koje su prve pojasnile koncept i/ili postavile terminologiju
- identifikacija metodoloških postupaka
- podrška tvrdnjama
- referiranje na rezultate drugih istraživanja radi usporedbe
- ispravka pogrešaka u ranijem radu
- kritika ili opovrgavanje informacija u drugim publikacijama
- ukazivanje na odabranu literatura koja nije direktno vezana za glavnu temu
- bibliografske reference na slabije poznate ili teže dostupne publikacije
- informiranje drugih istraživača o publikacijama koje će se tek objaviti

Kao što vidimo, iako je citat intuitivan mehanizam u znanosti, precizno i jednoznačno korištenje pojma nije jednostavno. Također, popis naveden gore uključuje samo razloge citiranja koji se mogu smatrati znanstvenima. Kada se mehanizam citiranja ne samo osvijesti već i koristi za kasniju procjenu rada, moguće je i citiranje citiranja radi tj. svjesno igranje "igre citata" poput ciljanog citiranja određenih publikacija kako bi se utjecalo na kasniju metriku ili uključivanje citata radi društvenih normi više nego tematske relevantnosti.

Ipak, dok svakako treba biti svjestan opsega i problematike koncepta, analiza citata proučava ključan mehanizam znanstvene literature. Također, podaci često pokazuju da razlozi citiranja nisu toliko međusobno različiti i pružani po slučaju kako ne bi bili pouzdani pokazatelji odjeka (Van Raan, 1998). Abt (2000), na primjer, pokazuje da kad stručnjaci odaberu radove koji smatraju važnima (bez uvida u citiranost) ti radovi su znatno citiraniji od slučajno odabranih radova.

Citatne analize u scientometriji potiču od Mertonovog modela znanosti kroz rad Eugena Garfielda (1955) i ključan su pristup u analizi znanstvenih publikacija. Originalna ideja povezivanja radova putem citata u citatnom indeksu temelji se na ideji tematske relevantnosti citiranih radova za citirajući rad. Preciznije rečeno, originalna namjena SCI citatnog indeksa je pomoć pri snalaženju u sve većem broju znanstvene literature, a citat je interpretiran kao tematska poveznica između znanstvenih radova koja je korisna za pronalaženje literature. Kasnije, često pragmatično, korištenje podataka o citatima nalazi im nove svrhe.

U svakom slučaju, osnovna postavka citatnih analiza je da su citati prebrojivi te da se iz tako kvantificiranih podataka može doći do relevantnih informacija o znanstvenoj literaturi pa tako i o znanosti i znanstvenom procesu. Dok ova pretpostavka stoji u korištenju kvantitativnih citatnih informacija u mnoge svrhe, važno je biti svjestan i problematike vezane uz razloge citiranja i značenje citata. Na primjer, osnovna postavka tradicionalne citatne analize je pridodavanje iste vrijednosti svakom citatu, tj. svaki citat vrijedi "1 citat" bez obzira u koje svrhe je korišten u radu. Skraćeno rečeno, kao i kod mnogo toga drugog u scientometriji, važno je učiniti konceptualnu razliku između pokazatelja (koji treba biti razvijen i korišten u posebne svrhe (Van Raan, 2005) i čija informativnost i validnost često raste u kombinaciji s drugim pokazateljima) od egzaktnih mjera koje same po sebi daju nedvosmislenu i preciznu informaciju.

Citatne analize su nedvojbeno među osnovama scientometrijske metodologije (Zitt, 1991). Među važnim radovima za njihov status u scientometriji svakako treba spomenuti: (Narin, 1976; Garfield, 1979; Cronin, 1984). Narin (1976) je prikazao mogućnost upotrebe citatnih podataka u svrhu procjene rezultata znanstvenog rada. Garfield (1979) pokazuje kako je citatna analiza legitiman i koristan alat ukoliko se razborito koristi. Cronin (1984) je demonstrirao da citati, kao i mnogi scientometrijski pokazatelji, funkcioniraju u širem društvenom kontekstu (Sugimoto, 2014).

Važni koncepti u razvoju citatnih analiza su koncepti bibliografskog uparivanja (Kessler, 1963) i ko-citiranja (Small, 1973). Bibliografsko uparivanje uparuje radove pomoću njihovih popisa literature tj. radovi se uparuju ako citiraju neke iste radove. Bibliografski parovi opisuju statičnu informaciju jer su "upareni" u trenutku objave i taj status se ne mijenja.

Vremenski dinamičnije je pratiti parove iz drugog smjera tj. smjera citiranosti gdje su radovi upareni kroz citate koje dobivaju u novoj literaturi. Small (1973) i Marshakova (1973) neovisno su razvili ko-citatnu analizu 1973. godine. Kocitatna analiza se može jednostavno opisati na sljedeći način: ako su dva rada zajedno citirani u nekoj publikaciji, ta dva rada su međusobno povezani. U ovom smislu frekvencija ko-citiranja može se shvatiti kao mjera sličnosti između dva rada. Ta mjera, za razliku od bibliografskog uparivanja, može rasti kroz vrijeme kako se objavljuju novi radovi koji ko-citiraju postojeću literaturu.

Ovakav pristup povezivanju publikacija često se naziva "mapiranjem znanosti". Mapiranje znanosti nastoji otkriti strukturu i dinamiku znanosti koristeći atribute znanstvene komunikacije, posebice znanstvenih publikacija (van den Besselaar i Heimeriks, 2006). U ovom kontekstu, dva glavna pristupa mapiranju znanosti u scientometrijskoj literaturi su kocitatna analiza i, u manjoj mjeri, analiza supojavnosti riječi (eng. *co-word analysis*).

Seol i Park (2008) razvrstavaju objavljene radove koji se bave citatnim analizama u šest tematskih područja:

1. priroda citata i citiranja
2. samo-citiranje
3. visoko citirani radovi
4. analiza specijaliziranih područja
5. aspekti korištenja rezultata
6. ko-citiranost

U ovaj popis mogla bi se dodati i tema starosti citata tj. zastarijevanja literature koja je važan dio citatnih analiza s vlastitim prepoznatim pokazateljima poput polu-života citata ili medijana starosti literature. Proučavanje zastarijevanja literature, na primjer, pokazuje primjerenost vremenskog okvira za citatne analize.

Tema visoko citiranih radova posebno je značajna za ovaj rad, jer u svrhu promatranja područja kroz visoko citirane radove promatramo ne samo vjerojatno najznačajniji dio nekog korpusa već i umanjujemo šum prisutan u samim razlozima citiranja koji je među najčešće zamjeranim problemima citatnoj analizi. Drugim riječima, analiza visoko citiranih radova je zahvaća radove važne za razvoj predmeta promatranja, u ovom slučaju discipline, ukoliko ga, naravno, ulazni skup radova kvalitetno reprezentira.

Interpretacija značenja citata među najčešćom je kritikom citatnim analizama (Phelan, 1999). Problem je sličan problemu "produktivnosti" u scientometrijskom smislu nasuprot produktivnosti neke osobe u radnom odnosu. U ovom kontekstu, u scientometrijskim i to posebno sociološkim studijama znanstvene literature, brojem citata se može aproksimirati koncept kvalitete izlaza tj. produktivnosti (Henk F. Moed, 2005). Kao i kod produktivnosti, za istraživanja na makro razinama ovo je svrsishodna operacionalizacija, ali problem nastaje kada se koncept koristi kao da je precizno informativan na mikro razinama, na primjer za kvalitetu rada pojedinaca. U ovom kontekstu, šum je još naglašeniji nego kod proučavanja produktivnosti zbog dodane razine problema značenja citata. Tvrdnja da su rezultati rada autora koji je dobio jedan citat manje kvalitetni od autora koji je dobio dva citata nasljeđuju sve probleme već opisane kod produktivnosti, ali uključuju i dodatnu nesigurnost: autor koji je dobio dva citata mogao je biti i dva puta citiran u negativnom kontekstu. Mnogi od ovih problema nestaju ukoliko se citati ne uzimaju kao mjera kvalitete već kao mjera užeg koncepta poput odjeka ili utjecaja i to pogotovo pri proučavanju na makro razinama ili visoko citiranih radova (Phelan, 1999).

1.1.2.2.3 Koautorstvo i suradnja

Pojam suradnje među znanstvenicima odnosi se na zajednički rad istraživača u svrhu postizanja zajedničkog cilja koji je otkrivanje novih spoznaja i stjecanje znanja (Jokić, 2005). U posljednjih nekoliko desetljeća dolazi do dramatičnog povećanja znanstvene suradnje, što se očituje u sve većem apsolutnom i relativnom broju višeautorskih radova u odnosu na jednoautorske radove. U skladu s time, i zbog prevladavajućeg mišljenja da je suradnja pozitivna pojava koju treba poticati u okviru znanstvene politike, mnogi istraživači iz različitih disciplina su počeli istraživati taj fenomen.

Cainelli i suradnici (2010) su na temelju literature o preko 20 godina istraživanja suradnje napravili listu faktora koji determiniraju i potiču suradnju: specijalizacija zbog rasta znanja, potreba za multidisciplinarnim pristupom, potreba za specifičnim tehnološko-analitičkim vještinama, sinergija - u smislu da više ljudi zajednički može doći do ideje do koje nitko od njih ne bi samostalno došao, pritisak za što većim brojem objavljenih radova (tzv. "publish or perish" strategija), veća vjerojatnost prihvaćanja rada za objavljivanje, te poticaj od strane znanstvene politike.

Osim tih objektivnih faktora, lista sadrži i dva subjektivna: smanjenje nesigurnosti zbog procesa prihvaćanja rada i bijeg od izolacije te održavanje motivacije kroz druženje s kolegama. Usto, napredak u informacijskoj i komunikacijskoj tehnologiji umanjuje prepreke u suradnji zbog geografske udaljenosti.

Znanstvena suradnja je složen fenomen koji je moguće istraživati na različitim razinama i na koji utječe veliki broj faktora. Važnost i mogućnost primjene rezultata istraživanja znanstvene suradnje u praćenju kognitivnog i socijalnog procesa razvoja znanosti prepoznate su i prate se u posljednjih 30-ak godina. Prema Katzu i Martinu (1997) do danas su obrađivani različiti aspekti suradnje: kako mjeriti suradnju općenito i koautorstvo, koji faktori utječu i potiču porast istraživačke suradnje, koji su izvori suradnje (koja je uloga komunikacije i učinci fizičke i društvene bliskosti na sklonost suradnji), te koji su učinci suradnje na znanstvenu produktivnost.

Koautorstvo se smatra formalnom manifestacijom intelektualne suradnje u znanstvenim istraživanjima (Acedo et al., 2006). U većini slučajeva predstavlja sudjelovanje dva ili više autora u produkciji objavljenog rada. Koautorstvo na objavljenim radovima često se koristi kao indikator suradnje i to zbog dostupnosti podataka, stabilnosti mjerenja, mogućnosti provjere te neintruzivnosti i nereaktivnosti takve mjere (Sonnenwald, 2007). Unatoč nedostatcima koautorstva kao mjere suradnje to je i dalje najčešće korištena mjera. Kuzhabekova (2011) je podijelila istraživanja koautorstva u dvije skupine: deskriptivna i eksplanatorna istraživanja. Novija istraživanja često pripadaju u obje skupine pa podjela više odgovara fazama pojedinog istraživanja. Deskriptivna istraživanja koautorstva su usmjerena na procjenu učestalosti pojave koautorstva kod istraživača kroz vrijeme i u različitim disciplinama i grupama istraživača. Neka istraživanja mjere koautorstva pokušavaju povezati s sociodemografskim osobinama znanstvenika (dob, spol, zemlja u kojoj djeluju). Po svojoj prirodi ta istraživanja su najčešće kvantitativna. Prema novijim podacima (Feinberg et al.,

2011), u većini društvenih znanosti dolazi do rasta višeautorskih radova, no taj je rast izražen u različitoj mjeri.

Eksplanatorna istraživanja koautorstva pokušavaju razumjeti razloge zbog kojih dolazi do koautorstava, te nerijetko koriste kvalitativnu metodologiju. Široko prihvaćeno objašnjenje povećanog trenda koautorstva jest da ono ima prednosti u odnosu na jednoautorske publikacije jer timski rad povećava istraživačku produktivnost u terminima kvantitete i kvalitete objavljenih publikacija. Ductor (2011) se bavio empirijskim istraživanjem koautorstva i akademske produktivnosti i zaključio da rezultati istraživanja ne dovode do jednoznačnog odgovora i ne postoji slaganje o tome da li je taj odnos pozitivan, negativan ili nepostojeći. Pretpostavka je da produktivnost raste jer koautorstvo dovodi do mnogo koristi uz manje ulaganja no što je to slučaj u samostalnom radu (Duque et al., 2005; He et al., 2009).

Značajan metodološki napredak u istraživanju znanstvene suradnje nastao je primjenom metodologije analize mreža. U slučaju mreže koautora možemo pričati o društvenoj mreži i o analizi društvenih mreža (eng. *social network analysis* - SNA). Društvene mreže su grupe međusobno povezanih pojedinaca koji imaju neki zajednički atribut. SNA je interdisciplinarni pristup koji uključuje skup metoda, mjernih koncepata i teorija koje omogućuju empirijsko mjerenje društvenih struktura i okoline unutar koje pojedinac funkcionira (Borgatti et al., 2009). Prema Percu (2010), društvene mreže su zanimljiva vrsta složenih veza, a koautorstvo znanstvenika predstavlja njen jasan i gotovo prototipski primjer. Znanost se može definirati kao društvena mreža (struktura) znanstvenika: koja se manifestira kroz odnose suradnje među znanstvenicima i kao kognitivna mreža (struktura) znanja: sastoji se od odnosa među znanstvenim idejama (Boerner et al., 2012).

U scientometriji i sociologiji znanosti se često koriste analize mreža za razumijevanje koautorstva i proučavanje razvoja znanstvenih disciplina. Pritom je fokus na nastanku novih pod-disciplina te povezivanju među postojećim i stvaranje novih zajednica (Moody, 2004; Kronegger et al., 2012; Perc, 2010). Također, kroz ovaj pristup moguće je definirati ključne autore kao i pratiti razvoj područja u odnosu na njihovo povezivanje. Mreža nekog uskog područja koja sadrži više nepovezanih grupa daje drugačiju informaciju nego mreža područja u kojemu postoji jedna velika grupa kojoj pripada većina autora. U ovom pristupu se uzorak autora često definira preko autora svih radova objavljenih u određenom časopisu u nekom određenom razdoblju (npr. Chen, Börner, & Fang, 2012).

Koncepti analize kompleksnih mreža koja polazi od teorije grafova i analize društvenih mreža nisu dio scientometrije sami po sebi, ali budući da su strukture u znanosti među glavnim temama u scientometriji, s vremenom analiza mreža postaje prominentan pristup za istraživanje mnogih tema pa se i u ovom radu bogato koristi kako je prikazano u metodologiji.

1.1.3 O časopisu *Scientometrics*

Pojava prvog časopisa za neko istraživačko područje važan je prvi korak u institucionalizaciji područja (Jokić i Zauder, 2013). Ovaj trenutak u načelu označava prisutnost kritičnog broja znanstvenika zainteresiranih specifično za područje. (Schummer, 2004). Početak izlaženja časopisa *Scientometrics*, kako se Van Raan (Van Raan, 2005) slikovito izrazio, označava emancipaciju kvantitativnih studija znanosti.

Časopis *Scientometrics* počinje izlaziti 1978. godine i zajednički ga objavljuju izdavači Akadémiai Kiadó i Springer. Prvi broj časopisa otvaraju pozdravne riječi, između ostalih, Dereka de Sole Price i Eugena Garfielda što govori o čvrstoj povezanosti časopisa s glavnim akterima u razvoju discipline.

Prema opisu izdavača, časopis *Scientometrics* je međunarodni časopis usmjeren na sve kvantitativne aspekte znanosti o znanosti. Časopis objavljuje originalne istraživačke radove, kratke komunikacije, preliminarne izvještaje, pregledne članke, pisma uredniku i recenzije knjiga vezanih uz scientometriju. Tematski fokus su istraživanja u kojima se razvoj i mehanizmi znanosti proučavaju statističkim tj. matematičkim metodama.

Scientometrics od svojih početaka ne samo da izlazi kontinuirano već uslijed sve većeg interesa za scientometriju kontinuirano povećava broj članaka koji objavljuje svake godine. Osnivač časopisa (Van Raan, 2005) i urednik od početka izlaženja je Tibor Braun, koji nastavlja biti urednikom časopisa u cijelom promatranom razdoblju, dakle preko 30 godina. 2014. godine, *Scientometrics* dobiva novog glavnog urednika (Glänzel, 2014), Wolfganga Glänzela koji je kao autor s časopisom od samih početaka. Glänzel je čest Braunov suradnik u cijelom promatranom razdoblju, a zajedno s Andrásom Schubertom ova tri autora čine najproduktivniju suradnju kroz koautorstva u časopisu *Scientometrics*, što je vidljivo u poglavlju Rezultati i rasprava.

Časopis *Scientometrics* kroz cijelo promatrano razdoblje jedini je časopis u potpunosti specijaliziran upravo za ovakva proučavanja znanosti. Među ostalim časopisima relevantnim za scientometriju svakako možemo izdvojiti *Journal of the American Society for Information Science (and Technology)*, *Information Processing and Management*, *Journal of Documentation*, *Journal of informetrics*, *Research Policy* i *Research Evaluation*.

2012. godine počinje izlaziti i *Journal of Scientometric Research*[3], ali ovaj časopis je doživio tek dva broja. Promatranjem časopisa *Scientometrics*, dakle, promatramo glavno tijelo scientometrijskih radova.

1.2 O ovom istraživanju

Cilj ovog istraživanja je iskoristiti versatilno operacionaliziran scientometrijski metodološki postupak za istraživanje glavnog tijela scientometrijske literature. Cilj istraživanja je dakle dvojak: 1. kvantitativno opisati glavno tijelo scientometrijske literature s posebnim fokusom na razvoj kroz vrijeme i 2. opisati korake u istraživanju, implementirati glavne scientometrijske postupke dinamičkim programiranjem (s fokusom na zajedničke pod-probleme) i povezati ih s analizom mreža i analizom cjelovitog teksta.

Prvi doprinos ovog rada je dakle u razumijevanju i uopće validaciji scientometrije kao zasebne pod-discipline informacijskih znanosti kroz primarno kvantitativnu obradu glavnog tijela scientometrijske literature tj. svih radova objavljenih u časopisu *Scientometrics* od početka objavljivanja 1978. do 2010. godine s posebnim naglaskom na članke.

Drugi doprinos je u prikazu, implementaciji i povezivanju tradicionalnih metodoloških scientometrijski postupaka s analizom mreža i tekstova radova. S obzirom na kompleksnost ulaznih podataka koji u većini slučajeva nisu stvoreni pod kontrolom istraživača, u sklopu prikaza metodologije posebno je detaljno prikazana i priprema podataka budući da ista može znatno utjecati na rezultate.

U svrhu prikaza rezultata obrade radova objavljenih u časopisu *Scientometrics* korištena je standardna scientometrijska metodologija s naglaskom na analizu mreža, koja se u posljednje vrijeme sve češće koristi i u scientometrijskim istraživanjima. Dodatno, s obzirom da su primarni objekti koji se u scientometriji promatraju njihovim metapodatkovnim opisima, tj.

[3] http://www.jscires.org/

znanstveni radovi, također digitalni objekti u sve većoj mjeri, ovaj rad nastoji pružiti temelje za uključivanje tekstova radova u scientometrijska istraživanja. Budući da se metodologija uključuje veći broj kompleksnih postupaka na više razina i s obzirom na ranije iskustvo u izradi alata za potrebu projekta *Izrada modela vrednovanja znanstvenog rada u RH za sva znanstvena područja*, sva obrada je provedena u programskom jeziku visoke razine Python koji se sve češće koristi u svrhu obrade podataka s mnogim dostupnim modulima specifično namijenjenima u znanstvene svrhe. Također, ne samo da ne postoje gotova lako dostupna rješenja s grafičkim sučeljima u ove svrhe, već je i upitno kako dizajnirati ista radi potrebne versatilnosti pristupa individualnim slučajevima. Ovo se u prvom redu odnosi na potrebne korake u pripremi podataka, ali i na korištenje, testiranje i razvijanje najnovijih analitičkih indikatora koji se počinju pojavljivati brže no što ih je u standardiziran softver moguće uključivati.

Sam pristup ovog rada u skladu je sa pristupom mnogih članaka u časopisu *Scientometrics* koji kroz prikaz rezultata istraživanja koji su sami po sebi tematski zanimljivi prikazuju i neki inovativan metodološki postupak ili problematiziraju istu. Razlog ovome je u tome što sama scientometrijska metodologija nije (još) standardizirana, ali i u širem kontekstu računalne obrade digitalnih zapisa tj. brzog razvoja novih mogućnosti i metodoloških pristupa u obradi podataka općenito.

Rad pruža kratak uvod o znanosti i istraživanjima znanosti s fokusom na proučavanje znanosti kroz znanstvenu literaturu. Nakon toga slijedi poglavlje o metodologiji podijeljeno u dva dijela: priprema i analiza. Radi promatranja ovog tijela literature s više aspekata, rezultati i rasprava su prikazani zajedno i slijede nakon metodologije. Rad završava kratkim zaključkom i popisom korištene literature, a u dodatku 1. je prenesen Python kôd kojim su implementirani postupci opisani u metodologiji, kao i proizvedene tablice i većina slika koji se koriste u tekstu.

2 METODOLOGIJA

Metodologija ovakvih istraživanja je vrlo kompleksna kada se u potpunosti opiše. Razlog tome je informacijsko bogatstvo bibliografskih metapodataka koji se temelje na tekstualnim zapisima te sadrže kompleksne odnose među entitetima. Dodatno, analitički postupci iz godine u godinu dobivaju nove dimenzije uslijed novih spoznaja o računalno potpomognutoj obradi podataka kao i sve više digitalno dostupnih izvora podataka.

Kod većine scientometrijskih istraživanja upravo je priprema podataka vrlo složen proces koji zahtijeva puno vremena, koji uvelike utječe na rezultate, njihovu interpretaciju i, povratno, na postavljanje istraživačkih problema. Kako se upravo u ovom procesu otkrivaju mnogi problemi koji mogu utjecati na kasniju analizu podataka, priprema je detaljno opisana.

U ovom radu svi procesi od najosnovnije pripreme do izrade podatkovnih tablica i vizualizacija su u potpunosti izvedene u programskom jeziku Python (uz pomoć nekoliko posebnih modula, koji su pobrojani kasnije) i potpuno su ponovo provedivi od sirovih podataka do izlaznih tablica. Kroz sve procese postepeno su prikupljani podaci o nekonzistentnostima i, ako je broj potrebnih provjera bio velik te ako je to bilo moguće, razvijen je automatiziran postupak provjere tj. ispravljanja. U suprotnom podaci su provjereni i ispravljeni ručno. U svakom slučaju, rezultati bilo ručne ili automatske provjere, bilježeni su u zaseban skup podataka koji popisuje sve promjene koje se provode nad podacima proizvedenim u procesu osnovne pripreme.

Ovakav pristup omogućuje:

- **postupno prikupljanje** potrebnih promjena ručnim pregledavanjem, automatiziranim procesima ili kombinacijom
- **ponovnu provedivost** cijelog procesa nakon dodataka novih promjena; osigurava ponovljivost automatiziranih procesa radi novih spoznaja, neočekivanih slučajeva, previda te za potrebe testiranja procesa
- **kontrolu promjena**, uvid u sve direktne promjene vrijednosti koje su učinjene na podacima, mogućnost ponovne provjere

- **uvid u problematiku:** putem uvida u podatke prikupljenih u procesu pripreme mogu se dodatno ispitati zanimljivi slučajevi, a analizom skupa promjena se može utvrditi koji su procesi u pripremi podataka bili potrebni i kako bi mijenjali rezultate da nisu implementirani

Dodatak 1. prenosi ključne dijelove ovog kôda. Izostavljeni su popisi promjena podataka jer bi zauzeli previše mjesta kao i pomoćni dijelovi kôda za izvještavanje, uključivanje podataka u tekst doktorata, formatiranje teksta doktorata i slične postupke koji nisu vezani na postupke pripreme ili scientometrijske analize samih podataka.

U skladu s ciljem doktorata, metodologija je prikazana posebno detaljno kako bi prikazala cjelovit pristup ovakvim istraživanjima. Metodologija počinje s kratkim terminološkim bilješkama, a glavni dio je podijeljen u dva glavna poglavlja: Preuzimanje i priprema te Analiza.

2.1 Terminologija

U tekstu ovog rada pazilo se kako bi terminologija bila čim jasnija. U nastavku slijedi pojašnjenje termina koji su važni za razumijevanje podataka i rezultata.

Važna distinkcija tiče se svih radova i priloga u nekom časopisu i članaka u časopisu. U scientometrijskim istraživanja, važno je razlikovati publikacije koje prenose primaran sadržaj od publikacija koje prenose druge vrste sadržaja. Na primjer, u časopisu Scientometrics postoji takav autor koji ima više desetaka svih radova i priloga, ali niti jedan znanstveni članak jer se radi o osobi odgovornoj za objavljivanje priloga u rubrici novosti. Na hrvatskom se često koristi izraz "svi radovi i prilozi" kad se želi zbirno nazvati sve individualne predmete unutar časopisa bez obzira na njihovu vrstu. "Radovi" se u načelu koriste za sve predmete koje prenose primaran sadržaj pri čemu su u razgovornom jeziku često sinonim za "članke". Kako bi se učinila što jasnija distinkcija, u ovom radu koristio se izraz "radovi i prilozi" za sve predmete bez obzira na sadržaj (dakle od ispravke do originalnog istraživačkog članka), a izraz "članci" za sve predmete koji prenose primaran sadržaj. Kratak izraz "radovi" se koristio u općenitom smislu kada za potrebe rasprave nije bilo važno ukazati na vrstu publikacije, odnosno on označava sve radove i priloge bez obzira na vrstu rada. Korištena klasifikacija vrste radova i priloga za potrebe proučavanja radova objavljenih u časopisu *Scientometrics* prikazana je u metodologiji.

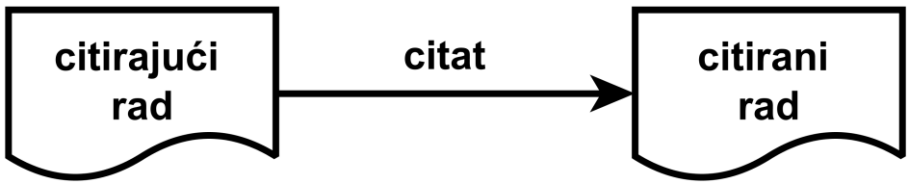
S obzirom da se radi o tiskanoj serijskoj publikaciji, za identifikaciju individualnih svezaka koriste se broj i godište[4]. U ovom radu napušten je uobičajen izraz godište u korist izraza "volumen". Razlog ovome je kako bi se izbjegle konstrukcije poput "Časopis *Scientometrics* godišnje objavljuje tri godišta."

U žargonu vezanom uz znanstvene tekstove često se koristi engleski izraz *full text* kako bi se razlikovao tekst rada od teksta, na primjer, sažetka i kako bi se omogućio razgovor o "pristupu cjelovitim tekstovima radova" za razliku od pristupa samo bibliografskim metapodacima. U direktnom preuzimanju s engleskog jezika, na hrvatskom se često upotrebljava izraz "cjeloviti tekst". U ovom radu, riječ "cjeloviti" je ispuštena kao nepotrebna i koristi se izraz "tekstovi radova" jer je "tekst rada" upravo to tj. riječ "cjelovit" ne doprinosi značenju te fraze.

Također važna distinkcija koja je problematična i na engleskom jeziku tiče se primanja i pružanja citata tj. radova koji te citate primaju i pružaju. Često se koriste distinkcije između citata i referenci, gdje su citati primljeni iz drugih radova, a reference pružene prema drugim radovima. Drugim riječima, ono što je nekom radu referenca (odnosno citat koji taj rad pruža prema drugom radu) je radu na koji se ta referenca odnosi citat (odnosno citat koji je taj rad primio od strane nekog drugog rada). Ovakva terminologija može biti zbunjujuća radi šire upotrebe riječi referenca pa se često izbjegava i na engleskom jeziku (Diodato, 1994) te se koriste pojmovi ulaznih i izlaznih citata (eng. *ingoing and outgoing citations*). Ta terminologija se dobro prenosi na radove koji te citate primaju ili pružaju jer možemo koristiti izraze citirani i citirajući radovi (eng. *cited and citing papers*). Upravo je ovaj izbor terminologije odabran u ovom radu kao najpreciznji i koji se dobro prenosi na koncepte mreža (ulazne i izlazne veze). Slika 3. ilustrira korištenu terminologiju.

[4] digitalni časopisi postavljeni kao baza podataka ne trebaju ovakvu nomenklaturu za identifikaciju, ali ju često koriste kako bi se vezali na prijašnje koncepte i zadovoljili potrebe korištenja literature koju objavljuju e.g. znanstveni časopisi (kao što su PLoS časopisi) postavljeni u obliku baze podataka dodaju nove radove u bazu u grupama s dodijeljenim brojem i volumenom koji zadovoljavaju potrebe standardnog navođenja literature i predstavljaju identifikaciju dodavanja sadržaja u bazu. U ovakvom sustavu, radovi bi mogli biti dodavani u bazu i individualno što bi odbacilo koncept sveska koji je potreban za diseminaciju i pristup u tiskanom obliku

Slika 3. Citirajući rad, citat i citirani rad



U ovom radu kao citirani radovi se primarno promatraju članci iz časopisa *Scientometrics*, a kao citirajući radovi svi radovi i prilozi objavljeni u časopisima indeksiranim u WoS SCI, SSCI i A&HCI citatnim indeksima. Navedeno je podrobnije objašnjeno u metodologiji prilikom opisa podataka i citatnih analiza.

2.2 Preuzimanje i priprema bibliografskih metapodataka

Preuzimanje i priprema podataka za kasniju analizu procesi su od ključnog značaja jer mogu drastično utjecati na dobivene rezultate. Neki od problema su:

- ujednačavanje tekstualnih nizova različitog podrijetla i sadržaja (imena autora, nazivi ustanova, naslovi, ključne riječi ...)
- preuzimanje podataka gdje su upiti temeljeni na ovim tekstualnim nizovima
- kasnija metrika koja se u velikoj mjeri bazira na prebrojavanju i povezivanju ovih tekstualnih nizova
- velik broj provjera potreban za sigurnost obrade
- potreba povezivanja automatskih procesa s ručnim provjerama

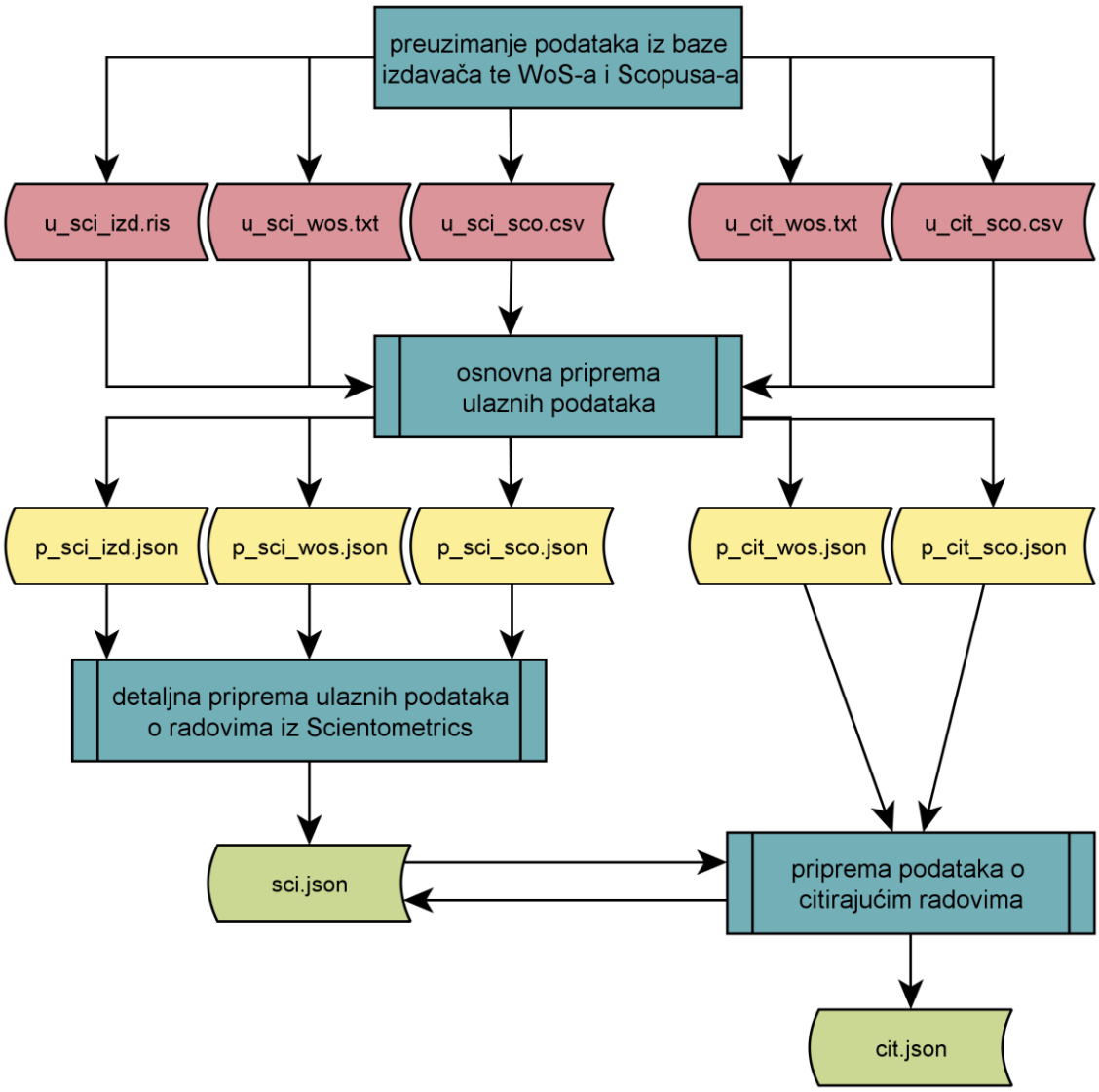
Poseban problem se u mnogim slučajevima javlja kod preuzimanja podataka putem slabije kontroliranih imena, poput imena autora ili ustanova. Kod nekih autora, posebno autorica, postoji i do desetak različitih inačica imena pod kojima se mogu pronaći. Mnoge od ovih inačica nastaju promjenama prezimena te uslijed pogrešaka u pisanju, ali i algoritamskom obradom gdje su, na primjer, algoritmi prilagođeni američkim imenima. Potonje je slučaj kod WoS citatnih indeksa, što predstavlja problem pretraživanju putem imena autora i što je posebno zabrinjavajuće s obzirom da se upravo ti citatni indeksi često koriste za procjene rada pojedinaca.

Kod časopisa je situacija jednostavnija, ali ipak, neki imaju više ISSN (International Standard Serial Number) identifikatora kao i promjena u imenu pa i kod njih treba biti pažljiv. Kod časopisa *Scientometrics* nije bilo problema ni potreba za posebnim strategijama pretraživanja jer se radi o mlađem časopisu jednostavnog naziva koji se nije mijenjao. Dodatno, svi podaci su preuzeti izravno od nakladnika, pa tek zatim iz citatnih indeksa što kroz kros-validaciju ovih skupova omogućuje sigurnost u kompletnost podataka. Ujednačavanje imena autora i ustanova je provedeno tek nakon spajanja ulaznih skupova i to na najinformativnijim imenima autorima kao što je opisano niže u metodologiji.

S obzirom na navedene probleme začuđuje da se scientometrijski radovi češće ne bave ovom problematikom. Iako su recenzije izvora podataka relativno dostupne (Jacsó, 2005), opis procesa pripreme podataka je često izostavljen iz radova što čini istraživanja neponovljivim. Prema vlastitim iskustvima na projektu *Izrada modela vrednovanja znanstvenog rada u RH za sva znanstvena područja* (Jokić et al., 2012), kao i vezanim istraživanjima, strategija preuzimanja i pripreme podataka može prouzročiti i do višestruko više ili manje uključenih radova nekog autora, što je posebno problematično kod procjena rada pojedinaca. Npr. kod nekih autorica, broj radova i citata je nakon intervencija porastao za više nego duplo. Shodno opisanome kao i ciljevima doktorata, priprema podataka za ovaj rad je bogato prikazana kako bi doprinijela budućim istraživanjima srodne vrste.

Krovni procesi u preuzimanju i pripremi bibliografskih metapodataka, kao i glavni ulazi i izlazi vidljivi su s dijagrama prikazanog na slici 4.

Slika 4. Preuzimanje i priprema ulaznih podataka za analizu



Dijagram prikazuje glavne procese u transformacijama podataka od preuzimanja do finalnih skupova na kojima se vršila analiza. Nazivima podatkovnih skupova pridodani su i nastavci kako bi se vidjelo u kojim formatima su bili dostupni i koji format je korišten kao temeljan za obradu. Nakon ovih procesa mogla se izraditi i baza podataka u smislu korištenog softvera (relacijska ili tzv. *key-value store*) na temelju razriješenih podataka. U sklopu ovog doktorata, obzirom da podaci nisu promjenjivi, da im pristupa samo jedan korisnik, da su dovoljno mali da stanu u radnu memoriju i da se analiza opet mora vršiti kakvim vanjskim rješenjima, baza je ocijenjena kao nepotrebna odnosno procijenilo se da nema dovoljno koristi od korištenja novog alata koji unosi dodatnu kompleksnost.

Tekstovi radova, kao drugačija problematika, su obrađeni i prikazani zasebno.

2.2.1 Izvori, ulazni skupovi podataka i njihovo preuzimanje

Kao glavni izvori podataka odabrani su:

- baza izdavača (pristup 05.09.2013)
    - metapodaci o svim radovima zaključno s 2012. godinom (uključujući i uredničke uvodnike, novosti i slično)
    - tekstovi radova u PDF-u zaključno s 2010. godinom
- WoS (pristup 10.10.2013, radovi preuzeti zaključno s 2012. godinom)
- Scopus (pristup 10.10.2013, radovi preuzeti zaključno s 2012. godinom)

Baza nakladnika je odabrana kao najpotpuniji izvor podataka o svim radovima u časopisu *Scientometrics*. WoS i Scopus citatni indeksi odabrani su kao dopuna nakladničkim metapodacima primarno radi podataka o citiranim i citirajućim radovima, ali i radi dodatne validacije podataka uključenih u istraživanje. Podaci o radovima objavljenim u *Scientometrics* su preuzeti zaključno s 2012. godinom kako bi se mogli pratiti samocitati za 2010-u godinu. Izdavački metapodaci su također preuzeti do 2012. godine kako bi poslužili za validaciju WoS i Scopus podataka.

Dodatno, iz ovih baza preuzeti su svi *različiti* radovi koji citiraju *Scientometrics* zaključno s 2012. godinom i to ne uključujući radove iz časopisa *Scientometrics* budući da su isti prisutni u već opisanim skupovima podataka o ovom časopisu. Obzirom da se radi o različitim radovima, broj citata na radove u *Scientometrics* je ukupno veći budući da svaki rad može citirati više radova iz ovog časopisa. Preuzimanje citirajućih radova stvorilo je dodatna dva skupa radova:

- radovi koji citiraju *Scientometrics*, a indeksirani su u WoS citatnim indeksima SCI, SSCI i A&HCI
- radovi koji citiraju *Scientometrics*, a indeksirani su u Scopusu

Od potencijalno korisnih izvora podataka izostavljen je jedino Google Scholar. Odluka za nekorištenje ovog izvora donesena je iz nekoliko razloga:

- nejasan opseg
- nemogućnost preuzimanja podataka u formatima za razmjenu tj. masovnog izvoza
- velik broj grešaka i prevelika vremenska zahtjevnost kvalitetne obrade ovih podataka (Jacsó, 2010; Zauder et al., 2011)
- nepotrebnost nadopune korištenog skupa podataka; u nekim slučajevima ovo je najveća snaga Google Scholara budući da donosi inače nevidljive izvore

Kao zanimljiv primjer može poslužiti prvi rezultat iz Google Scholara (na dan: 15.12.2013.) na upit "Little science, big science" koji prijavljuje 3 256 citata i čiji BiBTeX zapis je:

```
@book{de1986little,
  title={Little science, big science... and beyond},
  author={de Solla Price, Derek John and de Solla Price, Derek John and de Solla Price, Derek
John and de Solla Price, Derek John},
  year={1986},
  publisher={Columbia University Press New York}
}
```

Ovaj zapis sadrži nekoliko grešaka među kojima je zanimljivo što i podložan link tj. predmet na temelju kojeg je zapis stvoren, vodi na uvod u jedno od izdanja knjige "Little science, big science" naslovljen "Little science, big science... and beyond" kojeg su potpisali Garfield i Merton, a ne na samu knjigu. Distinkcija između ovog teksta i knjige je tako izgubljena. Također, kompletnost zapisa iz Google Scholara neusporediva je s podacima nakladnika, a kamoli onih iz citatnih indeksa ili pak iz knjižničnih kataloga.

Što se podataka u ovom istraživanju tičem svakom ulaznom, prijelaznom i finalnom skupu podataka dodijeljena je šifra kako bi se na njih moglo lakše referirati. Šifra je opisna kako bi se izbjegla potreba šifrarnika za razumijevanje njenog značenja. Šifra je u formi "status_sadržaj_izvor". Status početnih skupova podataka je "u" za "ulazni podaci". "Sadržaj" označava da li su podaci o radovima u časopisu *Scientometrics* ili o radovima koji su ga citirali, a izvor da li se radi o podacima izdavača ili iz WoS-a tj. Scopusa. Šifre su formirane na taj način da budi i validni nazivi za imena datoteka i programske varijable što je pomoglo pri imenovanju svih tranzicijskih stanja do pripremljenog skupa podataka.

Ovakvim pristupom došlo se do pet skupova bibliografskih metapodataka koji se mogu podijeliti na sljedeći način:

Podaci o radovima u *Scientometrics*:

- izdavački metapodaci (**u_sci_izd**, N = 3588)
- WoS podaci (**u_sci_wos**, N = 3376)
- Scopus podaci (**u_sci_sco**, N = 3162)

Podaci o radovima koji citiraju radove u časopisu *Scientometrics*, a nisu u njemu objavljeni:

- Web of Science podaci (**u_cit_wos**, N = 6641)
- Scopus podaci (**u_cit_sco**, N = 8095)

Nakladnički metapodaci su preuzeti u jednostavnom tekstualnom formatu za razmjenu metapodataka RIS[5]. Prilikom preuzimanja zapisa o radovima objavljenim u časopisu *Scientometrics* primijećeno je da je svim radovima od početka izlaženja dodijeljena DOI (eng. *Digital object identifier*) vrijednost koja omogućuje jednoznačnu identifikaciju rada u različitim skupovima podataka. U istom trenutku preuzeti su i tekstovi radova u PDF formatu koji su u preuzimanju imenovani DOI-em rada kako bi se lako povezali na bibliografske metapodatke. Web of Science i Scopus podaci o radovima u *Scientometrics* i citirajućim radovima preuzeti su u tekstualnom tabličnom obliku (*tab delimited* tj. csv formatu) koji je uključivao sva polja koja WoS tj. Scopus dopušta za preuzimanje.

Zanimljivo je spomenuti da ni WoS ni Scopus nisu isporučili potpuno validne podatke za obradu. WoS dopušta preuzimanje do 500 radova po datoteci pa se kod izvoza citirajućih radova moralo pristupati inkrementalno. Svaka od preuzetih datoteka (osim zadnje) trebala je sadržavati 500 radova što nije bio slučaj jer je kasnijom provjerom utvrđeno da su neke od njih imale dvostruko manje zapisa. Razlog ovome su bili radovi koji nisu validno zapisani u WoS bazi što ne začuđuje s obzirom na njenu veličinu. Ono što začuđuje, pa i zabrinjava, jest činjenica da je izvoz podataka kad bi naišao na zapis koji sadrži strukturalnu pogrešku jednostavno stao, ali ne bi prijavio nikakvu grešku odajući dojam da je preuzeto svih 500 radova.

[5] format originalno razvijen za potrebe Reference Manager softvera za upravljanje referencama koji je danas, kao i WoS, u vlasništvu tvrtke Thomson Reuters. Od razvitka formata, isti se često koristi kao jednostavan format za razmjenu bibliografskih metapodataka uz slične formate poput formata BibTeX koji je razvijen u sklopu sustava za pripremu dokumenata LaTeX

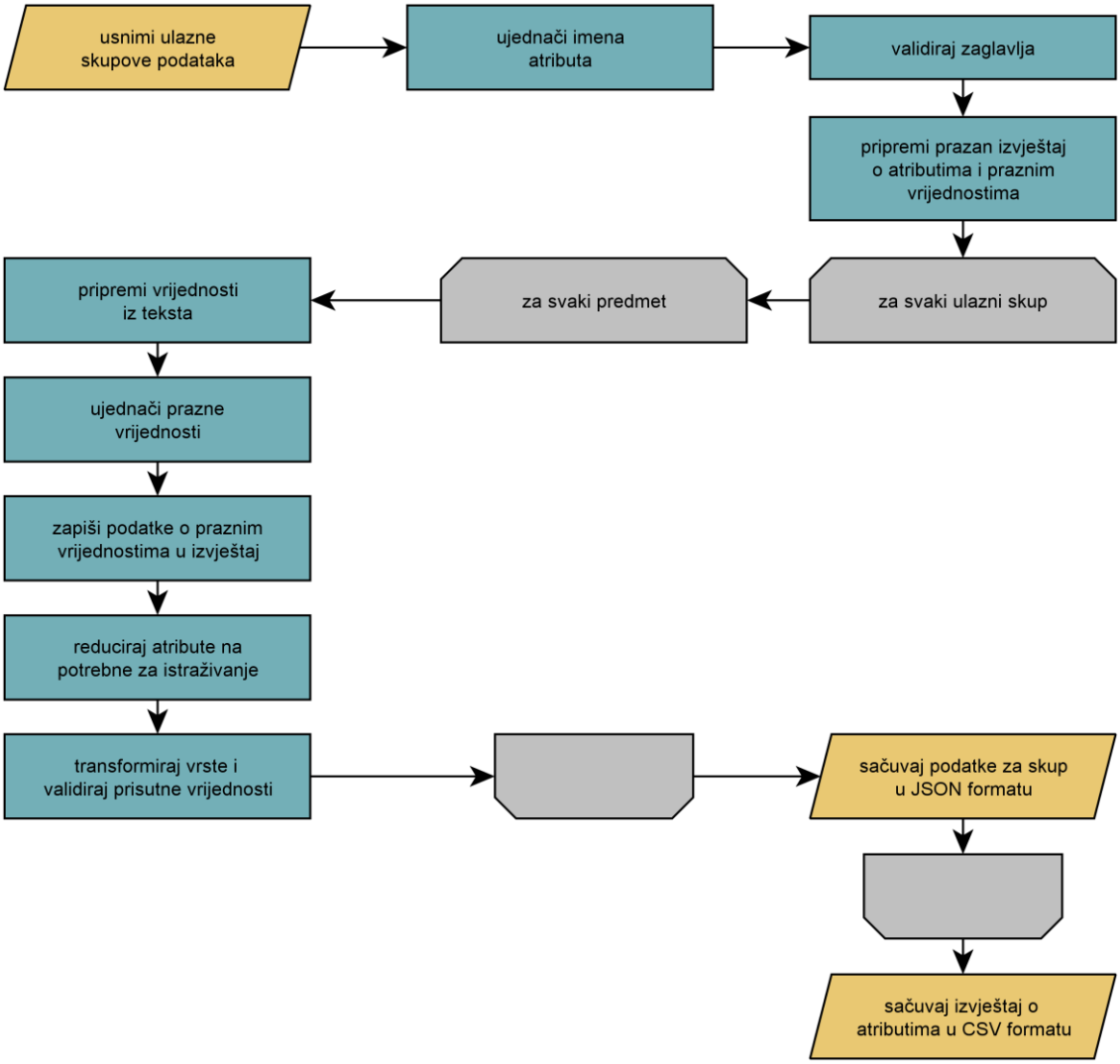
Radi ovakvih grešaka oko 150 radova iz cijelog skupa citirajućih radova (i veći broj citata) je bilo u potpunosti nedostupno. Kod Scopusa je situacija bila nešto lakša, ali je pronađeno nekoliko redaka koji nisu validno preneseni u tekst pa su imali više "atributa" nego zaglavlje datoteke. S obzirom da podaci sadržavaju velika tekstualna polja poput referenci i sažetaka, ovim greškama je bilo vrlo teško ući u trag. Iskustvo, dakle, pokazuje da i na takve greške treba paziti te da treba provjeravati i podatke iz tzv. prestižnih izvora, odnosno onih koji su gotovo standard za scientometrijska istraživanja.

2.2.2 Osnovna priprema ulaznih podataka

U ovom koraku, provedena je osnovna priprema i karakterizacija podataka o radovima u časopisu *Scientometrics* kao i citirajućih radova. Svi procesi su automatizirani i ne zahtijevaju vanjske podatke za operaciju tj. riječ je o mehaničkim radnjama s vrijednostima dobivenim iz različitih baza putem izvoza podataka poput pretvaranja u potrebne vrste podataka i pripreme znakovnih nizova.

Proces je prikazan na slici 5, a implementiran skriptom koja je prenesena u dodatku 1.

Slika 5. Preliminarna priprema svih skupova podataka



Procese ovog koraka pripreme podataka je lako zanemariti jer podaci često mogu odavati dojam "pripremljenosti i validnosti". Ovo vrijedi pogotovo kada sami podaci nisu preuzeti u "sirovom" obliku već su obrađeni kroz kakvo *online* sučelje koje pruža gotove rezultate analiza. Potonji slučaj čini bazu i proces izračuna crnom kutijom pa istraživač povjerenje u preciznost i uopće značenje i relevantnost dobivenih podataka prepušta proizvodu koji često ne samo da ne objašnjava i dokazuje preciznost internih postupaka već su isti često i poslovna tajna.

Svaki propust u ovom koraku može praviti velike probleme kasnije. Najgori od ovih problema su oni koji ne javljaju grešku tj. oni kod kojih kasnije djeluje da je sve pošlo po planu. Jedan od primjera je kada se dva tekstovna niza evaluiraju različitim radi velikih i malih slova, korištenja posebnih znakova ili pak radi različitih *unicode* znakova. Na primjer, nazivi

časopisa "Quality and quantity" i "Quality & quantity" kao i vrijednosti atributa volumen "1-2", "1–2" i "1−2" su različite vrijednosti prema naivnoj računalnoj evaluaciji. Bez standardizacije ovih nizova znakova, raspolažemo s informacijama o dva različita časopisa tj. tri volumena časopisa. Radi navedenog, velik broj ujednačavanja i provjera u ovom koraku rješava mnoge potencijalne probleme kasnije i pruža dodatnu sigurnost u preciznost dobivenih pokazatelja.

Nakladnički metapodaci su iz RIS formata prebačeni u tablični oblik koji je podobniji za obradu. Jednostavan Python RIS parser koji je korišten u sklopu ovog istraživanja može se pronaći u dodatku 1.

WoS i Scopus podaci preuzeti su u tabličnom obliku, pa ovaj korak nije bio potreban. S obzirom da se radi o razgraničenim tekstualnim podacima koji sadrže kompleksna polja poput "reference" ili "sažetak", kod WoS i Scopus podataka je provjerena validnost ulaznih redaka u odnosu na zaglavlje tj. učinjena je provjera da svaki redak ima isti broj atributa kao i zaglavlje. Na ovaj način pronađene su greške u Scopus podacima koje su zatim ispravljene. WoS podaci pak, relativno nestandardno, sadrže praznu vrijednost na kraju svakog retka što je provjereno i utvrđeno da je konzistentan slučaj i zatim je "zadnji stupac" ignoriran. Također, ovaj korak validira i potencijalne propuste u ručnom preuzimanju podataka. Dok se ovdje radi o znanju inherentnom računalnoj obradi podataka i to relativno jednostavnim operacijama, a ne samoj scientometriji, ta problematika se ovdje prenosi kako bi se prikazala važnost tretiranja pripreme podataka kao zasebne problematike više nego procesa s kojeg treba čim prije preći na analizu. Algoritamski pristup na kvalitetno pripremljenim ulaznim podacima čini kvantitativne postupke u analizi visoko efikasnim te zahtijeva daleko manje vremena od pripreme podataka. Potonje, naravno, izuzima slučajeve koji razvijaju nove analitičke postupke.

Ujednačavanje imena atributa, uklanjanje nepotrebnih kao i provjera prisutnosti i vrste njihovih vrijednosti preduvjet su za razumijevanje podataka i njihovu kasniju obradu. Također, s obzirom da se radi o skupovima podataka bogatima informacijama i da su podaci iz citatnih indeksa preuzeti u najpotpunijem obliku, imenovanje i izbacivanje atributa, a pogotovo provjera vrijednosti nije jednostavan proces budući da se kod WoS-a radi o čak 58 atributa, a kod Scopusa o 40. Mnogi od ovih atributa nisu relevantni za svaki skup podataka jer sadrže informacije o imenima kemikalija, PubMed šiframa radova i sličnome. Kod informacija od nakladnika situacija je jednostavnija budući da se radi o podacima koji sadrže

dovoljno informacija za citiranje rada, tj. o osnovnim bibliografskim poljima uz nekoliko dodatnih poput sažetka i DOI-a rada.

Radi tekstualne prirode svih vrijednosti, prvo je u vrijednostima normaliziran prazan prostor bilo koje vrste uključujući i znakove za prijelome retka, tabulatore, ne-prelamajuće razmake i slično. Ovim postupkom su se tekstovni nizovi tipa "riječ-razmak-tabulator-prijelom retka-riječ-razmak-prijelom retka" standardizirali u "riječ-razmak-riječ". Nakon ovog postupka, svi prazni tekstualni nizovi i posebne vrijednosti tipa "[No abstract available]" su označeni standardnom vrijednošću[6] čije značenje je "nedostaje vrijednost". Kod ovog postupka se pazilo i da neke nedostajuće vrijednosti nemaju posebno značenje kao što je to slučaj kod broja citata iz Scopusa kod kojeg nedostajuća vrijednost označava 0 citata.

Provjera prisutnosti i vrste vrijednosti ukazala je na važne nedostatke ulaznih podataka, poput praznih vrijednosti kod jedinstvenih identifikatora tipa DOI ili podataka o izdanju i stranicama, što je omogućilo razvoj strategije spajanja podatkovnih skupova i informiralo o potrebnim nadopunama koje su obavljene u idućem koraku pripreme podataka. Nakon provjera prisutnosti vrijednosti donesena je i konačna odluka o tome na kojim atributima će se temeljiti kasnija obrada. Na primjer, ključne riječi su nažalost isključene iz obrade jer su bile prisutne samo za četvrtinu radova.

Nakon ovog procesa svaki od ulaznih bibliografskih skupova podataka sadržavao je atribute koji su prikazani u tablici 1, uz postotke prisutnih vrijednosti za te atribute tj. uopće postojanju atributa kod različitih skupova.

[6] konkretna korištena vrijednost je bila Python vrijednost None, koja se kasnije u JSON formatu kodirala kao null

Tablica 1. Prazne vrijednosti preuzetih atributa u ulaznim skupovima podataka

atribut	sci_pub	sci_wos	sci_sco	cit_wos	cit_sco
godina izdanja	95.9	100	100	100	100
autori	98.2	99.8	99.7	99.8	99.9
naslov	100	100	100	100	100
naslov časopisa	100	100	100	100	100
volumen	95.9	100	99.6	98.4	91.6
broj	95.9	100	99.6	96	86.7
vrsta rada	100	100	100	100	100
broj citata	n/a	100	100	100	100
broj referenci	n/a	n/a	n/a	100	n/a
reference	n/a	94.8	84.1	100	100
adrese ustanova	n/a	71.6	n/a	n/a	n/a
konferencija	n/a	13	0	n/a	n/a
početna stranica	95.9	100	100	96.8	93
završna stranica	95.9	100	99.2	96.8	92.4
DOI	100	93.8	65.2	76.7	67.7
wos_id	n/a	100	n/a	100	n/a
sco_id	n/a	n/a	100	n/a	100
scopus_url	n/a	n/a	100	n/a	n/a
sažetak	55.1	73.3	92.4	n/a	n/a

n/a = atribut nije prisutan

Za svaki od atributa provedena je provjera i transformacija vrste vrijednosti. Neki atributi su u ovom koraku sačuvani samo radi provjere ulaznih podataka. Konkretne radnje i upotreba atributa iz tablice 1 bile su sljedeće:

Atributi **doi**, **wos_id**, **scopus_url** smatrani su jednoznačnim identifikatorima za izdavačke, WoS i Scopus podatke pa je provjereno da se pojavljuju u 100% slučajeva i da među njima nema duplikata unutar pojedinih skupova. S obzirom da je scopus_url poprilično velika vrijednost od koje je samo mali dio jedinstven, iz ove vrijednosti je preuzeta vrijednost za **scopus_id** atribut koji je korišten kao primarni ključ za Scopus podatke. U ovom smislu primijećeno je da svi radovi i prilozi objavljeni u časopisu *Scientometrics* kod izdavača imaju dodijeljen DOI. DOI stoji za Digital Object Identifier i globalno je jednoznačan identifikator za objekte kojima je dodijeljen, a može se koristiti i za nedvosmislen pristup *online* resursima. DOI je stoga korišten kao glavni identifikator radova objavljenih u časopisu *Scientometrics* te

povezivanje radova iz različitih izvora kao i povezivanje citirajućih radova s citiranim radovima se može shvatiti kao pridruživanje bibliografskih zapisa određenim DOI vrijednostima. Dok je DOI često korišten i relativno kvalitetan jednoznačan identifikator i kod njega su zamijećene nekonzistentnosti. Nekoliko radova u časopisu *Scientometrics* imaju više dodijeljenih DOI vrijednosti kod kojih je izdavač koristio jedne, a WoS druge.

Rječnik WoS i Scopus klasifikacija za *vrstu rada*, je ujednačen. Budući da se radi o malom broju opisnih vrijednosti gdje različiti nazivi odgovaraju jedni drugima, bilo je dovoljno ujednačiti nekoliko vrijednosti s jedne strane kako bi odgovarale drugoj. U ovom slučaju, neke WoS vrijednosti za vrstu rada su prilagođene Scopus klasifikaciji.

Godina objave važan je podatak za sva scientometrijska istraživanja, a posebno za ovakva koja se bave praćenjem publikacija, autora i trendova kroz vrijeme. Osim osnovne provjere da je godina cijeli broj i da je prisutna, potrebno je provjeriti da je u odgovarajućem rasponu te da su individualne godine jednako pokrivene u svim izvorima. Radi praznih vrijednosti, ova provjera je provedena u fazi detaljne pripreme ulaznih podataka tj. nakon nadopunjavanja praznih vrijednosti.

Naslov časopisa je za skupove zapisa o radovima u časopisu *Scientometrics* provjeren radi sigurnosti da se uvijek radi o časopisu *Scientometrics* tj. radi sigurnosti da nije došlo do grešaka u pretraživanju tj preuzimanju podataka. Kod skupova citirajućih radova, provjereno je suprotno jer su svi radovi iz *Scientometrics*, zajedno s pripadajućim referencama, već prisutni u *u_sci_wos* tj. *u_sci_scopus*.

Polja **autori**, **reference** i **ključne riječi** razgraničena su za kasniju upotrebu. Za vrijednosti atributa **broj citata**, **godište** i **broj** utvrđeno je da sadrže samo dopuštene vrijednosti. Prazna vrijednost za **broj citata** je kod Scopus skupova pretvorena u vrijednost 0, budući da je ručnom provjerom ustanovljeno da prazna vrijednost za ovaj atribut kod Scopusa znači 0 prije nego nepoznatu informaciju. Što se podataka o broju tiče, s obzirom da su isti važni za kasnije spajanje radova među skupovima pažnja je posvećena svim prisutnim znakovima u ovom polju kako bi se popravili potencijalni teško uočljivi nedostaci poput korištenja različitih unicode crtica za dvobroje.

Nakon provedbe opisanih procesa, ulazne skupove podataka se smatralo pripremljenim za daljnju obradu, a pripremljeni skupovi su imenovani s prefiksom "p" te pohranjeni u JSON[7] formatu. JSON za razliku od razgraničenog teksta podržava različite ugniježđene strukture i vrste vrijednosti, ali je još uvijek jednostavan (i.e. direktno ljudski čitljiv) i interoperabilan tekstualan format što ga čini podobnim za ovu vrstu i količinu podataka.

2.2.3 Detaljna priprema podataka o radovima iz *Scientometricsa*

Nakon osnovne pripreme i karakterizacije ulaznih podataka, radovi iz časopisa *Scientometrics* su detaljno pripremljeni kroz nekoliko složenih procesa:

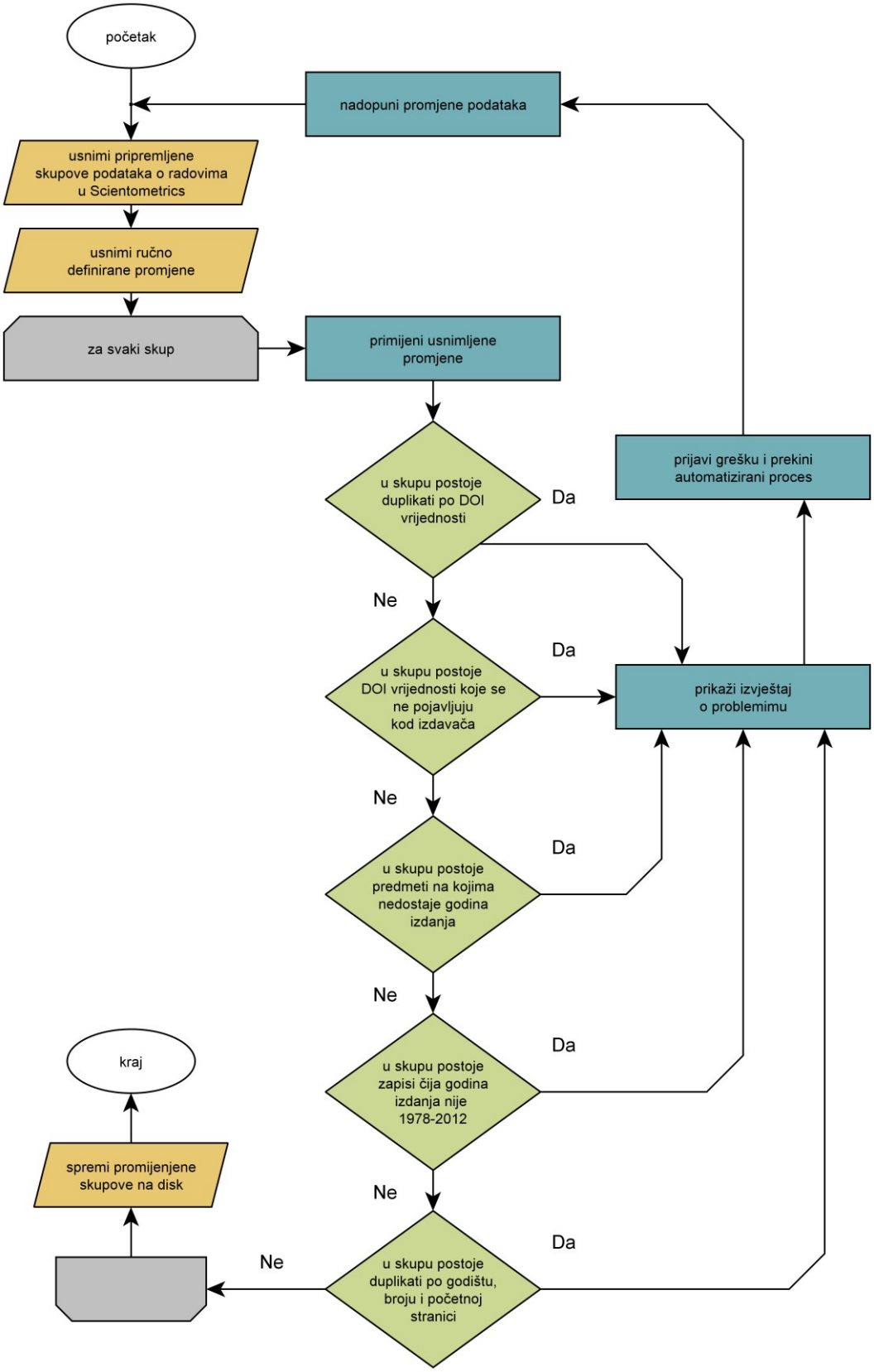
- validacija i ispravljanje ulaznih skupova na individualnoj razini i u odnosu na izdavačke metapodatke
- povezivanje podataka o radovima u *Scientometrics* iz različitih izvora (izdavač, WoS, Scopus)
- krosvalidacija i spajanje vrijednosti za pojedine radove
- ujednačavanje imena autora
- klasifikacija radova

2.2.3.1 Validacija i ispravljanje ulaznih skupova

U prvom koraku provjerene su značajke svakog skupa posebno kao i pojedinih WoS i Scopus skupova u odnosu na izdavačke metapodatke. Ispravljanje podataka se odvijalo prema dijagramu na slici 6, a proces je implementiran skriptom koja je prenesena u dodatku 1.

[7] http://www.json.org/

Slika 6. Provjera vrijednosti za različite skupove podataka o radovima iz časopisa *Scientometrics*



Kroz ovaj proces izrađena je tablica koja sadrži pobrojane nekonzistentnosti. Podaci su vidljivi u tablici 2. Uvid u ove brojeve je poslužio za odluku o strategiji spajanja podataka i za odabir strategije rješavanja problema navedenih u tablici.

Tablica 2. Nepravilnosti u ulaznim skupovima podataka

problem	izdavač	WoS	Scopus
duplikati u DOI vrijednosti	0	0	8
vrijednost "DOI" nedostaje	0	210	1100
vrijednost "broj časopisa" nedostaje	148	0	13
vrijednost "početna stranica" nedostaje	148	0	1
vrijednost "volumen" nedostaje	148	0	13
vrijednost "godina izdanja" nedostaje	148	0	0
DOI vrijednosti koje nisu prisutne kod izdavača	0	33	27
duplikati po volumenu, broju i prvoj stranici	3	6	3

Za nadopunu izdavačkih podataka korištene su DOI vrijednosti koje su preuzete putem Crossref usluga. Za ostale podatke rađen je ispis koji pruža uvid u nekonzistencije te iz kojeg su nakon pregledavanja preuzete potrebne ispravke. U slučaju nedoumica konzultiran je tekst radova koji je uvelike pomogao prilikom ispravaka te pružio uvid u neke zanimljive slučajeve. Prilikom ispravaka dana je prednost podacima izdavača odnosno predmetima kako su identificirani DOI-em i sabrani u PDF dokumente.

Nadopunom podataka je zaključeno da zapisi kod kojih nedostaje informacija o godini objave kod izdavača su predmeti koji su objavljeni *online* prije no što je izašao broj časopisa kojeg su formalno dio. WoS i Scopus bilježe godinu objave u tisku. Dok bi prema standardnim definicijama objave bilo primjerenije koristiti godinu objave na webu, pogotovo za potrebe citatne analize, u ovom istraživanju su ovi slučajevi ipak uklonjeni budući da se pojavljuju tek u zadnjoj godini promatranog razdoblja koja je uključena samo radi analize samocitata časopisa. Radi se naime o radovima objavljenima u tiskanoj inačici časopisa 2013. godine koji kao takvi nisu preuzeti iz WoS tj. Scopus citatnih indeksa budući da je godina izdanja prema citatnim indeksima izvan pretraživanog razdoblja (1978-2012). Ovo bi bio znatno veći problem da je prisutan veći broj slučajeva i u većem dijelu promatranog razdoblja pa je očekivan kao jedan od problema koji će buduća istraživanja morati prepoznati i rješavati.

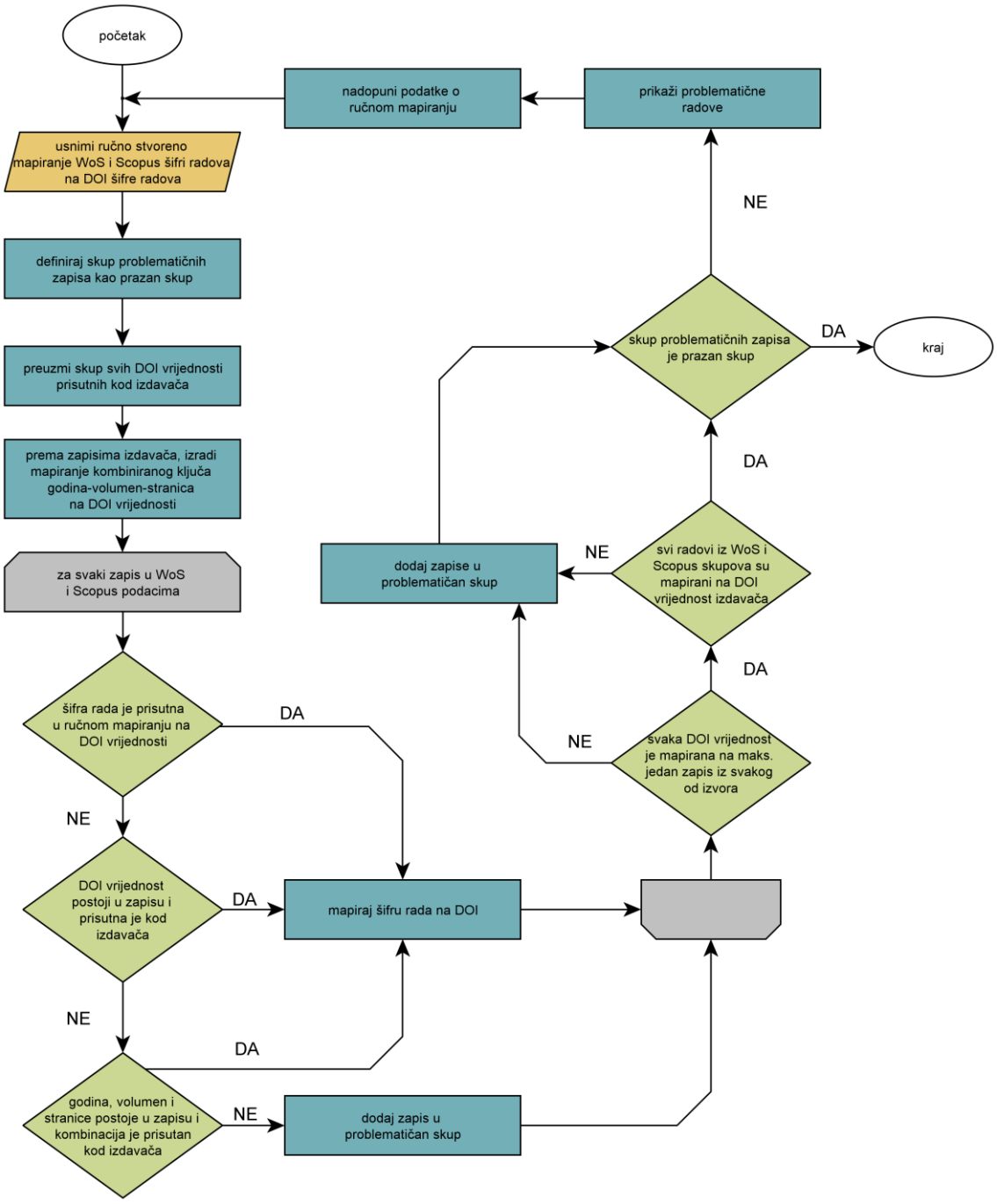
Drugi zanimljiv problem je bio što neki "radovi" sadrže više integralnih dijelova. Na primjer pismo i odgovor na to pismo, kratka priopćenja sabrana pod isti naslov ili recenzije knjiga sabrane u zajedničku sekciju. Kod WoS podataka ovakvi predmeti su katalogizirani kao više predmeta. U jednom ekstremnom primjeru, predmet "Vijesti" koji kod izdavača ima jedan DOI i dostupan je kao jedan PDF dokument u WoS-u je registriran kao čak 14 jedinica. Budući da se radi o predmetima koji su uglavnom bili recenzije knjiga i vijesti pa nisu relevantni za većinu analiza, isti su ispušteni kod WoS podataka kako ne bi unosili šum u kasnije spajanje radova. Informacija o takvim radovima odnosno razina granularnosti zapisa uključena je u analizu prema metapodacima izdavača.

Kod ostalih predmeta ispravljene su greške poput pogrešnih DOI vrijednosti, pogrešaka u pisanju broja (kod izdavača dvobroji su bilježeni samo kao prvi broj) i nekoliko dupliciranih predmeta. Sve DOI vrijednosti u WoS i Scopus skupovima zapisa pronađene su kod izdavača izuzev malog broja predmeta koji su imali grešku u DOI identifikatoru (poput nedostajućih znakova) i nekoliko predmeta koji su imali alternativnu DOI vrijednost u WoS podacima.

2.2.3.2 Mapiranje radova među skupovima

Nakon ispravaka pojedinih skupova izrađeno je mapiranje identifikatora radova među skupovima kako bi se utvrdilo koji zapisi opisuju iste radove. Upravo ovaj proces je glavni razlog detaljnim gore navedenim provjerama gore jer je šum u ovom koraku teže *ad-hoc* razlučivati i razrješavati. Također, u samom dizajnu procesa su korištene informacije iz prijašnjih koraka (npr. o prisutnosti vrijednosti) i isti pretpostavlja prethodne provjere jednoznačnosti korištenih identifikatora poput DOI vrijednosti i kombiniranih ključeva unutar svakog od skupova. Radovi su mapirani kao što je prikazano na slici 7.

Slika 7. Spajanje zapisa o radovima iz časopisa *Scientometrics*



Mapiranje radova je dalo uvid u nove greške svih skupova podataka. Prvi uvid je dobiven kroz nemogućnost pronalaska nekog rada u drugim skupovima podataka. Nakon ispravaka se ustanovilo da je slučaj o radovima koji radi grešaka u zapisu (npr. pogrešna godina uz nedostajuću DOI vrijednost) nisu pravilno upareni.

Nakon uparivanja dobiven je uvid u članke kod kojih se vrijednosti različitih polja ne slažu među različitim izvorima podataka. Kod nesuglasnosti u vrijednosti većine atributa poput godine, volumena, naslova i slično zaključeno je da se radi o manjem broju slučajeva pa je napravljen ispis svih onih kod kojih se postojeće vrijednosti za neki zapis razlikuju među izvorima (izuzev podataka o autorima, što je tretirano i opisano kao zasebna problematika). Veći broj grešaka je otkriven kod izdavača za atribut "broj časopisa" jer su kod izdavača dvobroji bilježeni samo prvim brojem. Budući da se radi o sustavnoj grešci, ista je ispravljena automatski. Nakon pregledavanja opisanih slučajeva nadopunjeni su podaci o ispravcima radova, a cijeli postupak pripreme podataka je automatski ponovljen kako bi se uzeli u obzir novi ispravci.

Budući da su svi zapisi iz WoS-a i Scopusa povezani sa zapisima izdavača, rezultat ovog procesa se može interpretirati kao mapiranje radova iz časopisa *Scientometrics* na DOI vrijednost koju koristi izdavač. Drugim riječima, u ovom slučaju WoS i Scopus nakon ispravaka, očekivano, ne uključuju radove koji nisu prisutni u metapodacima izdavača. Suprotno bi ukazalo na nekompletnost izdavačkih podataka.

Nakon mapiranja radova spojen je glavni skup o radovima u *Scientometricsu*. Za svaki zapis u tom skupu vrijednosti su dobivene na sljedeći način:

- **doi**, **godina**, **volumen**, **broj**, **početna stranica** i **završna stranica** preuzete su iz nakladničkih metapodataka, a nadopunjeni vrijednostima iz WoS odnosno Scopus podataka samo ako nedostaju u izdavačkim
- **wos_id** tj. **sco_id** su preuzete iz WoS tj. Scopus podataka, **broj citata** je preuzet iz oba skupa podataka i diferenciran jer se odnosi na različite podatke (i.e. ne na rad u scientometricsu nego na sve zabilježene slučajeve njegova citiranja)
- **naslov** je posebno obrađen za svaki skup (izbačeni su tzv. HTML entiteti i slično), a zatim je preuzet onaj s najviše znakova
- **vrsta predmeta** postoji samo za WoS i Scopus, kako se radi o atributu važnim za obradu koji sadržava mali broj različitih vrijednosti pregledane su svi slučajevi u kojima su obje vrijednosti poznate i različite, a dodane ukoliko su nepoznate
- **ustanove** su preuzete iz WoS podataka
- **reference** su izdvojene u poseban skup za kasniju obradu
- **autori** su preuzeti iz sva tri skupa kako bi pomogli u kasnijem razrješavanju imena autora

Podaci koji su dobiveni kao rezultat opisanih procesa predstavljaju početne spojene podatke o svim radovima i prilozima objavljenim u časopisu *Scientometrics* od 1978. do 2012. godine. Radovi do 2010. godine uključeni su kao predmet proučavanja, a radovi iz 2011. i 2012. godine su korišteni kao radovi koji vjerojatno citiraju radove u časopisu *Scientometrics* u promatranom razdoblju. Drugim riječima, ovi radovi su zadržani radi mogućnosti samocitata časopisa.

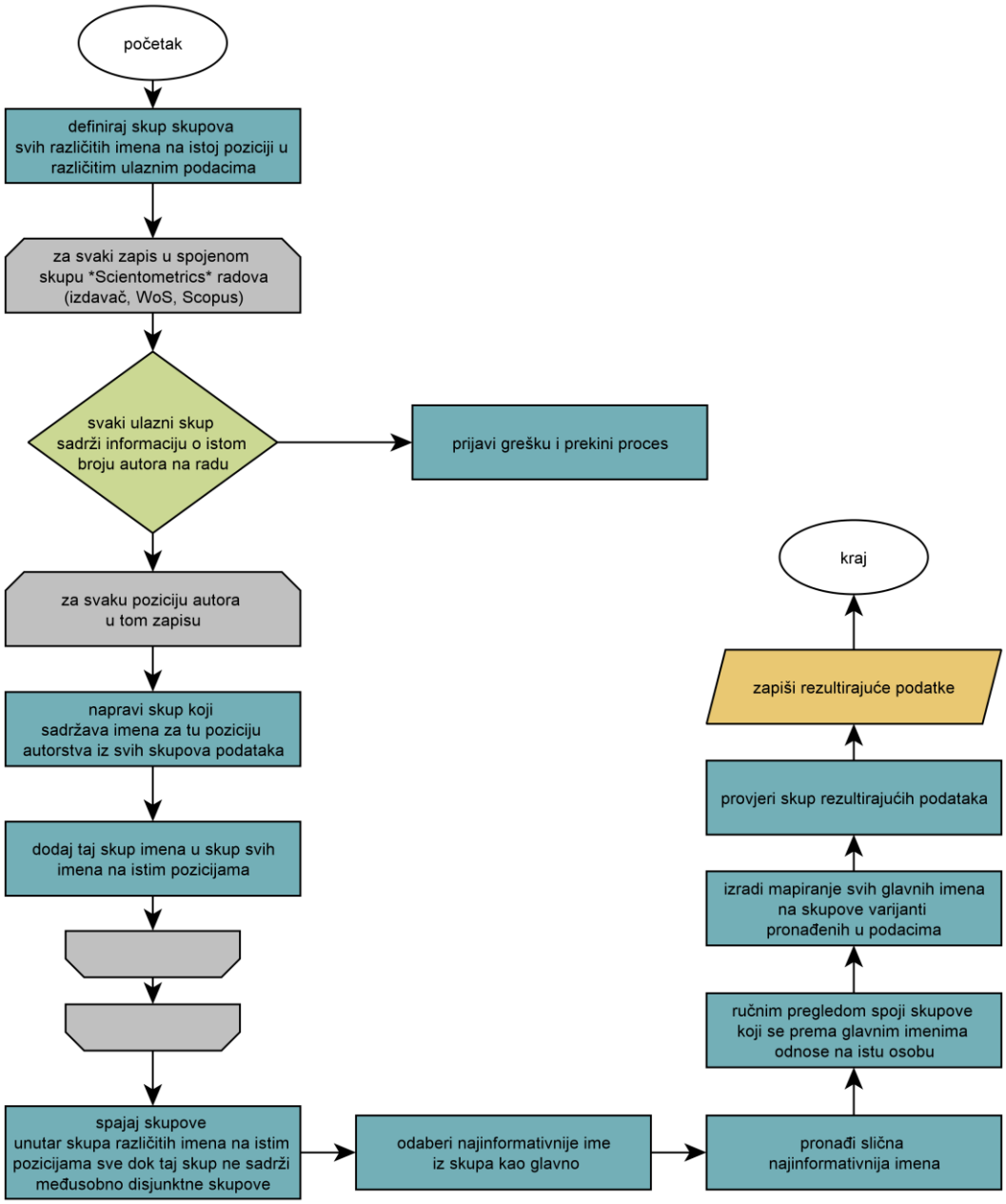
2.2.3.3 Ujednačavanje imena autora

Nakon spajanja ulaznih skupova o radovima u *Scientometrics*, bilo je nužno ujednačiti imena autora budući da ista nisu ujednačena unutar skupova, a pogotovo ne među njima. Jednoznačna detekcija autora od velikog je značaja u scientometrijskim istraživanjima jer se mnoga temelje na promatranju autora, na primjer kod promatranja suradnje i produktivnosti. Ona istraživanja kojima je primaran cilj promatranje neke skupine znanstvenika bi ovaj problem trebala rješavati prilikom izrade strategije pretraživanja tj. pretraživanjem i validacijom za svako zasebno ime (van Leeuwen, 2005). Ovakav pristup jamči preciznije podatke za promatrane autore, ali i dalje je prisutan problem sa svim imenima autora s kojima su promatrani autori surađivali, ali nisu unaprijed poznata jer se ne nalaze u podatkovnom skupu promatranih autora.

U eksplorativnom istraživanju nekog područja, autori često nisu unaprijed poznati. Dapače jedan od ciljeva je često detekcija ključnih autora. Drugim riječima, podaci su preuzeti putem nekog drugog kriterija, a osobe su detektirane putem imena prisutnim u preuzetim podacima. Ovaj pristup ima manju preciznost kada se promatraju individualni autori jer čak i kada je postotak greške za cijeli skup malen, greška može biti lokalizirana na nekoliko problematičnih autora što u slučaju individualne evaluativne bibliometrije može biti problem. U promatranju većih znanstvenih jedinica poput časopisa ili područja ova greška nije toliko značajna (van Leeuwen, 2005). Također, kontrolirano pretraživanje na razini individualnih autora zahtijeva daleko veći angažman istraživača prilikom preuzimanja podataka, što ga čini iskoristivim za manje skupove promatranih autora ili pak velike istraživačke projekte. Dodatan problem je što validacija jednoznačne identifikacije autora prilikom pretraživanja često zahtijeva poznavanje područja rada tog autora, što strategiju čini teže provedivom za istraživanja koja obuhvaćaju velik broj različitih područja.

U ovom istraživanju dodatna preciznost za detekciju autora putem osobnih imena je postignuta uparivanjem informacija o autorstvu iz tri različita skupa ulaznih metapodataka tj. iz izdavačkih, WoS i Scopus metapodataka. Polje autor je stvoreno tako da sadržava najdulje ime autora tj. ime koje ima veću šansu da bude jednoznačno ili omogući uvid da se iza nekih inicijala krije više različitih imena, ali nije gubilo informaciju o ostalim inačicama imena. Proces koji je odabirao najinformativnije ime autora prikupio je i sva imena kandidate koji nisu odabrani kako bi se iskoristile informacije dobivene mapiranjem više izvora. Ovaj proces je prikazan na slici 8.

Slika 8. Ujednačavanje imena autora



Rezultat procesa je mapiranje glavnih odabranih imena na sve varijante korištene u podacima. U ovom mapiranju provjereno je da:

1. su zastupljena sva imena autora koja se pojavljuju u izdavačkim, WoS i Scopus metapodacima
2. su sva glavna imena autora različita
3. su skupovi varijanti za sva glavna imena međusobno disjunktni

U prvoj fazi dobiveni su svi različiti skupovi imena autora gdje je svaki skup dobiven prikupljanjem poznatih imena autora u svim skupovima za neku poziciju u autorstvu rada. Podaci koje je ovaj proces proizveo izgledaju ovako:

```
...
{'Glanzel, W.', 'Glaknzel, W.', 'Glänzel, W.'}
{'Glanzel, W.', 'Glanze, W.', 'Glänzel, W.'}
{'Glanzel, W.', 'Glanzel, Wolfgang', 'Glänzel, Wolfgang'}
{'Glanzel, Wolfgang', 'Glänzel, Wolfgang'}
{'Glanzela, W.', 'Glänzela, Wolfgang', 'Glanzel, Wolfgang'}
{'Glaser, J.'}
{'Gleiser, S.', 'Gleiser, Sonia'}
{'Glenisson, P.', 'Glenisson, Patrick'}
{'Glänzel, Wolfgang'}
{'Gläser, Jochen', 'Glaser, J.'}
...
```

Dobiveni skupovi predstavljaju sve različite kombinacije izdavačkih, WoS i Scopus podataka za individualna autorstva. Minimalan ovakav skup će imati jedan element, ukoliko svi izvori podataka sadrže identično ime autora ili se vrijednost atributa "autor" pojavljuje u samo jednom od izvora podataka. Maksimalan skup će imati tri elementa ukoliko se vrijednost atributa "autor" pojavljuje u sva tri skupa i u sva tri je različita.

Ovi skupovi su zatim spojeni na sljedeći način: ukoliko presjek dvaju skupova iz popisa gore nije bio prazan skup u popis je dodana unija tih skupova, a ti skupovi su izbačeni iz popisa. Ovaj proces je rekurzivno ponavljan sve dok svi skupovi iz popisa nisu bili međusobno disjunktni. Rezultirajući popis je predstavljao sva poznata imena iz svih skupova grupiranih po supojavnosti na radovima. Primjer rezultata ovog procesa na skupovima navedenim gore vidljiv je iz sljedećeg popisa.

```
{'Gleiser, S.', 'Gleiser, Sonia'}
{'Glenisson, Patrick', 'Glenisson, P.'}
{'Glänzel, Wolfgang', 'Glanzela, W.', 'Glänzel, W.', 'Wolfgang, Glänzel', 'Glaknzel, W.',
 'Glanzel, Wolfgang', 'Glanze, W.', 'Glazel, W.', 'Glaenzel, Wolfgang', 'Glänzela, Wolfgang',
'Glanzel, W.'}
{'Gläser, Jochen', 'Glaser, J.'}

...

{'Bailón-Moreno, R.', 'Bailón-Moreno, Rafael', 'Bailon-Moreno, R.', 'Bailon-Moreno, Rafael'}
{'Moreno, Rafael', 'Bailon Moreno, Rafael', 'Moreno, RB.'}
```

Ovaj popis je dobra aproksimacija različitih autora, ali je osjetljiv na sljedeće slučajeve:

- međusobno disjunktni skupovi koji se odnose na istog autora (pojavljuju se kad različite inačice pisanja nekog autora ne dijele varijante među različitim radovima)
- autore s istim imenima i prezimenima, u rijetkim slučajevima i autore s istim prezimenima i inicijalima (kod autora koji se u svim skupovima pojavljuju samo s inicijalima)

Prvi slučaj je riješen na sljedeći način. Za svaki skup imena odabrano je ime s najvećim brojem znakova kao najinformativnije. Taj popis imena je grupiran po sličnosti, a grupe su zatim pregledane, a iz njih preuzete potrebne ispravke. Kao mjera dodatne sigurnosti izmjerene su i razlike među varijantama pojedinog autora, a pregledane one za koje su te razlike bile velike. Ovaj proces otkrio je nekoliko grešaka u poretku autora na ulaznim radovima koje su zatim ispravljene.

Što se drugog slučaja tiče provjereno je da ne postoji takav rad na kojemu se neko ime autora pojavljuje više puta u popisu autora tog rada. Ostaje naravno mogućnost tretiranja više autora kao jednog, ako nisu niti jednom surađivali na radovima u časopisu Scientometrics. Ovaj slučaj je gotovo nemoguće detektirati jer se radi o dva ili više autora istih imena i prezimena koja djeluju u istom uskom području. Informacije iz tekstova radova kao i informacije o ustanovama autora često ne pomažu puno jer tekst radova može sadržavati čak manje informacija nego bibliografski zapis (e.g. u starijim izdanjima časopisa *Scientometrics* autori su i na radove potpisani inicijalima, bibliografski zapisi mogu, pak, sadržavati puna imena tih autora).

Ipak, s obzirom da se radi o uskoj tematici i jednom časopisu možemo pretpostaviti da je broj ovakvih slučajeva malen, ako i postoje. Spomenuto postaje veći problem kako se šire podaci, e.g. na istraživanje širokih područja poput medicine ili velikog broja časopisa iz različitih područja. U širim podacima postoji i više informacija pa je moguće drugačije dizajnirati

proces. Na primjer putem detekcije interesa autora kroz časopise u kojima su objavljivali i zatim detekciju "istih" autora koji su objavljivali u drastično različitim područjima i ručnom provjerom istih.

Brojevi različitih imena ukupno i za svaki skup su se kroz ovaj proces kretali na sljedeći način:

- Izdavački, WoS i Scopus podaci su sadržavali 3 834, 3 997 odnosno 3 467 različitih imena autora. Broj elemenata odnosno različitih imena autora u uniji tih skupova bio je 9 003 što ukazuje na različite načine pisanja autora u svakom od skupova. Jedini ispravci provedeni na autorima do sad jesu izbacivanje viška praznog prostora i razgraničavanje Scopus autora zarezom budući da nisu tako razgraničeni u podacima dobivenim iz Scopusa.
- Nakon osnovnog čišćenja imena autora koje je uključivalo ujednačavanje veličine slova, razmaka unutar imena i interpunkcije ove brojke su pale na 3 793, 3 800 odnosno 3 443 različita imena za izdavačke, WoS i Scopus podatke, a broj imena u uniji skupova je iznosio 7 556.
- Nakon grupiranja po poziciji autora, broj različitih kombinacija imena, ne uzimajući u obzir redoslijed bio je 4 107.
- Broj različitih skupova imena dobiven spajanjem ovih 4 107 skupova u međusobno disjukntne skupove bio je 3 353.
- Nakon odabira najduljeg imena za svaki skup i zatim ujednačavanja po sličnosti sa odabranim imenima iz ostalih skupova, broj različitih skupova imena bio je 3 305 i taj broj je smatran brojem različitih autora u cijelom skupu podataka o radovima u časopisu *Scientometrics*.

Podaci pokazuju da je Scopus imao najujednačenija imena, ali je sadržavao i najmanje informacija (to jest samo inicijale imena) što povećava šanse imenjaka, mnogi od kojih su u ovom istraživanju ispravljeni kroz informaciju o punim imenima autora dobivenu iz drugih skupova podataka.

Na kraju procesa, izrađen je registar autora koji sadrži glavna imena autora, poveznice na članke putem DOI vrijednosti i sva imena koja se smatraju varijantama tog autora. Varijante imena su sačuvane za kasniju reviziju ili ponovno korištenje. Korištena su npr. za detekciju samocitata autora kod radova koji su citirali *Scientometrics*, ali nisu u njemu objavljeni.

Na temelju ovog popisa izrađen je ujednačen atribut "autor". U svakom zapisu, svako ime u atributima "autor" za svaki od ulaznih skupova zamijenjeno je glavnim imenom prema prikupljenim varijantama. Ukoliko je ovaj proces rezultirao istim poljem "autor" za sva tri skupa, dobivena vrijednost je korištena kao razriješeno ime autora. U suprotnom, prikazana je informacija o autorstvu iz različitih skupova podataka, a greška je ispravljena ručno. Radilo se o nekoliko radova koji su u jednom od skupova imali drugačiji redoslijed autora. Nakon tih ispravaka nije postojao niti jedan rad kod kojih je razrješavanje polja autor proizvodilo različite vrijednosti za različite izvore podataka.

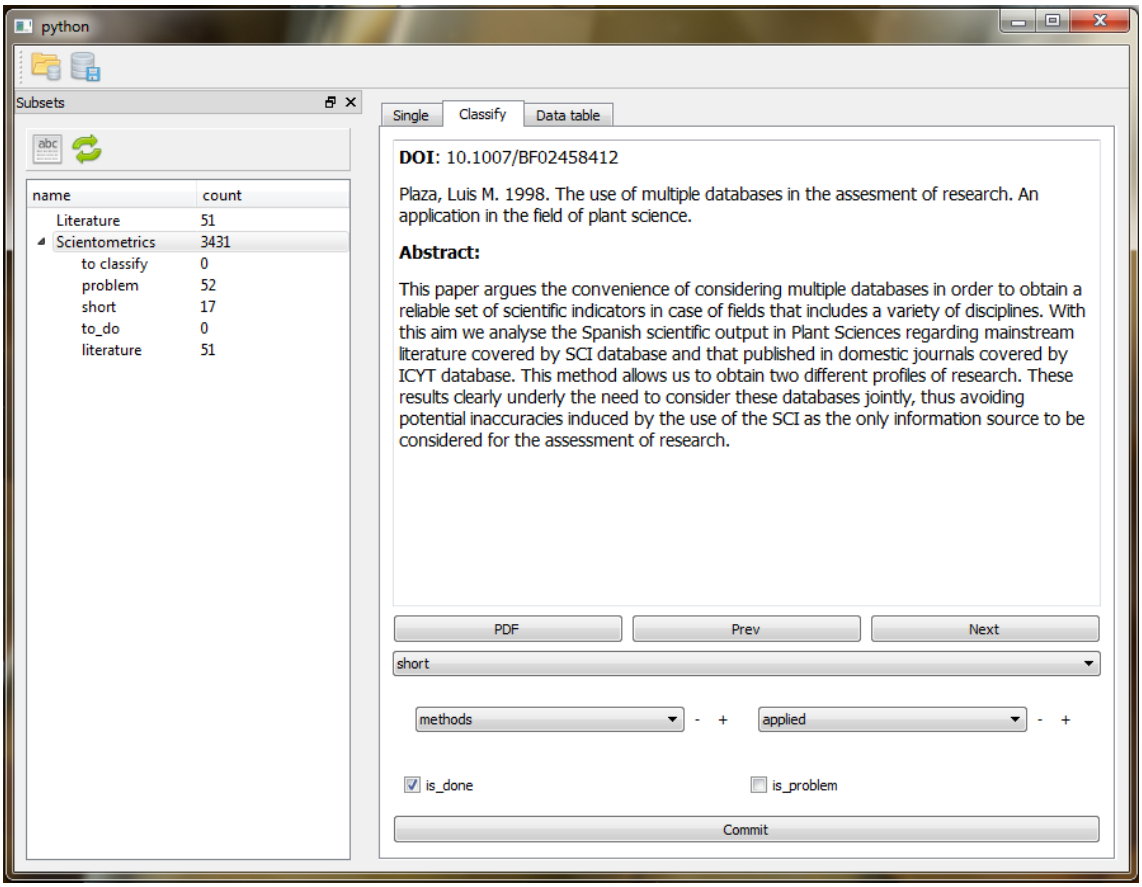
2.2.3.4 Klasifikacija radova po vrstama radova i priloga i tematici članaka

Uz razrješavanje autora, bilo je prijeko potrebno ustanoviti vrste rada jer se iste koriste za kasniju filtraciju radova i odlučivanje o primjerenosti različitih analitičkih postupaka. Na primjer, člancima je dana najveća pažnja dok su novosti i recenzije knjiga isključene iz analiza o autorstvu, suradnji i citiranosti. Kako bi isto bilo moguće bila je potrebna ovakva klasifikacija radova i priloga objavljenih u časopisu *Scientometrics*. Dok WoS i Scopus sadrže podatke o vrsti radova i autor ovog rada ih se nadao iskoristiti za potrebe istraživanja, provjera je pokazala da iako se radi o označivanju s malim brojem naoko jasnih kategorija poput "Article", "Letter" ili "Book Review", ovi se podaci uvelike razlikuju između WoS-a i Scopusa. Provjera tekstova radova je pokazala da se ne može pridodati veća važnost jednoj od klasifikacija te da katkad i obje griješe.

Iz navedenih razloga pristupilo se polu-automatskoj klasifikaciji koja je jednostavnim heuristikama pretpostavljala vrstu rada u odnosu na broj stranica, prisutnost i broj referenci, WoS i Scopus klasifikaciju te riječi iz naslova. Pretpostavke nisu međutim automatski primijenjene već su pružene na uvid uz prikaz osnovnih podataka i lagan pristup tekstu rada. Potom su odobrene ručno što je osiguralo preciznost i efikasnost procesa, ali i brz razvoj *ad hoc* kôda kao svrsishodnu pomoć, a ne kao zaseban alat šire namjene što bi zahtijevalo znatno više vremena za razvoj. Još jedan razlog ručnom pregledavanju radova bila je dodatna klasifikacija prema orijentaciji rada koja je zahtijevala ručna pregledavanja kao što je opisano niže.

U svrhu lakšeg pregledavanja i konzistentne dodijele klasifikacijskih oznaka, ali i izbjegavanja prikaza informacija koje bi utjecale na klasifikaciju prema osnovnoj tematici, poput podataka citiranosti razvijeno je grafičko sučelje prikazano na slici 9.

Slika 9. Prikaz sučelja razvijenog za efikasnu klasifikaciju radova i priloga u časopisu *Scientometrics*



Budući da se kôd za ovaj alat odnosi gotovo isključivo na programiranje grafičkog sučelja, a ne na podatkovnu problematiku, isti nije naveden u dodacima. Sučelje je omogućilo jednostavna grupiranja i filtriranja radova i priloga kroz unos Python funkcija te brz pristup tekstovima radova. Upravo je pristup tekstovima radova riješio mnoge nedoumice te omogućio uvid u zanimljive slučajeve.

Za označavanje vrsti radova odabrane su sljedeće klase:

- *članci* - u širem smislu riječi, bilo da se radi o originalno znanstvenom ili preglednom; radovi kojima je cilj znanstveni doprinos kroz otkriće, problematizaciju ili sintezu
- *kratak članak* - svi radovi koji prenose primaran sadržaj, a u časopisu su označeni kao "kratak rad" ili "kratka komunikacija"
- *korespondencija* - uključujući pisma, komentare, rasprave i ostale radove koji su namijenjeni izravnoj komunikaciji među autorima i koji su često vezani uz neki specifičan tekst objavljen u istom časopisu

- *tekstovi uredništva* - tekstovi koji nemaju znanstveni doprinos kao primaran cilj već oslovljavaju zajednicu poput uvodnika, tekstova za posebne brojeve ili za posebne zgode i slično
- *novosti* - radovi koji prenose vijesti poput onih o konferencijama ili o dobitnicima nagrada
- *podatkovni izvještaji* - radovi koji prenose ekstenzivne podatke uz malo objašnjenja tj. kojima je svrha prijenos podataka
- *ispravke* - radovi koji prenose ispravke za ranije objavljene članke
- *recenzije knjiga*
- *bibliografije*

Svaki rad svrstan je u samo jednu od gore navedenih skupina. Za potrebe većine analitičkih postupaka promatran je samo skup članaka tj. onih radova koji prenose primaran sadržaj časopisa. Za potrebe proučavanja trendova objavljivanja i citiranja, pogotovo pojedinih znanstvenika, upravo ovi radovi predstavljaju ono što se proučavanjem literature aproksimira tj. objavljeno znanstveno znanje. Kod drugih vrsta radova često se radi o potpuno drugim vrstama sadržaja ili autorstva. Na primjer, vijesti i uredničke tekstove često potpisuju suradnici na objavljivanju časopisa. U scientometrijskim istraživanjima ne možemo uspoređivati kao istu vrstu doprinosa produktivnost dva autora od kojih svaki ima više desetaka radova, ali prvi ima više desetaka članaka, a drugi priloga poput rubrike vijesti i niti jedan članak.

U grupu članaka za potrebe svih analiza za koje je smisleno da promatraju samo članke su naknadno pridodani i kratki članci. Tokom ručne klasifikacija radova i priloga utvrdilo se da je riječ o radovima koji prenose istu vrstu sadržaja i čija "kratkoća" nije toliko jednoznačna budući da postoje kratki članci dulji od "punih" članaka odnosno onih koji u časopisu nisu označeni kao kratki članci. Radi navedenog ti radovi su uključeni u iste analize kao i članci. S obzirom da je udio ovih radova bio vrlo malen odlučeno je da ih nema dovoljno kako bi bili informativni kao zasebna skupina pa se kod analiza nije radila distinkcija između "kratkih članaka" i "članaka".

Uz klasifikaciju po vrstama publikacija koja je temeljna za bibliometrijsku obradu, provedena je i klasifikacija u odnosu na orijentaciju rada na teorijski ili metodološki napredak ili na prikaz rezultata primijenjenih istraživanja. Ovakvu podjelu je zanimljivo za istražiti kada se

promatra dulji vremenski period koji dobrim dijelom odgovara ukupnom periodu razvoja neke discipline. Za ove potrebe definirane su tri kategorije u koje su članci razvrstani:

- **teorijski članci** - svi članci kojima je cilj pojasniti, problematizirati ili sistematizirati scientometrijske postavke ili samu scientometriju kao disciplinu
- **metodološki članci** - svi članci kojima je cilj doprinijeti operacionalizaciji budućih scientometrijskih istraživanja, bilo kroz nove postupke, sintezu starih, strategije preuzimanja podataka i slično
- **primijenjeni članci** - svi članci koji prenose rezultate primijenjenih istraživanja poput istraživanja nekog specifičnog područja, časopisa, produktivnosti zemalja i ustanova i slično

Svakom članku dodijeljene su sve kategorije koje su u njemu zastupljene. Kategorija je dodijeljena ukoliko je procijenjeno da članak podržava taj sadržaj i ako se maknu sve ostale komponenete. Na primjer, članku koji prenosi rezultate primijenjenog istraživanja **i** opisuje korišten inovativan metodološki proces, dodijeljena je kategorija primijenjenog članka ukoliko se procijenilo da kad se iz njega preuzmu samo dijelovi o rezultatima primijenjenog istraživanja da oni stoje kao cjelina za sebe. Isto vrijedi za metodološku komponentu. Kao primarni izvori informacija za dodjeljivanje kategorija korišteni su sažetak i puni tekst rada.

Uz samo dodjeljivanje ovih kategorija one su i rangirane prema naslovu i sažetku tj. tekstu članka u slučaju nedoumica. Neki članak je mogao biti primijenjeno-metodološki ili metodološko-primijenjeni ovisno na koji aspekt su autori stavili veće težište. U velikom broju slučajeva, autori su sami u sažetku rangirali ciljeve izrazima poput "Primaran cilj ovog rada je ... ." ili "Dodatno, rad prikazuje ... .". S obzirom na navedeno, članci se po ovim informacijama mogu grupirati na više načina:

- prema bilo kojoj kategoriji koja je na radu zastupljena (u tri preklapajuće skupine),
- po primarnoj kategoriji (u tri skupine koje se ne preklapaju),
- i po svim dobivenim kombinacijama kategorija (u maksimalno 15 skupina koje se ne preklapaju).

S obzirom na relativnu kompleksnost podatkovnih prikaza često je korištena podjela po primarnoj orijentaciji rada, a gdje je su ove kategorije bile posebno zanimljive prikazani su i podaci po kombinacijama kategorija.

Radi obima posla, kao i kod drugih procesa pripreme podataka poput ujednačavanja imanena autora, rezultirajući skup podataka, odnosno mapiranje klasa na DOI vrijednosti, je posebno sačuvan radi revizije i ponovnog korištenja.

2.2.4 Priprema citirajućih radova i citata na radove u časopisu *Scientometrics*

Nakon pripreme podataka o radovima u časopisu *Scientometrics* izvršena je priprema skupa podataka o radovima koji citiraju radove u časopisu *Scientometrics* tj. skupa citirajućih radova.

Citirajući radovi prikupljeni su zasebno iz baza WoS i Scopus. Budući da WoS podaci pokrivaju cijelo promatrano razdoblje, na njima su se vršile glavne citatne analize. Scopus podaci o citiranosti radova u časopisu *Scientometrics* mogu se smatrati relevantnim tek od 1996. godine pa nadalje tj. za polovicu promatranog razdoblja. Kao što je u Uvodu opisano, Scopus selektivno indeksira i radove iz ranijih perioda, ali tek se pregledavanjem podataka u časopisu *Scientometrics* utvrdilo da je to kod njega slučaj za samo nekoliko izoliranih članaka što je onemogućilo usporedbu i agregaciju podataka o citiranosti iz WoS-a i Scopusa. S obzirom na glavnu temu rada, prednost je dana podacima koji su relevantni za cijelo promatrano razdoblje (odnosno WoS podacima), a ne podacima koji zahvaćaju širi okvir citiranosti u kraćem razdoblju (odnosno agregaciji WoS i Scopus podataka). Agregacija WoS i Scopus podataka predstavlja potencijalnu nadopunu citatne slike o člancima u časopisu *Scientometrics*, ali radi opsega rada ovakve analize su ostavljene za buduća istraživanja. U analizu citiranosti uključeni su dakle samo WoS podaci i to sve vrste radova i priloga u kojima WoS prati citirane radove tj. razrješava reference.

Podaci su pripremljeni na sličan način kao i podaci o radovima u časopisu *Scientometrics* sa sljedećim razlikama:

- radi se o pripremi jednog ulaznog skupa podataka; informacije nisu preuzimane iz spajanja skupova radova kao kod podataka o radovima i prilozima u časopisu *Scientometrics*
- reference radova su uparene s radovima u časopisu *Scientometrics*; u ovaj postupak uključeni su i svi radovi i prilozi u časopisu *Scientometrics*

- autori su ispravljeni na temelju postojećih ispravaka iz pripreme podataka o radovima u *Scientometrics* i zatim na temelju sličnosti nizova unutar skupa citirajućih radova

Svaki citirajući rad tj. svaka jedinica u ovom skupu podataka citirao je jedan ili više radova u časopisu *Scientometrics*. Kako bi se utvrdilo koliko i koje radove u časopisu *Scientometrics* citiraju citirajući radovi, obrađena je svaka referenca citirajućih radova. Kod WoS podataka reference su kodirane na sljedeći način:

```
Jensen LB, 2011, INT J PHARMACEUT, V416, P410, DOI 10.1016/j.ijpharm.2011.03.015
Berlow EL, 2004, J ANIM ECOL, V73, P585, DOI 10.1111/j.0021-8790.2004.00833.x
AINLEY WM, 1993, PLANT MOL BIOL, V22, P13, DOI 10.1007/BF00038992
LIMA R.A., 2007, PERSPECTIVAS CIENCIA, V12, P50
Bonaccorsi A, 2007, RES EVALUAT, V16, P66, DOI 10.3152/095820207X218141
ALBRECHTSEN H, 1992, THESIS ROYAL SCH LIB
Bjorneborn L, 2001, SCIENTOMETRICS, V50, P65, DOI 10.1023/A:1005642218907
Kurtz M. J., 2007, ASTROPHYSICAL J, V709
Selzer J., 1993, UNDERSTANDING SCI PR
Luque-Martinez T, 2005, CITIES, V22, P411, DOI 10.1016/j.cities.2005.07.008
```

WoS reference ne pružaju puno podataka za identifikaciju radova koji nisu objavljeni u časopisima, ali kod radova u časopisima često sadrže DOI vrijednost. Za potrebe istraživanja preuzete su sve reference koje citiraju časopis *Scientometrics* u razdoblju 1978-2010. Odabran je period do 2010. godine kako bi objavljeni radovi imali vremena dobiti citate, odnosno kako bi se osigurala relevantnost podataka o citiranosti. Budući da je skup objavljenih radova poznat iz pripremljenog skupa o radovima u *Scientometrics*, omogućena je validacija prepoznatih citata. Strategija spajanja ovakvih nizova tekla je na sljedeći način:

- preuzmi sve različite reference iz rada
- preuzmi reference u kojima se spominje riječ "SCIENTOMETRICS"
- izbaci reference u kojima je godina > 2010
- ako referenca ima DOI, pokušaj spojiti na DOI vrijednost
- ako se godište:stranica iz reference pojavljuju u scientometrijskim podacima točno jednom, spoji na DOI vrijednost
- ako se prvo prezime autora:godina pojavljuju u scientometrijskim podacima točno jednom, spoji na DOI vrijednost
- izdvoji za ručno povezivanje

Na 6 641 citirajućih radova pronađeno je 291 529 referenci i to 172 320 različitih vrijednosti. Među različitim vrijednostima pronađeno je 3 144 vrijednosti koje sadrže niz "SCIENTOMETRICS". Iz ovih je uklonjena 151 referenca s godinom nakon 2010. te 54

reference koje se ne odnose na časopis *Scientometrics*. Među preostalima, 2 764 reference su gore opisanom strategijom povezane s DOI brojem automatskim putem, a ostatak je povezan ručno. Ukupno 2 939 referenci je povezano na 2 280 različitih DOI vrijednosti radi varijanti u kodiranju referenci. Niti jedna referenca nije povezana na više DOI vrijednosti.

Posebno zanimljivi kod ovih podataka su podaci o tome koliko različitih citiranih radova je prisutno prema WoS kodifikaciji, a koliko nakon uparivanja na DOI vrijednosti. WoS sadrži 2 939 različitih nizova tipa "Kurtz M. J., 2007, ASTROPHYSICAL J, V709" za radove u časopisu *Scientometrics*, a nakon mapiranja utvrđeno je da se radi o 2 280 različitih radova. S obzirom da se radi o prestižnom citatnom indeksu koji ne objavljuje metode pripreme podataka, ovakvi podaci su zabrinjavajući za procjenu preciznosti tog citatnog indeksa. Ukoliko postoji više vrijednosti poput "Kurtz M. J., 2007, ASTROPHYSICAL J, V709" za neki rad indeksiran u WoS-u, koji su od njih brojani u ukupan broj citata prijavljen za taj rad? Odgovor na ovakva pitanja, međutim, ne samo da zahtijeva opširne prikaze podataka na primjerenijem uzorku već bi trebao biti i transparentan u opisu citatnog indeksa naročito ukoliko se taj citatni indeks koristi za procjenu rada znanstvenika. Za potrebe ovog istraživanja, podacima o citiranosti je podignuta preciznost mapiranjem na provjerene DOI vrijednosti radova.

Uz povezivanje citirajućih radova sa citiranim radovima, autori svih 6 641 radova su ujednačeni kroz varijante dobivene u procesu spajanja podataka o radovima iz *Scientometrics*. Ovaj proces je ujednačio inačice imena autora koji su objavili i u časopisu *Scientometrics* i omogućio detekciju samocitata autora.

2.3 Priprema tekstova radova za analizu

Kao što je ranije opisano, scientometrijska metoda se uvelike zasniva na bibliometrijskoj obradi znanstvenih tekstova. Bibliometrijska metoda, shodno imenu, se pak zasniva na bibliografskim zapisima u svrhu operacionalizacije istraživanja. Za kvantitativno promatranje nekog korpusa znanstvenih publikacija potrebno je izraditi kakve podatkovne surogate za same publikacije pa su standardizirani bibliografski zapisi od velike važnosti kao ulazni podaci.

Ipak, od polovice prošloga stoljeća, računalna obrada tekstova značajno napreduje uz razvoj informacijskih tehnologija tj. mogućnosti i paradigmi koje iz njih proizlaze te razvojem

posebnih područja poput računalne lingvistike. S obzirom na napredak u mogućnostima procesiranja prirodnog jezika te računalne analize teksta općenito, kao i sve dostupnije alate, scientometrija dolazi do mogućnosti uključivanja samih predmeta promatranja tj. znanstvenih publikacija u kvantitativne analize. Ova mogućnost se, međutim, još uvijek rijetko koristi.

Razlozi su vjerojatno višestruki:

- Bibliografski metapodaci su sami po sebi kompleksni tj. s velikim brojem vezanih entiteta i velikim brojem atributa te i sami po sebi uključuju neka tekstualna polja poput naslova i sažetka. Već obrada bibilografskih metapodaka, kao što je prethodno poglavlje prikazalo, zahtijeva puno truda i specifična znanja.
- Znanstveni radovi su tekstovi s kompleksnom strukturom sadržaja (npr. dva stupca teksta s tablicama koje zauzimaju oba stupca) i posebnim značajkama teksta (npr. citiranje).
- Cjeloviti tekstovi znanstvenih radova su često dostupni samo u PDF formatu. PDF je namijenjen prezentaciji i prenosivosti bez gubitaka oblikovanja, a ne strukturiranju sadržaja što, pogotovo u kombinaciji s kompleksnom strukturom radova, predstavlja teškoće u pripremi tekstova za obradu. Navedeno je predstavljalo problem i u ovom istraživanju, a detaljnije je opisano niže u tekstu.
- Prisutan je i problem diseminacije samog korpusa kad se bazira na tekstovima dostupnim u komercijalnom pristupu.

2.3.1 Ekstrakcija iz PDF formata

Jedan od temeljnih problema uključivanja znanstvenih tekstova u istraživanja je priprema tekstova iz PDF formata što, ovisno o standardima korištenim pri izradi tih PDF datoteka, može biti znatan problem. U sklopu ovog rada testiran je velik broj alata za ekstrakciju teksta iz ovog uzorka PDF datoteka (i.e. radovi iz časopisa Scientometrics). U velikom broju slučajeva, rezultati nisu bili ni približno iskoristivi. Nakon nezadovoljavajućih rezultata dostupnih uporabom specijaliziranih alata, do rješenja se pokušalo doći zaobilaznim putem, kroz rasterizaciju u visoku rezoluciju pa zatim OCR rezultata. Razlog za ovaj pokušaj bio je što je struktura stranice često zbunjivala ekstraktore teksta, a ozbiljan OCR softver posjeduje mogućnost analize strukture i to više iz izgleda dokumenta nego loše strukturiranog PDF sadržaja.

Drugi razlog je bila dostupnost alata za OCR visoke preciznosti Tesseract. Preliminarni testovi ovog alata pokazali su bolju kvalitetu izlaznog teksta nego slobodno dostupni specijalizirani alati. Razlog tome najvjerojatnije leži u specifičnostima PDF dokumenata časopisa Scientometrics koji se kroz vrijeme koristi različitim rješenjima za izradu PDF dokumenata poput PdfGeanie i Jaws PDF creator.

Nakon testiranja OCR pristupa ekstrakciji dokumenata zaključeno je da se na ovaj način propuštaju neki detalji koji su važni za detekciju citata u tekstu. Kao što je već spomenuto, važna značajka znanstvenih tekstova je kultura citiranja. Iz perspektive scientometrije, upravo nam je ova značajka znanstvenih tekstova posebno zanimljiva. Citatni indeksi kodiraju citate bez obzira na njihovo korištenje u tekstu, a upravo je kontekst njihova korištenja u tekstu među glavnim kritikama citatnim analizama. Također, unutar citatni navodi unutar teksta su posebni dodatni dijelovi rečenica i korisno ih je izdvojiti kao takve prije daljnje jezične obrade.

Problem kodifikacije citatnih navoda kroz tekst je problem njihovog prepoznavanja u tekstu dokumenta. Pri ovome, treba obratiti pažnju da se citatni stilovi bitno razlikuju po lakoći prepoznavanja u tekstu. Na primjer, numerirani stil u uglatim zagradama lakše je raspoznati nego autor-datum stil navođenja. Što se časopisa Scientometrics tiče, s obzirom na njegovu tematiku pomalo je bilo iznenađujuće otkriće da časopis Scientometrics koristi veći broj stilova navođenja literature. U nastavku su prenesene slike teksta kako bi se prikazali razni stilovi korišteni u časopisu.

Slika 10. Stilovi navođenja literature korišteni u člancima u časopisu *Scientometrics*

This question has either been overlooked[15] or varying solutions have been offered.[16–19] This paper attempts to clarify the state of methodology in the field. It

---

problems, while the latter is normally associated with questions having theoretical implications.[5,11]

Although the validity or usefulness of this dichotomy has been challenged by numerous authors (e.g., *Amick,*[1] *Levins,*[12] *Price*[14]), several studies have demonstrated

---

version of 2001) [1, 2] . The CSTP decided the Governmental Policy Evaluation Act (2002, revised on 2005) [3, 4] and the Law on the General Rules of Incorporated

---

circulations of library materials see also *Burrell* (1985, 1986, 1987, 1990). For technical details of the construction see, for instance, *Parzen* (1962) or *Ross* (1996). Here we just

---

GRANOVETTER [1973, 1974], WELLMAN & BERKOWITZ [1991], and OTTE & ROUSSEAU [2002] view social networks as a way of representing social structures in terms of sets of members and sets of ties depicting relationships. This view implies that a social network is essentially a tool to study social structure [MARSDEN, 1987, 1990, 2003]. It describes social structure in terms of ties or relationships and interprets behavior in light of differing status and location within that structure [MARSDEN, 1990; WELLMAN & BERKOWITZ, 1991].

Dodatan problem je što se mnoge od ovih informacija gube prilikom ekstrakcije iz PDF formata u običan tekstualni format. Na primjer, "superskript" postaje normalan tekst čime se gubi informacija tj. superskript počinje izgledati kao korištenje broja unutar rečenice. Reference pisane u kurzivu ili smanjenim velikim slovima također gube distinkciju koju su vizualno posjedovali.

Nakon dodatnog testiranja i imajući na umu potrebu distinkcije ovakvih informacija, za ekstrakciju podataka iz PDF datoteka odabran je PDFlib TET (Text Extraction Toolkit). TET je profesionalan komercijalan alat za pouzdanu ekstrakciju teksta iz pdf datoteka. Uz razne naprednih značajku poput detekcije strukture stranice i tablica, ovaj alat je odabran iz dva glavna razloga:

1. Od svih korištenih alata, TET je pokazao preciznost ekstrakcije teksta kojoj se približio jedino *Tesseract*
2. Uz samu ekstrakciju "sirovog texta", TET je u mogućnosti proizvesti XML bazirani TETML format koji sadrži dodatne informacije o strukturi teksta te o oblikovanju znakova

Uz navedeno, TET nudi API kroz više programskih jezika među kojima je i Python, što je ovaj alat povezalo s pristupom implementaciji metodoloških postupaka. Budući da Python uključuje standardan modul za rad s XML datotekama, nije bilo problema s korištenjem TET-a ni za ekstrakciju ni za kasniji programski pristup rezultatima ekstrakcije.

Sam TETML format može opisivati preuzeti tekst na više razina. Za potrebe ovog istraživanja preuzeta je najdetaljnija razina, tj tekst je opisan do razine znaka. Stvaran primjer preuzet iz jednog od *Scientometrics* radova za tekst "impact factor[1,3,6, 10] is" u TETML formatu izgleda ovako:

```
<Word>
 <Text>impact</Text>
 <Box llx="155.60" lly="397.80" urx="182.60" ury="407.76">
  <Glyph font="F3" size="9.96">i</Glyph>
  <Glyph font="F3" size="9.96">m</Glyph>
  <Glyph font="F3" size="9.96">p</Glyph>
  <Glyph font="F3" size="9.96">a</Glyph>
  <Glyph font="F3" size="9.96">c</Glyph>
  <Glyph font="F3" size="9.96">t</Glyph>
 </Box>
</Word>
<Word>
 <Text>factor</Text>
 <Box llx="187.42" lly="397.80" urx="210.64" ury="407.76">
  <Glyph font="F3" size="9.96">f</Glyph>
  <Glyph font="F3" size="9.96">a</Glyph>
  <Glyph font="F3" size="9.96">c</Glyph>
  <Glyph font="F3" size="9.96">t</Glyph>
  <Glyph font="F3" size="9.96">o</Glyph>
  <Glyph font="F3" size="9.96">r</Glyph>
 </Box>
</Word>
<Word>
 <Text>1,3,6,10</Text>
 <Box llx="210.68" lly="400.84" urx="237.15" ury="408.88">
  <Glyph font="F3" size="8.04" sup="true">1</Glyph>
  <Glyph font="F3" size="8.04" sup="true">,</Glyph>
  <Glyph font="F3" size="8.04" sup="true">3</Glyph>
  <Glyph font="F3" size="8.04" sup="true">,</Glyph>
  <Glyph font="F3" size="8.04" sup="true">6</Glyph>
  <Glyph font="F3" size="8.04" sup="true">,</Glyph>
  <Glyph font="F3" size="8.04" sup="true">1</Glyph>
  <Glyph font="F3" size="8.04" sup="true">0</Glyph>
 </Box>
</Word>
<Word>
 <Text>is</Text>
 <Box llx="242.00" lly="397.80" urx="248.64" ury="407.76">
  <Glyph font="F3" size="9.96">i</Glyph>
  <Glyph font="F3" size="9.96">s</Glyph>
 </Box>
</Word>
```

U ovom primjeru značajka citatnih navoda je da su pisani u "superskriptu" te da su cijeli brojevi koji su potencijalno razdvojeni zarezima ili spojeni u raspone crticom. Drugim riječima, "1", "1-3" i "1,2, 5-7" su validni navodi dok oni koji uključuju ostale znakove nisu. Dodatno, korištenje ovakvog stila implicira i stil navoda u popisu literature. Ukoliko su navodi u popisu literature u istom dokumentu precizno razdvojeni, moguće je validirati i brojke koje se pojavljuju u navodima kroz tekst. Dakle, ukoliko je pronađeno 12 navoda u popisu literature, navodi u tekstu, nakon razrješenja nabrajanja i raspona, moraju biti svi brojevi od 1 do 12. Dok ovaj pristup dodatnom validacijom povećava preciznost postupka i ukazuje na moguće greške u postupku preuzimanja informacija iz teksta isti je osjetljiv na greške koje su prisutne u dokumentu tj. na one koje nisu nastale pogreškama u preuzimanju iz teksta.

Situacija kod prezime-godina stila navođenja je nešto teža. Sami navodi su sličniji normalnom tekstu rečenica, a često su i sami njegov sastavni dio. Budući da su u časopisu Scientometrics, kao što je često slučaj, navodi posebno vizualno istaknuti, strategija identifikacije navoda je u ovom slučaju kombinirala informacije o fontu i oblikovanju znakova u tekstu s metodama prepoznavanja dijelova teksta putem strukture, poput regularnih izraza.

2.3.2 Priprema tekstova radova

Cijeli proces, od PDF datoteke do pripremljenog teksta zahtijevao je mnoge postupke. S obzirom na kompleksnost postupaka, procesi su prikazani u obliku popisa. Krovni procesi u pripremi mogu se podijeliti na sljedeći način.

1. preuzimanje teksta iz PDF datoteka
2. karakterizacija i razrješavanje dijelova teksta
3. razrješavanje referenci
4. jezična priprema teksta

U nastavku slijedi detaljniji opis svakog procesa kroz popis postupaka koji je svaki od njih uključivao.

Preuzimanje teksta iz PDF datoteka

Kao prvi korak procesa pripreme tekstova radova preuzete su TETML datoteke koje su zatim razriješene. U ovom koraku preuzet je "sirov" tekst radova, a iz njega su izlučeni strukturalno problematični elementi i oni elementi koji ovise o stranicama. Nakon ovog koraka tekst je tek doveden do forme u kojoj se njime može računalno baratati. Koraci uključeni u ovaj proces su:

1. proizvedi TETML iz PDF datoteke
2. prikaži različite TETML elemente objektno,
    - klase odabrane za prikaz različitih razina teksta u ovom trenutku bile su Znak->Riječ->Paragraf->Stranica->Dokument
    - svaka razina funkcionira kao filtar prijašnje
    - izbacuju se prazni i pogrešno strukturirani elementi te svi nepoznati znakovi tj. oni koje nije bilo moguće interpretirati prema zadanoj kodnoj stranici
    - sve izbačeno je dostupno za pregled i provjeru kasnije
3. razrješi posebne elemente dokumenata poput sadržaja tablica i slika
4. Pročisti sve elemente ovisne o stranici, na primjer zaglavlje, brojevi stranica, posebne informacije na prvoj stranici ...

Karakterizacija i razrješavanje dijelova teksta

Nakon ekstrakcije iz PDF-a, u ovom koraku se smanjuje kompleksnost programske reprezentacije teksta te se sadržaj dokumenta dodatno ispravlja i označava. Proces uključuje sljedeće korake:

1. preuzmi informacije iz prošle faze i ispusti nepotrebne elemente, to jest "odmakni se" od TETML-a i PDF-a; reprezentiraj tekst kao Dio -> Dokument, gdje je "dio" (u ovom smislu se na engleskom često koristi izraz *chunk*) neki niz znakova koji reprezentira više riječi s poznatim granicama riječi te dodatnim informacija o fontu znakova
2. napravi izvještaj o korištenju fontova u svim tekstovima
3. razdvoji dijelove koji sadržavaju različite elemente (e.g. naslov i dio paragrafa u istom chunku)
4. spoji dijelove koji su prelomljeni usred rečenice (e.g. spoji paragrafe koji su prelomljeni stranicama)

5. karakteriziraj dijelove prema sadržaju i izgledu: detektiraj naslove sekcija te naslove tablica i figura
6. podijeli tekst na sljedeće dijelove:
    1. početne dodatne informacije, odnosno naslov, abstract, autori ...
    2. tekst rada, odnosno od uvoda do referenci
    3. reference, odnosno bibliografija na kraju rada
    4. dodaci

Razrješavanje referenci

Nakon osnovne pripreme teksta, može se pristupiti ekstrakciji referenci u tekstu radova prema sljedećim koracima:

1. pripremi sve reference u popisu referenci
2. koristeći se kombinacijom strukturalne analize sadržaja (e.g. regularni izrazi) i informacija o fontovima pronađi sve navode unutar teksta
3. međusobno validiraj sve navode iz teksta s navodima iz popisa referenci, prijavi greške na uvid
4. izdvoji sve reference iz teksta kao posebne dijelove izuzev slučaja kad je ime autora dio rečenice

Jezična priprema teksta

Nakon što je tekst pripremljen i što su iz njega uklonjeni posebni dijelovi poput tijela tablica i citatnih navoda u rečenicama, može se provesti jezična analiza na glavnom tekstu rada.

1. usitni dijelove iz "paragrafa" u rečenice
2. karakteriziraj svaku rečenicu koristeći Stanford CoreNLP koji uključuje tokenizaciju, POS označavanje, NER i koreferenciranje

S obzirom na kompleksnost cijelog opisanog postupka i opseg ovog istraživanja, informacije dobivene kroz jezičnu pripremu teksta namijenjene su kao preliminaran test ovih postupaka. Ručnim provjerama manjeg broja tekstova utvrđeno je kako ovi postupci pružaju relativno visoku preciznost i bez posebnog treniranja Stanford CoreNLP softvera. Ipak, korištenje ovih podataka je ostavljeno za buduća istraživanja radi vremenske zahtjevnosti dodatnih provjera.

2.4 Struktura istraživanja i korišteni alati

Gotovo svi postupci pripreme i analize podataka implementirani su programskim jezikom Python. Ulazni resursi za ovo istraživanje su postavljeni na sljedeći način:

- **podaci**: preuzeti i izvedeni podaci; ulazi i izlazi za skripte
- **alati**: općeniti alati (i.e. python funkcije) koje se koriste u kasnijim postupcima
- **promjene**: sve promjene provedene nad podacima, vanjsko bilježenje promjena omogućava provjeru i upravljanje istima te jasan uvid u sve radnje provedene u obradi podataka
- **skripte**: implementacija postupaka opisanih i provedenih u sklopu ovog rada; koriste alate i promjene
- **tekst**: tekst ovog rada u ReStructuredText[8] formatu koji je automatski povezan na one izlaze skripti koji su namijenjeni za prikaz podataka kao i na sam kôd alata, promjena i skripti

Alati i skripte su u potpunosti prenesene u Dodatku 1. Promjene su podatkovni skup koji je stvoren ispravljanjem nekonzistentnosti. S obzirom da se radi o "sirovim" podacima, isti nisu preneseni.

Glavni razlog opisanom pristupu je nedostupnost kvalitetnih rješenja za obradu, posebno pripremu, ove vrste podataka. Također, upitno je do koje mjere općenita rješenja bazirana na grafičkim sučeljima uopće mogu biti razvijena s obzirom na potrebnu versatilnost pri obradi različitih skupova ove vrste podataka.

Posljedica i prednost korištenja ovakvog pristupa je potpuna ponovljivost procesa iz ulaznih podataka[9] te analiza, ponovne provjere i izbacivanja / dodavanja svih promjena od trenutka preuzimanja podataka nadalje. Naravno, ova tvrdnja stoji samo ako nikakve ručne promjene nisu unesene direktno u podatke nakon njihova preuzimanja, što se u ovom radu poštivalo izuzev ispravka nekoliko strukturalnih grešaka koje su priječile usnimavanje podataka.

Korišteni alati koji nisu razvijeni u sklopu ovog doktorata tj. o kojima ovo istraživanje ovisi su:

[8] http://docutils.sourceforge.net/rst.html
[9] Ipak, ulazne podatke vjerojatno nije moguće ponovno preuzeti iz baza jer se baze mijenjaju, ali cijeli proces se može ponoviti na "zamrznutom" stanju prikupljenih podataka

Python 3.3 (i Pythonova standardna zbirka modula)

Dodatni Python moduli:

- igraph (za rad s grafovima tj. mrežama)
- matplotlib (za izradu grafikona)
- numpy (je pružio neke osnovne statističke funkcije)
- docutils (za formatiranje teksta rada i povezivanje s podatkovnim izlazima)

Ostali alati:

- PdfLib TET (za ekstrakciju teksta iz PDF formata)
- Stanford CoreNLP (za jezičnu pripremu preuzetog teksta)
- Microsoft Office Word (za prijelom finalnog teksta doktorata po stranicama)

U nastavku prvo se opisuje preuzimanje i detaljna priprema bibliografskih metapodataka kao i podataka iz PDF dokumenata koji su sadržavali tekstove radova, a zatim postupci korišteni pri analizi pripremljenih podataka.

2.5 Analiza

U scientometrijske svrhe, shodno imenu, primarno se koristi kvantitativna metodologija u kojoj centralnu poziciju ima bibliometrijska analiza znanstvenih publikacija. U objavljenim scientometrijskim istraživanjima vidljivo je da se u osnovi radi o istraživanjima radova u znanstvenim časopisima. Glavni razlog tome je što je članak osnovni nositelj znanstvene komunikacije u područjima prirodnih znanosti, tehnologije i medicine. Prvi citatni indeks znanstvenih tekstova, SCI, je nastao upravo za potrebe literature takozvanih STM područja (eng. *science, technology, medicine*) i temelji se na indeksiranju članaka, a kasniji citatni indeksi, SSCI i A&HCI, nastaju po uzoru na SCI. Dodatan razlog usmjerenosti na radove u časopisima su dakle dostupni izvori podataka. Citatni indeksi koji uključuju i knjige javljaju se tek nedavno i to s nejasnom pokrivenošću (Google Scholar) ili ograničenog opsega (Book Citation Index kao dio WoS-a).

Analitički pristup korišten u ovom doktoratu može se opisati kao bibliometrijska analiza publikacija mjerena analizom mreža tj. analizom podataka na grafovima gdje god je to bilo svrsishodno. Dodatno, istraživanje uključuje eksperimentalno povezivanje bibliometrijskih

metoda s korpusom tekstova znanstvenih radova. Stoga je metodološki pristup analizi podataka bio kroz spomenute tri cjeline: bibliometrijsku analizu, analizu mreža i uključivanje tekstova radova. Naglasak u rezultatima i većini postupaka je na bibliometrijskim analizama (kroz analize mreža). U smislu analiza tekstova radova, već rad na pripremi korpusa ovakvih tekstova predstavlja važan pomak u metodološkim mogućnostima. Nažalost, zbog zahtjevnosti posla, pogotovo s obzirom na značajke PDF datoteka časopisa *Scientometrics*, detaljnija analiza i prikaz rezultata koji bi se koristio širom paletom podataka dobivenih iz tekstova radova, ostavljena je za buduće istraživanje.

Bibliometijska analiza znanosti temelji se na pogledu na znanost kroz njene publikacije. Ulazna jedinica bibliometrijske obrade je bibliografski zapis. Dok se ovi podaci u scientometrijskim istraživanjima publikacija često povezuju s vanjskim podacima poput podataka o znanstvenicima i ustanovama radi nadopune i/ili definicije ulaznog skupa bibliografskih zapisa, upravo su bibliografski zapisi oni koji ih povezuju u cjelinu i glavna su prizma kroz koju se promatra.

Provedene analize odnosno pokazatelji na kojima se analiza temeljila, podijeljene su na sljedeći način:

1. **bibliometrijski pregled** - osnovni bibliometrijski pregled radova i priloga objavljenih u časopisu *Scientometrics*
2. **autorstvo i produktivnost**
    - *autorstvo radova* - analiza radova u odnosu na broj autora
    - *produktivnost autora* - analiza autora u odnosu na broj radova
    - *suradnja* - proučavanje ko-autorstva među autorima i zemljama njihovih ustanova
3. **citatne analize**
    - *radovi u časopisu Scientometrics kao citirani radovi*
        - analiza radova objavljenih u časopisu *Scientometrics* u odnosu na citate koje su primili
        - analiza radova koji pružaju citate prema časopisu *Scientometrics* i časopisa u kojima su ti radovi objavljeni
        - analiza autora prema citatima koje su dobili na radove objavljene u časopisu *Scientometrics*
    - *radovi u časopisu Scientometrics kao citirajući radovi*

- starost citirane literature
- citirani časopisi u časopisu *Scientometrics*

U tekstu koji slijedi najprije su opisani procesi i koncepti centralni svim pokazateljima, a zatim su pobliže opisane kategorije iz prethodno navedenog popisa.

2.5.1 Prebrojavanje

S obzirom na bibliometrijsku srž i proučavanje znanstvene literature kao fokusa većine istraživanja, ulazni podaci za scientometrijska istraživanja su bibliografski zapisi o znanstvenoj literaturi, a kvantificiraju se prebrojavanjem. Iako pojam brojanje na prvi pogled zvuči više nego jednostavno (van Raan, 2005), u scientometrijskim istraživanjima je jedan od temeljnih indikatora koji je vrlo kompleksan i kojemu ovise ishodi.

"Broj radova" je tako centralan konstrukt koji prebrojavanjem u određenom okviru daje podlogu scientometrijskim pokazateljima. Na primjer, možemo razlikovati broj radova nekog entiteta (e.g. autora, časopisa, ustanove, zemlje) koji ukazuju na produktivnost tog entiteta te broj radova koji su citirali radove tog entiteta koji ukazuju na odjek tih radova.

Dok prebrojavanje radova može djelovati jednostavno, a dobivene brojke visoko precizne i relativno lako interpretabilne, treba uzeti u obzir:

- korištene metode brojanja
- problem računalnog prebrojavanja tekstovnih nizova
- bibliografske distinkcije različitih "radova"

Gauffriau i ostali (2007) definiraju tri vrste brojanja oko kojih postoji konsenzus: cijelo brojanje, frakcionalno brojanje i brojanje prvih. Kod cijelog brojanja svaki entitet na radu dobiva cijeli bod za rad. Na primjer, za rad koji ima tri autora, svakom od autora se bilježi jedan rad. Kod frakcionalnog brojanja, svi entiteti dijele bod. Kod brojanja prvih cijeli bod se dodjeljuje prvom entitetu. Dok ove tri vrste prebrojavanja ugrubo opisuju glavne pristupe, u praksi se često koriste specifični izračuni s razlikama u nijansama. Na primjer, kod frakcionalnog brojanja svakom autoru se može dodijeliti isti udio u radu ili se pak može uzeti i redoslijed autora u obzir pa prvi autor dobiva najveći udio. Metode se mogu i kombinirati, na primjer za radove s manjim brojem autora može se koristiti cijelo brojanje, a za radove s većim brojem autora frakcionalno brojanje.

U svakom slučaju, odabir metode mora odgovarati predmetu istraživanja. Na primjer, za potrebe istraživanja suradnje preporučuje se cijelo brojanje jer se suradnji ne pridodaje značaj radi "primarnosti" autora na radu. Drugim riječima, cilj je utvrđivanje mreže suradnje među znanstvenicima, a ne utvrđivanje značaja pojedinog autora za nastanak nekog individualnog rada. U ovom istraživanju, s obzirom da se radi o proučavanju u svrhu opisa područja s naglaskom na razvoj i da se na radovima u najvećem broju slučajeva nalazi "normalan" broj autora (kao što je opisano kasnije), korišteno je cijelo brojanje za sve analize.

S druge strane, ponekad podaci uvjetuju metodu. Na primjer, u opisu metodologije o pripremi citata prikazano je kako WoS kodira reference. U ovom istraživanju te reference su mapirane na pune bibliografske zapise što je omogućilo cijelo brojanje citata po autorima prema informacijama iz ujednačenih bibliografskih zapisa o radovima objavljenim u časopisu *Scientometrics*. U WoS bibliografskim zapisima (kakvim im se može pristupiti) vrijednosti atributa "reference" sadrže podatak samo o prvom autoru. U ovom istraživanju, da citati iz svih radova koji su citirali časopis *Scientometrics* ne vode na pune bibliografske zapise koji su također uključeni u istraživanje, bilo bi moguće koristiti samo "brojanje prvih" prilikom brojanja citata po citiranim autorima jer ne bi postojala informacija o ostalim autorima. Također, nakon povezivanja WoS citatnih navoda na razrješen skup radova informacija o broju citiranih radova se smanjila s 2 939 citirana rada u časopisu *Scientometrics* na 2 280, odnosno za 22,4%, što nije zanemariv podatak. Početna brojka je nešto veća od *svih* radova i priloga objavljenih u časopisu *Scientometrics*, uključujući i priloge poput eratuma ili novosti. Razlika u ovim brojevima prikazuje važnost pripreme podataka i dovodi nas do idućeg problema.

Drugi problem je vezan uz operacionalizaciju prebrojavanja pomoću računala. S obzirom da se uglavnom radi o prebrojavanju tekstovnih nizova koji su prikupljeni raznim procesima i nisu strogo kontrolirani potrebno je biti pažljiv u pripremi podataka. Nizove se mora ujednačiti do puno veće mjere no što je to potrebno za ručno pregledavanje, što je opisano u prethodnom poglavlju. Također, javljaju se problemi sinonimije i homonimije, koji su relativno često prisutni u zapisima o radovima. Neki od njih su nastali uslijed automatiziranog baratanja tekstom poput automatskog skraćivanja imena autora i to često algoritmima prilagođenima drugim jezicima. Kao primjer za potonje mogu poslužiti imena hrvatskih autora i autorica kod kojih prvo prezime često završava kao drugo ime ukoliko se ne piše s crticom.

Suptilne razlike, poput prisutnosti dijakritika, u računalnoj obradi često stvaraju sinonime. Također, isti nizovi mogu označavati različite entitete s ograničenim mogućnostima provjere ovisno o istraživanju i podacima. Na primjer, u istraživanju nekog skupa autora koje sadrži vanjske informacije o njima, poput registra znanstvenika, biti će lakše razriješiti imena promatranih autora nego u istraživanju koje ne sadrži dodatne informacije o autorima uz one koje su dostupne kroz imena s publikacija.

Manji, ali svejedno prisutan, problem je granularnost bibliografske obrade. Jednostavno rečeno, za neki skup bibliografskih zapisa potrebno je odgovoriti na pitanje: Da li je razina obrade, i.e. "jedan rad", prihvatljiva i da li se slaže s pretpostavkama istraživača? Dok kod većine radova u znanstvenim časopisima nema problema s jednostavnom identifikacijom oko toga što je *jedan* rad, postoje slučajevi u kojima može doći do različitih interpretacija. Ekstreman primjer je "rad" iz časopisa *Scientometrics* naslovljen "News"[10] koji ima jedan DOI i diseminira se kao jedan PDF dokument. Radi se o sabranim kratkim vijestima, svaka od kojih je paragraf teksta odvojen od ostalih s "***". U WoS-u je ovaj dokument predstavljen s 11 zapisa od kojih svaki odgovara zasebnoj vijesti. Kada istraživanje krene uključivati i druge vrste publikacija dobivenih iz izvora općenitijeg tipa, na primjer nacionalnih bibliografija ili kataloga nacionalnih knjižnica, problem postaje naglašeniji. Provjera, na primjer, razine obrade knjiga kod kojih je autorstvo po poglavljima, uvelike utječe na dizajn istraživanja (Jokić et al., 2012).

Opseg u kojem se broji je određen izvorom tj. izvorima ulaznih podataka. Većina bibliografskih baza ima jasan opseg u smislu da je poznato koje publikacije su uključene. Međutim, čak i kada se baza temelji na potpuno jasnom opsegu, poput nekog skupa časopisa, teško je ustanoviti relevantnost baze za pojedinog autora ili ustanovu ukoliko nisu prikupljeni i podaci o ostalim radovima koje je autor tj. ustanova objavila. Primjerenost baze će varirati ovisno o području istraživanja kao i drugim čimbenicima poput moguće lokalne orijentacije znanstveno-istraživačkog rada. Slično tome, relevantnost citatnih podataka za nekog autora, ustanovu ili područje ovisi o odnosu trendova citiranja i nekog izvora citatnih podataka. Na primjer, broj citata koji je izmjeren na literaturi iz časopisa pruža precizniju reprezentaciju stvarnog stanja za područja u kojima se glavnina citata odnosi na radove u časopisima, za razliku od područja u kojemu to nije slučaj.

[10] DOI: 10.1007/BF02097179

2.5.2 Grupiranje

Kao što je već više puta rečeno, temeljna ulazna jedinica za veliku većinu scientometrijskih istraživanja je publikacija. Publikacija je jedinica koja povezuje aktere, područje i primarne zapise znanja, koja se može kvantificirati prebrojavanjem i čije značajke se mogu kvantitativno promatrati u odnosu na skupove radova. Operacionalizacija promatranja nekog aktera ili drugog entiteta putem radova provodi se, dakle, primarno grupiranjem radova.

Drugim riječima, dok je temeljna ulazna jedinica rad, čest je slučaj da je fokus istraživanja neki akter vezan za rad poput znanstvenika ili znanstvenih ustanova. Kao primjer može poslužiti bibliometrijska metodologija proučavanja obrazaca objavljivanja znanstvenika. U ovom smislu, znanstvenike promatramo u ulozi autora. U podatkovnom smislu, autora možemo za potrebe bibliometrijskih istraživanja shvatiti kao grupaciju radova. Scientometrijski pokazatelji skupa radova nekog autora su scientometrijski pokazatelji o tom autoru. Kao primjeri mogu poslužiti broj radova, broj citata, *h*-indeks (koji je opisan kasnije u tekstu) i slično. Dodatni atributi autora se mogu koristiti za predviđanje i/ili interpretaciju ovih podataka, poput spola ili godine rođenja, ali u većini scientometrijskih istraživanja ovi su nadopuna (tj. dodatni kriteriji grupiranja) scientometrijskim indikatorima proizašlim iz radova. Isto vrijedi i za ostale promatrane aktere, za vremenska razdoblja, vrste radova ili bilo koju vrijednost koja grupira radove.

S obzirom na značajke bibliografskih metapodataka, postoji velik broj mogućih grupacija ulaznih podataka o publikacijama. Na temelju zastupljenih vrijednosti u podacima moguće je grupirati podatke na načine koji zadovoljava fokus istraživanja, na primjer:

- autor
    - ustanova
        - zemlja
- časopis
- godina
- vrsta rada
- tematska klasifikacija

Uz same vrijednosti atributa, zapise je moguće grupirati i po značajkama ili agregacijama vrijednosti atributa, na primjer po vremenskim razdobljima ili po broju autora na radu (e.g. jednoautorski, dvoautorski ...). Dobivena, često hijerarhijska, grupiranja te agregacija vrijednosti atributa radova i vezanih entiteta, podloga su kvantitativnim analitičkim postupcima koji zadovoljavaju potrebe istraživanja. Tzv. evaluativna bibliometrija se često usmjerava na autore i ustanove. Za potrebe ovog istraživanja koriste se podaci dobiveni grupiranjem po svim navedenim atributima kako bi se omogućio dovoljno širok uvid za potrebe opisa centralnog korpusa scientometrijske literature.

Što se odnosa među entitetima koji se ovako grupiraju tiče, u terminima relacijskog modela podataka (ER modela), odnos između autora i radova je više prema više što je slučaj i kod mnogih drugih entiteta vezanih uz scientometrijske radove. Grupiranje radova po autorima i svim vezanim atributima proizvodi međusobno preklapajuće skupine. Na primjer, iz informacije da su autori A i B bili svaki autorima 2 rada. ne možemo znati ukupan broj radova u skupu radova koje su napisali A ili B. Ukoliko se koristilo cijelo brojanje, tih radova je 2-4. Drugim riječima, brojevi radova različitih autora ne smiju se zbrajati. Navedeno može u kompleksnijim hijerarhijskim grupacijama i analizi koautorstva prouzročiti probleme, a u slučaju nepažnje i greške, ukoliko se pažljivo ne operacionalizira. U ovom istraživanju, kako bi se izbjeglo navedeno, ovakvi podaci za potrebe izračuna pokazatelja nisu reprezentirani brojem već skupovima šifri.

Na primjer, kao temeljna podloga kasnijih izračuna pokazatelja o autorima je za svakog autora stvoren skup šifri radova e.g. a:{1, 2} i b:{2, 3}. Kardinalnost nekog skupa je broj radova, a skup sadržava dovoljno informacija za povezivanje na ostale atribute radova kao i za spajanje grupa i detekciju supojavnosti.

Na primjer, ako su a i b skupovi šifri radova i ako koristimo cijelo brojanje, skupovi nam otvaraju mogućnosti poput:

- ukupan broj radova autora a jest n(a)
- ukupan broj radova koji su napisali a ili b jest n(a∪b)
- ukupan broj radova na kojem su surađivali a i b jest n(a∩b)

Opisan pristup efikasno rješava probleme u radu s podacima agregiranim po autorima ili po drugim atributima, a istovremeno omogućava i jednostavno proučavanje koautorstva tj.

supojavnosti vrijednosti. Ovaj problem je važan i u interpretaciji i korištenju podataka o produktivnosti autora. Na primjer, ako je 10 autora napisalo 10 jednoautorskih radova, prosječan broj radova po autoru je 1. Ako je pak na svaki od radova potpisano svih 10 autora, tada je prosječan broj radova po autoru 10. Prilikom pregleda podataka, ukoliko su predočene samo informacije o brojevima radova po autoru, moglo bi se zaključiti da su autori u potonjem slučaju 10 puta produktivniji, iako je u oba slučaja isti broj autora odgovoran za isti broj članaka. Da smo pak koristili frakcionalno brojanje, dobili bismo drugačiju informaciju ovisno o točnoj strategiji podjele "boda" za jednu publikaciju. U slučaju da je svaki autor dobivao jednak udio na radu koji iznosi 1/broj autora, dobili bismo isti prosječan broj radova po autoru u oba skupa. Navedeno prikazuje važnost omogućavanja uvida u više vrsta podataka. Kod korištenog primjera te informacije su vezane uz autorstvo promatranih radova.

2.5.3 Supojavnost i mapiranje znanosti

Matrice supojavnosti (eng. *coocurrence*), poput matrica kocitiranosti, pružaju korisne podatke za mapiranje i razumijevanje struktura koje su prisutne u ulaznom skupu dokumenata (Leydesdorff i Vaughan, 2006).

U bibliometrijskom podatkovnom smislu, supojavnost je moguće mjeriti na bilo kojem atributu koji je u više-prema-više odnosu prema radovima ili koji se zasniva na tekstualnim informacijama. Slijede primjeri zajedno s objašnjenjima što se kroz njih proučava:

- supojavnost imena znanstvenika u polju "autor" - koautorstvo i suradnja
- supojavnost naziva ustanova ili zemalja u polju "ustanove" - međuinstitutcijska ili međunarodna suradnja
- supojavnost radova u popisu referenci - bibliografsko uparivanje tj. kocitiranost, ovisno o smjeru promatranja
- supojavnost riječi u naslovu, sažetku ili popisu - razvoj i povezanost tema
- supojavnost tematske specijalizacije autora u nekom časopisu - interdisciplinarnost časopisa

S obzirom na značajke podataka, korištenje informacija o supojavnosti vrijednosti čest je mehanizam u scientometrijskim istraživanjima. Očit primjer supojavnosti je koautorstvo, ali u scientometrijskim istraživanjima možemo proučavati i supojavnost citiranih ili citirajućih radova te riječi u različitim dijelovima zapisa ili dokumenta. O koautorstvu će biti više govora

u zasebnom poglavlju budući da se u ovom istraživanja koristi analiza mreža kao primaran pristup promatranju autora i autorstva u časopisu *Scientometrics* u razdoblju 1978-2010.

U načelu, teme koje se u scientometriji proučavaju na temelju supojavnosti vrijednosti u podacima o radovima su mapiranje znanosti te suradnja među akterima. Mapiranje znanosti nastoji otkriti strukturu i dinamiku znanosti koristeći atribute znanstvene komunikacije, posebice znanstvenih publikacija (van den Besselaar i Heimeriks, 2006). Sredinom 1960-ih, De Solla Price raspravlja o strukturalnim svojstvima časopisa, publikacija, autorstva i citata i time uvodi dinamično mapiranje znanosti pomoću informacija iz SCI tj. citatnih indeksa kao važnu temu u proučavanju znanosti kroz njene publikacije (Leydesdorff, 1987). Radi ovakvog shvaćanja mogućnosti koje nude podaci iz citatnih indeksa, važna ideja u scientometriji je da citatni indeksi reprezentiraju multidimenzionalne prostore (e.g. časopisa) koji odgovaraju disciplinama i specijalizacijama (Leydesdorff, 1987).

U smislu mapiranja znanosti najčešće se govori u smislu mapiranja putem citata ili ko-citata što je u ovom istraživanju obrađeno u sklopu citatnih analiza kako su objašnjene u metodologiji i prikazane u rezultatima.

Mapiranje znanosti često se koristi za istraživanje interdisciplinarnosti i multidisciplinarnosti. Postoji više različitih scientometrijskih pristupa mjerenju interdisciplinarnosti. Većina pristupa uzima članke (ili patente) kao predmete istraživanja te mjeri interdisciplinarnost u terminima zajedničkog pojavljivanja svega onog što se može definirati kao specifično za disciplinu – ključnih riječi, tematskih kategorija, interesa autora ili citata (Schummer, 2004). Osnovna postavka je da zajedničko pojavljivanje elemenata specifičnih za disciplinu na neki način otkriva snagu veza ili razmjene između srodnih disciplina.

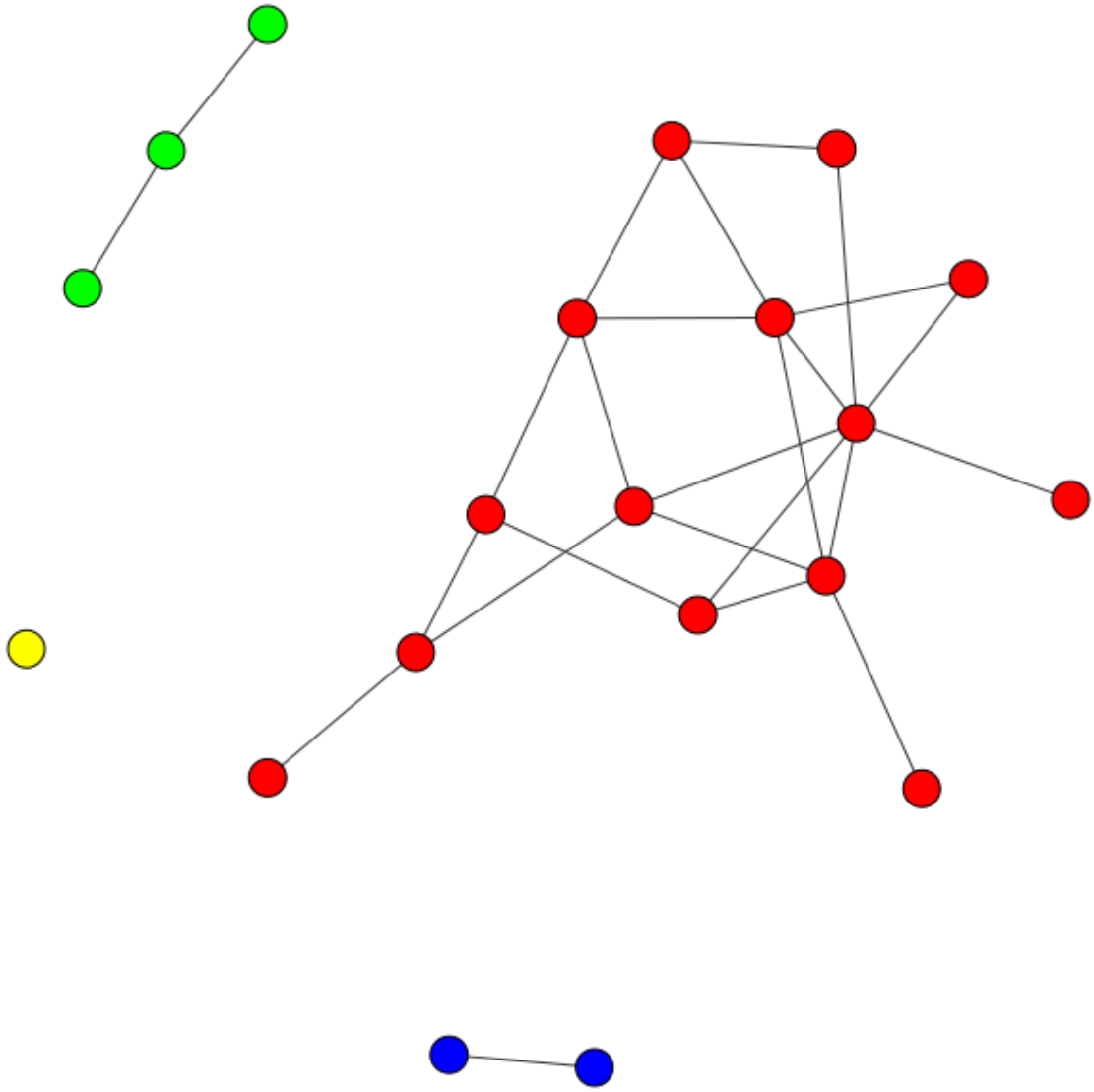
Suradnja među autorima, ustanovama i zemljama se unutar scientometrije najčešće tretira kao zasebna tema, što je slučaj i u ovom istraživanju. Česta i vrlo svrsishodna metoda koja se koristi za proučavanje suradnje je analiza mreža. U zadnje vrijeme, analiza mreža se počinje sve češće koristiti i u citatnim analizama poput analiza ko-citiranosti (White, 2003). Razlog ovome je što je analiza mreža prikladna za obradu relacijskih podataka odnosno, u ovom kontekstu, odnosa dobivenih kroz supojavnost. S obzirom i na druge pozitivne značajke i mogućnosti analize mreža, ona je važan dio metodologije ovog istraživanja.

2.5.4 Analiza mreža

Analiza mreža, u širem smislu, je moderan pristup koji ima korijene u dva glavna utjecaja: teoriji grafova i analizi društvenih mreža (Newman, 2010). Na temelju sve većeg broja radova iz ove problematike, u novije se vrijeme može se govoriti i o znanosti o mrežama kao mladoj disciplini koja se bavi razvojem jedinstvene teorije i empirijskim istraživanjem svih vrsta mreža (od društvenih mreža, do prometnih i internetskih mreža), a njene kvantitativne metode se temelje na teoriji grafova. Budući da se istraživanja najčešće baziraju na velikim i tzv. kompleksnim mrežama, znanost o mrežama se bavi i povezanim računskim i softverskim izazovima. Ovom pristupu osnovni cilj je opisati opća svojstva mreža na makro razini.

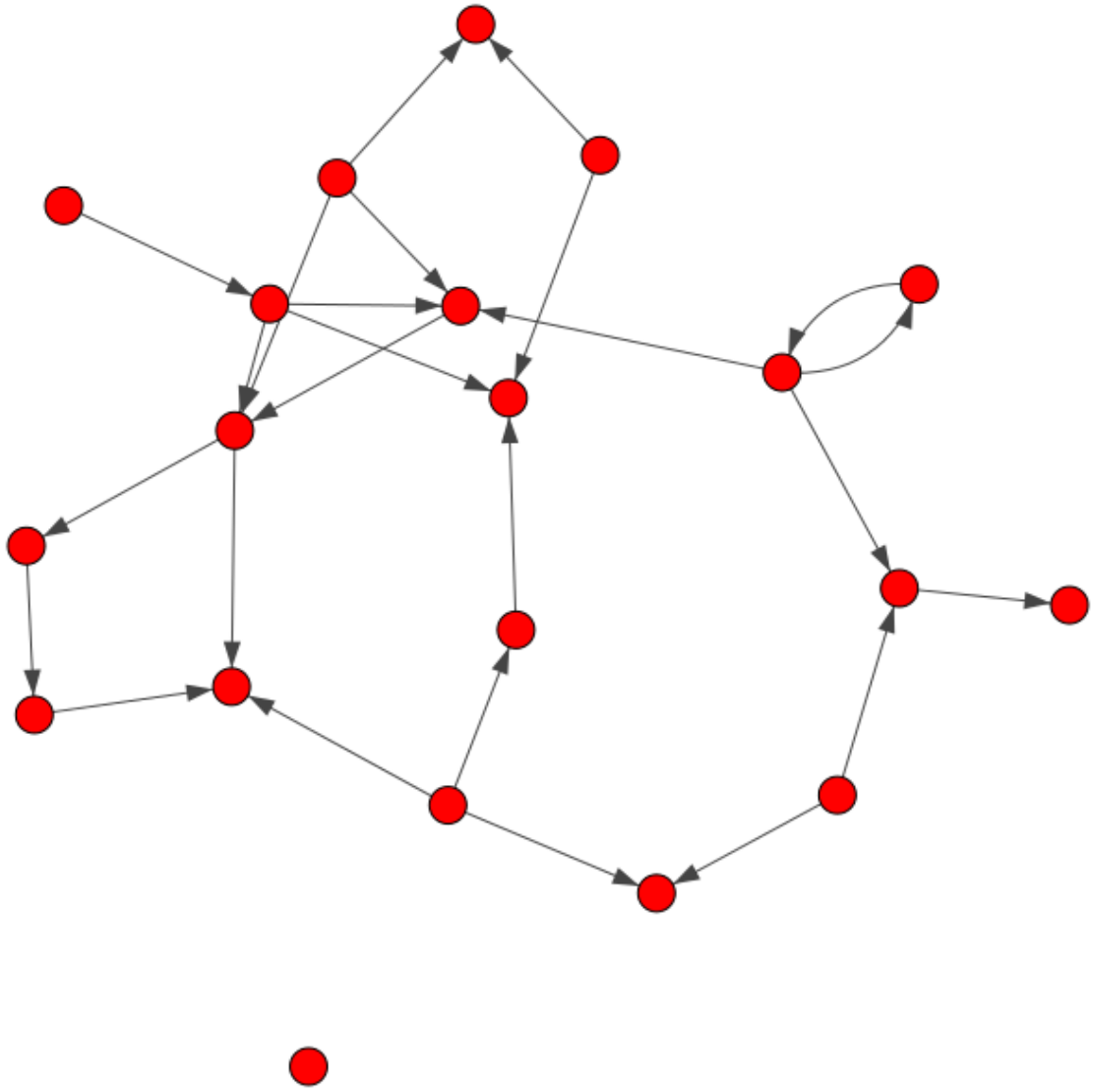
Konkretno, analiza mreža se provodi na podacima strukturiranim u matematičku strukturu grafa koji mogu biti usmjereni i neusmjereni. Kod neusmjerenih mreža nema razlike između odnosa a -> b i b -> a, tj. smjer odnosa ne postoji. Primjer ovakve mreže je mreža koautorstva gdje veza označava koautorstvo i nema konceptualne razlike između "a je surađivao sa b" i "b je surađivao sa a". Slučajno stvorena neusmjerena mreže prikazana je na slici 11.

Slika 11. Primjer slučajno generiranog neusmjerenog grafa



Usmjereni grafovi sadrže dodatnu informaciju o smjeru odnosa odnosno postoji razlika između a -> b i b -> a. Primjer ovakve mreže je citatna mreža u kojoj postoji razlika između "a citira b" i "b citira a". Slučajno stvorena usmjerena mreža je vidiljiva na slici 12.

Slika 12. Primjer slučajno generiranog usmjerenog grafa



Izraz "kompleksna mreža" često se koristi za velike mreže u strukturi grafa koje sadrže i dodatne informacije tj. atribute čvorova i veza. Ovakve mreže u načelu reprezentiraju neke stvarne podatke za razliku od mreža u teoriji grafova koje se često koriste za prikaz posebnih struktura i rješavanje matematičkih problema u sklopu tog teoretskog okvira. U podatkovnom smislu, "mrežu" možemo shvatiti kao generičku strukturu podataka u namjeni sličnoj tablici, ali koja eksplicitno opisuje i veze između "redaka". Ovakvu mrežu je moguće razlučiti u dvije tablice: atribute čvorova i atribute veza. Sve mreže korištene u ovom doktoratu su kompleksne mreže.

Kao što je uočljivo već u prvih nekoliko paragrafa ovog poglavlja, terminologija vezana uz mreže (odnosni grafove) može biti zbunjujuća. Na hrvatskom jeziku se pojam "graf" lako može zamijeniti za "grafikon", što je katkad slučaj i na engleskom. Drugim riječima, s obzirom na više disciplina koji se bave sličnom problematikom, kao i brz razvoj metodologije, u multidisciplinarnom području analize mreža postoji više terminoloških rješenja i na engleskom i na hrvatskom jeziku. U teoriji grafova, na primjer, uobičajeno je nazivati elemente grafa rubovima (eng. *vertex*) i bridovima (eng. *edge*). U analizi društvenih mreža uobičajeniji su nazivi su akter (eng. *actor*) i veza (eng. *tie*). Za "čvor" u mreži se vrlo često koristi i engleski izraz *node*. U sklopu ovog rada koristit će se nazivi čvor i veza kao, prema mišljenju autora, najneutralniji izrazi onoga što predstavljaju u mreži.

Što se bibliografskih i citatnih metapodataka tiče, oni podržavaju očite primjere mreža. Jedna od njih je citatna mreža tj. idejna podloga citatnom indeksu. U ovoj mreži radovi su povezani referencama tj. citatima. Radi se o usmjerenim vezama jer postoji distinkcija između odnosa a -> b i b -> a. Rad s lijeve strane je *citirajući rad*, a s desne *citirani rad*. U praksi, odnosi gdje a citira b i b citira a, biti će vrlo rijetki jer se u velikoj većini slučajeva radovi citiraju nakon objave tj. citirani radovi su objavljeni prije citirajućih.

Citatna mreža je primjer informacijske mreže u kojoj čvorovi predstavljaju informacijske jedinice umjesto društvenih aktera. Ovakve mreže, međutim, nisu podobne za vizualizaciju jer svaki čvor sadrži velik broj ulaznih i izlaznih veza, a i sam broj čvorova je obično velik.

Drugi, često korišten, slučaj mreža je koautorska mreža tj mreža u kojima su čvorovi autori, a veze označavaju koautorstvo na barem jednom radu u nekom skupu promatranih radova. U ovom grafu veze nisu usmjerene tj. a -> b i b -> a označavaju isti odnos. Kako je u ovom radu analiza mreže koautorstva primaran pristup analizi autora, autorstava i autorskih trendova na radovima u časopisu *Scientometrics*, taj je pristup i postupak podrobnije opisan u kasnijem poglavlju o analizi suradnje.

Uz mjere centralnosti i gustoće mreže, tradicionalno korištene u sociološkom pristupu tj. SNA (eng. *Social Network Analysis*), razvijene su i mnoge druge mjere koje se koriste za opis čitave mreže. Neke od njih su: koeficijent grupiranja, prosječna duljina puta i dijametar. Iste omogućuju opis topologije dobivene mreže i procjenu koji mehanizmi djeluju u procesu povezivanja. Navedeno se postiže usporedbom različitih svojstava mreže sa posebnim modelima koji opisuju različite tipove mreža. Na primjer, moguće je vidjeti razlikuje li se

dobivena mreža od mreže koja se na istom broju elemenata može dobiti slučajem (Erdős-Rényijev model, korišten za generaciju slučajnih mreža na slikama 10 i 11). Zatim, odgovara li dobivena mreža strukturi „malog svijeta“ (Watts-Strogatzov model). Naposljetku, moguće je utvrditi da li se čvorovi povezuju po principu referencijalnog povezivanja, odnosno s akterima koji imaju veći broj veza (Barabási-Albertov model).

Struktura malog svijeta kod koautorstva u znanosti opisuje mrežu gdje postoji visoka razina lokalnog grupiranja, što znači da suradnici jednog autora često i sami međusobno surađuju. Istovremeno, broj koraka između klastera je malen. To najbolje ilustrira jedan od najranijih primjera analiza društvenih mreža – projekt Erdösevog broja (Hoffman, 1987).

U kontekstu koautorstva mehanizam preferencijalnog povezivanje opisuje dobro utvrđenu pojavu da je distribucija stupnja centralnosti (broja izravnih veza, suradnji) nekog skupa autora izrazito pozitivno asimetrična, odnosno da je vjerojatnost povezivanja novog člana s nekime u mreži proporcionalna broju veza koje čvor već ima. Drugim riječima, novi autor koji ulazi u mrežu najvjerojatnije će biti koautor s nekim tko već ima veliki broj koautora. Prema tome, prominentni znanstvenici su odgovorni za povezivanje mreže. Ta pojava odražava tzv. Matejev efekt prema kojem oni koji imaju puno dobivaju još više. Važna karakteristika mreže formirane prema ovom principu jest da je centralizirana, ima nekoliko centralnih čvorova i veći broj perifernih, te omogućuje brzo širenje informacija.

Kako je opažanjem realnih podataka zaključeno da se mreže rijetko formiraju samo po principu slučaja ili preferencijalnog povezivanja, razvijene su različite vrste hibridnih modela. Oni opisuju nastanak mreža kombinacijom slučajnih i strategijskih mehanizama.

Analiza mreža, dakle, pruža bogat metodološki okvir za operacionalizaciju proučavanja odnosa među akterima tj. slučajevima ili predmetima. U okviru ovog rada analiza mreža se koristila prvenstveno za proučavanje koautorstava i suradnje, kako je detaljnije opisano kasnije u tekstu, te za proučavanje kocitiranosti radova i časopisa.

2.5.4.1 Mjere vezane uz analizu mreža

Mjere vezane uz analizu mreža možemo podijeliti na one koje se odnose na značajke mreže (tj. koje opisuju strukturu cjelovite mreže koju tvore više čvorova i veza) te one koje se odnose na značajke nekog čvora (tj. koje opisuju ego mreže ili/i poziciju pojedinca u cijeloj mreži).

U nastavku se nalazi popis odabranih mjera koje se tiču značajki čvorova tj. mreža. Mjere su ovdje prenesene s kratkim generičkim opisima kako bi se dao općenit uvid u mogućnosti analize mreža, a odabrane su mjere koje se koriste u ovom istraživanju. Sve navedene mjere proizlaze iz strukture mreža, a njihova konkretna interpretacija ovisi o konkretnom slučaju mreže - o tome što ona reprezentira. Konkretna upotreba ovih pokazatelja je pobliže opisana u poglavlju o koautorstvu.

Značajke čvorova:

- **stupanj** (eng. *degree*) - broj veza nekog čvora i.e. broj prvih susjeda
- **međupovezanost** (eng. *betweenness*) - koliko puta se neki čvor nalazi na putu između bilo koja druga dva čvora
- **blizina** (eng. *closeness*) - inverzna mjera udaljenosti pojedinog čvora od svakog čvora u mreži
- **artikulacijski čvorovi** (eng. *articulation points* ili *cut vertices*) - čvorovi koji povezuju inače nepovezane dijelove mreže.
- **lokalni koeficijent grupiranja** (eng. *local clustering coefficient*) - stupanj u kojem su susjedi određenog čvora međusobno povezani

Značajke mreže:

- **distribucija stupnja centralnosti** – frekvencija čvorova s *n* veza
- **gustoća** (eng. *density*) – broj ostvarenih veza u mreži u odnosu na maksimalni mogući broj veza
- **duljina puta** (eng. *path length*) – udaljenost čvorova mjerena brojem veza među njima
- **dijametar** (eng. *diameter*) - najveća zabilježena duljina puta između dva čvora u mreži
- **prosječan koeficijent grupiranja** (eng. *average clustering coefficient*) – vjerojatnost postojanja veze između dva čvora s kojima je povezan jedan čvor
- **glavna komponenta** (eng. *main component*) – najveći povezani dio mreže
- **asortativnost** – tendencija stvaranja veze među čvorovima s jednakim brojem veza

Velik broj navedenih mjera ima i svoje usmjerene inačice. Na primjer, stupanj povezanosti se u usmjerenim grafovima može podijeliti na ulazni i izlazni stupanj (eng. *in-degree* i *out-degree*), a neke mjere imaju posebnu formulu za usmjerenu inačicu. S obzirom da se u ovom istraživanju koriste isključivo neusmjereni grafovi, razlike u izračunu mjera koje imaju usmjerene i neusmjerene inačice nisu opisane.

2.5.5 Bibliometrijski pregled radova u *Scientometrics*

Bibliometrijski pregled radova i priloga objavljenih u časopisu *Scientometrics* pruža osnovne indikatore o tom časopisu u promatranom razdoblju tj. 1978.-2010. Radi se o jednostavnim, ali važnim indikatorima koji uključuju pregled količine radova i strukture objavljivanja časopisa *Scientometrics*. Preciznije, riječ je o:

- broju radova, volumena i brojeva te porastu kroz vrijeme
- pregledu vrsta radova i priloga u različitim vremenskim periodima
- pregledu brojeva teorijskih, metodoloških i primijenjenih članaka u različitim vremenskim periodima

2.5.6 Autorstvo radova, produktivnost autora i suradnja među autorima

Teme vezane na znanstvenike kao autore su, očekivano, među glavnim scientometrijskim temama. Proučavanjem autorstva proučava se možda najvažnija veza između aktera u znanosti i izlaza (eng. *output*) znanstvenih procesa. Proučavanje autorstva i autora možemo u načelu podijeliti na proučavanje radova u odnosu na broj autora, proučavanje autora u odnosu na radove te proučavanje suradnje.

2.5.6.1 Autorstvo radova

Proučavanje radova u odnosu na autore prvenstveno daje informaciju o broju autora na radu. Drugim riječima, proučava se distribucija broja autora po radovima u nekom skupu podataka. Ova distribucija informira o tipičnom autorstvu rada u nekom skupu radova. Možda je korisno i napomenuti da prilikom pregleda podataka valja biti pažljiv kako se ne bi zamijenile jezično slični konstrukti "broj autora po radu" i "broj radova po autoru" koji pak prenose posve različite informacije.

Opis distribucije broja autora po radu pruža važan uvid u trendove pisanja i suradnje u nekom području. Kao što je već ranije spomenuto, nije svejedno da li područje pokazuje "normalno" autorstvo (npr. s većinom radova do 5 autora i maksimalnim brojem autora do 20) ili "super" autorstvo (npr. s većinom radova s iznad deset autora i s maksimalnim brojem autora od nekoliko stotina). Informacije dobivene u ovom koraku znatno utječu na kasnije analize produktivnosti i suradnje tj. na odabir metode brojanja te kasniju interpretaciju rezultata.

U ovom istraživanju su stoga podaci o autorstvu među prvim obradama i prikazanim rezultatima. Ovdje se radi o jednostavnim pokazateljima poput deskriptivne statistike distribucije broja autora po radu kao i kontekstualni podaci poput broja jednoautorskih i višeautorskih (tj. dvoautorskih, troautorskih ...) radova. Navedeno je važno, ne samo pri interpretaciji podataka o produktivnosti, već i informira o obrascima suradnje putem koautorstva, odnosno nedostatku iste ako prevladavaju jednoautorski radovi.

2.5.6.2 Produktivnost autora

Kao što je opisano u uvodu, distribucija objavljenih radova neke skupine znanstvenika je izrazito asimetrična, gotovo eksponencijalna. Prema tome, za većinu radova odgovoran je mali udio znanstvenika. U smislu aplikacije ovog principa na publikacije autora, spomenuta pravilnost se zove Lotkin zakon. Lotkin zakon je pak posebna upotreba Zipfovog zakona koji je originalno formuliran kako bi opisao frekvencije pojavljivanja riječi u jeziku. Isti princip je u ekonomiji poznat kao Paretova distribucija.

U smislu proučavanja nekog tijela literature ili discipline, istraživanje produktivnosti autora prikazuje učestalost objavljivanja pojedinih autora u nekom vremenskom periodu. Kod proučavanja autora u časopisu *Scientometrics* za pretpostaviti je da će distribucija broja radova po autorima u načelu zadovoljavati Lotkin zakon. Također, za pretpostaviti je da velika većina autora koji su objavili u časopisu *Scientometrics*, vjerojatno i svi, da su objavili i u drugim časopisima. Dakle podaci o produktivnosti autora se u ovom kontekstu koriste u dvije svrhe:

1. prikaz trendova objavljivanja autora u časopisu *Scientoemtrics* u različitim vremenskim razdobljima prikazuje pojavljivanje ili nedostatak prepoznatljive skupine autora koji se bave scientometrijom u značajnoj mjeri, kao primarnom djelatnošću, a manje kao sporednim područjem rada
2. identifikacija udjela radova visoko produktivnih autora u časopisu *Scientometrics* pomaže u opisu centralnih autora za područje kao skupine i individualno

2.5.6.3 Suradnja među autorima i analiza mreže koautora

Analiza mreža, posebno analiza društvenih mreža, smatra se privilegiranim teoretskim pristupom za proučavanje trendova suradnje među znanstvenicima putem koautorstava (De Stefano et al., 2013). Ranija istraživanja su pokazala da se discipline razlikuju s obzirom na stupanj i vrstu suradnje (Melin i Persson, 1996; van Rijnsoever et al., 2008).

Važan metodološki korak u analizi društvenih mreža je definiranje granice mreže. Definiranje granice mreže se odnosi na problem određivanja granica skupa jedinica analize u mreži (čvorova). Granica mreže definira skup aktera za koje se pretpostavlja da čine kompletni skup aktera mreže koja je predmet istraživanja (Prell, 2011). Laumann i suradnici (1989; prema De Stefano et al., 2011) opisuju tri općenite strategije specifikacije granica kod istraživanja suradnje među znanstvenicima na temelju podataka o koautorstvu: pozicijski pristup, pristup utemeljen na sudjelovanju u nekoj vrsti relacijskih događaja i relacijski pristup. U ovom radu se koristi pristup utemeljen na sudjelovanju u nekoj vrsti relacijskih događaja (eng. *event-based*), u kojemu je predmet istraživanja sadržaj određenog izvora podatka, a koautorstvo se smatra relacijskim događajem. U ovom slučaju se ne polazi od liste autora, već aktere u mreži predstavljaju svi autori potpisani na radove objavljene u časopisu Scientometrics od 1978. do 2010.

Nakon razrješavanja imena autora, izrađena je mreža koautorstava na sljedeći način:

za svakog autora:

- ucrtaj čvor na graf
- dodaj jednoznačno ime autora kao atribut čvora
- dodaj skup šifri radova na kojima je ta osoba među autorima kao atribut čvora

za svaku kombinaciju bez ponavljanja 2. reda čvorova:

- ako presjek skupova radova tih dvaju čvorova nije prazan skup:
    - ucrtaj vezu
    - dodaj rezultat presjeka kao atribut veze

Svaki čvor dakle ima pripadajući skup šifri radova na kojima je ta osoba *među* autorima, a svaka veza ima pripadajući skup šifri radova na kojima su autori koje povezuje ostvarili koautorstva. S obzirom da je ovo istraživanje usmjereno na časopis, niti jedan od ovih skupova ne može biti prazan skup jer su informacije o postojanju i razlikovanju autora proizašle upravo iz tih radova. Ukoliko postoji takav autor čiji skup radova je prazan skup, tada se dogodila greška u obradi podataka jer nije postojao drugi izvor informacija o autorima, a koautorstvo među autorima postoji samo ukoliko su dva autora prisutna na barem jednom radu. Skup radova veze je uvijek podskup skupa radova svakog od čvorova koje povezuje. Mreža autora sadrži i autore koji su objavljivali samo jednoautorske radove. Ovakvi čvorovi su "izolati" tj. nisu povezani ni s jednim drugim čvorom.

U istraživanju s drugim fokusom, moguće je imati i graf koautorstva na kojemu postoje "autori" bez radova. Na primjer, u istraživanju koje polazi od nekog skupa znanstvenika mogu postojati takvi čvorovi u grafu koautorstava za koje nije pronađen niti jedan rad. Ovakvi čvorovi se uglavnom ne uključuju u krajnju analizu, a odluka o uopće njihovom uključivanju u obradu uopće ovisi o fokusu i operacionalizaciji istraživanja. Oni mogu informirati o relevantnosti izvora za istraživanje ili mogu sadržavati atribute koji nisu vezani uz graf, poput godine rođenja ili ustanove, kako bi bili uključeni u uzorak tj. u analitičke postupke koji koriste taj atribut. U potonjem slučaju, to može biti korisno ukoliko je graf, a ne tablica, primarna struktura na kojoj se provodi obrada.

U svakom slučaju, nakon stvaranja grafa, skupovi šifri radova su se koristile u dvije svrhe:

1. kardinalnost skupa, tj n radova tog autora ili te suradnje, koristila se kao težina čvora odnosno veze
2. šifre radova su omogućile povezivanje i agregaciju podataka vezanih uz radove, poput zbroja citata, *h*-indexa, godine objave i sličnog

Također, jednoznačno ime autora kao atribut čvora omogućuje povezivanje mreže s podacima o autorima, a oni se na grafu radi korištenja skupova jednoznačnih identifikatora mogu lako

grupirati bez dupliciranja. Jedan od primjera u kojem ovakav pristup može biti koristan, je grupiranje po ustanovama autora kako bi se vidjeli koautorski odnosi između ustanova.

U kontekstu obrade mreže autora, ranije navedene mjere su korištene na sljedeći način:

Broj čvorova i veza

Broj čvorova (n) i veza (L) u mreži su osnovni parametri kod opisa mreže. U mreži koautorstva, čvorovi predstavljaju pojedine autore, a veza između čvorova označuje da su zajednički koautori na barem jednom radu. Veze se u mreži koautorstva tretiraju kao neusmjerene. Veći broj suradnji (zajednička koautorstva dvojice autora) se grafički prikazuje kao veza (linija) proporcionalno veće debljine.

Mjere centralnosti

Postoje različite mjere koje opisuju poziciju pojedinog čvora u cjelokupnoj mreži (tzv. mikro mjere). Najpoznatije i najčešće korištene u mrežama koautorstva su:

- **stupanj** – opisuje koliko je čvor povezan direktnim vezama
- **međupovezanost** – opisuje važnost čvora u povezivanju ostalih čvorova
- **blizina** – opisuje koliko lako čvor može doći do ostalih čvorova

U ovom radu će se u svrhu opisa mreže koautorstva koristiti samo stupanj centralnosti. Stupanj centralnosti (eng. degree centrality) je broj direktnih veza koje ima čvor (kod neusmjerenih veza). U mreži koautorstva on znači broj različitih autora iz uzorka s kojima akter surađuje barem jednom. Za razliku druge dvije spomenute mjere, ova mjera centralnosti uzima u obzir samo izravne veze.

Prosječni stupanj centralnosti ovisi i o g, veličini mreže; što je mreža veća, veća je maksimalna moguća vrijednost stupnja centralnosti. Stoga, određena vrijednost stupnja centralnosti može značiti da je akter dobro povezan unutar manje mreže, ili da je povezan samo s nekoliko drugih autora u velikoj mreži.

Mjera stupnja centralnosti odgovara intuitivnom shvaćanju prema kojem je pojedinac s najviše direktnih veza najutjecajniji. Međutim,to je samo lokalna mjera, koja ne uzima u obzir poziciju aktera u čitavoj mreži niti njegove indirektne veze. Primjerice, pojedinac može imati mnogo direktnih veza, ali s akterima koji nisu u glavnoj komponenti mreže, pa je prema tome

njegova pozicija mnogo slabija od drugog pojedinca koji ima jednaki broj direktnih veza, ali s akterima koji su bolje povezani s ostatkom mreže.

Obrada najveće komponente

Glavna komponenta povezanosti je dio mreže koji nema izoliranih čvorova, odnosno gdje su svi čvorovi direktno ili indirektno povezani. Brojčano se izražava kao n čvorova u najvećoj komponenti (%). Obično zauzima najveći dio mreže i unutar nje se odvija najveća razmjena informacija (Newman, 2010). Mreže se sastoje obično od nekoliko komponenti različite veličine, od kojih su neke povezane s drugima, a neke su izolirane. Najveća komponenta ima najveću povezanu grupu pojedinaca u mreži komponenta, a njena veličina se može opisati postotkom čvorova mreže koji su povezani u jednu najveću grupu.

Prosječna udaljenost

Najkraći (geodezijski) put između dva čvora *i* i *j* je put s najmanjim brojem veza. Najkraći put se često zove i udaljenost čvora *i* i *j*, te se često bilježi s *d_ij* ili samo *d*. Obično se pronalazi više putova između dva čvora koji mogu biti iste ili različite duljine. Prosječna vrijednost najkraćeg puta za sve povezane čvorove u mreži daje prosječnu duljinu puta koja se naziva i prosječna udaljenost ili razmak.

Koeficijent grupiranja

Na razini pojedinog čvora tzv. lokalni koeficijent grupiranja opisuje stupanj u kojem su susjedi određenog čvora međusobno povezani.

Na razini cijele mreže, putem prosjeka lokalnih koeficijenata svih čvorova dobiva se globalni koeficijent grupiranja i može se interpretirati kao vjerojatnost da su dva susjeda slučajno izabranog čvora i sami međusobno direktno povezani. Koeficijent grupiranja u kontekstu mreža koautorstva daje informaciju o tome kolika je vjerojatnost da će suradnici bilo kojeg autora u mreži i sami međusobno surađivati.

Gustoća

Gustoća (g, eng. *density*) mreže predstavlja proporciju ostvarenih veza od svih mogućih veza. To je mjera općenite razine povezanosti među čvorovima u mreži, i služi kao općenita mjera kohezije. Razvojem neke grupe vremenom raste i gustoća, što sugerira da grupa postaje sve više kohezivna.

Dijametar

Dijametar (d_max ili δ) je najdulji najkraći (geodezijski put) put u mreži. Drugim riječima, najkraći put između dva najudaljenija čvora/autora.

Artikulacijski čvorovi

Artikulacijski čvorovi (eng. *articulation point*, *cut-point*) su akteri koji povezuju inače nepovezane dijelove mreže. Odnosno, to su osobe koje su povezane s barem dvije osobe koje međusobno nisu direktno ni indirektno u vezi. Bez tih aktera, mreža bi se raspala na veći broj nepovezanih komponenti. Prema Burtovoj teoriji (1992) strukturalnih pukotina, zbog pristupa međusobno nepovezanim izvorima, artikulacijski čvorovi posjeduju društveni kapital. Prema Moodyjevoj teoriji strukturalne kohezivnosti (2004), strukturalno kohezivnija mreža ima manji broj artikulacijskih čvorova, odnosno njena struktura ne ovisi o velikom broju pojedinih čvorova.

Asortativnost

Asortativnost ili koeficijent asortativnosti mjeri tendenciju čvorova u mreži da se povezuju s čvorovima koji su im slični u nekom kvantitativnom svojstvu. U kontekstu mreža koautorstva ta se mjera koristi najčešće za ispitivanje korelacije u stupnju centralnosti između svih povezanih čvorova u mreži. Računa se preko Pearsonovog koeficijenta korelacije, varira od -1 do +1: pozitivne vrijednosti ukazuju da se autori koji imaju veliki broj suradnika povezuju s drugim autorima koji imaju veliki broj suradnika, odnosno da je mreža asortativna. Kad su vrijednosti negativne, mreža se naziva disasortativnom, a kad je korelacija nepostojeća (blizu nuli), radi se o neasortativnoj mreži.

Očekivane vrijednosti

Očekivana prosječna udaljenost i očekivani koeficijent grupiranja su prosjeci ovih mjera na slučajno generiranim mrežama koje imaju zadane parametre jednake empirijski dobivenoj mreži (broj čvorova i gustoću). Kako je broj mogućih slučajnih mreža vrlo velik, obično se računaju na većem broju slučajnih mreža s istim zadanim parametrima. Za svaki prosjek mjera slučajnih mreža, u ovom je istraživanju generirano 1 000 slučajnih mreža.

2.5.7 Citatne i ko-citatne analize

Kao što je opisano u uvodu, citatne analize su važan sastavan dio bibliometrijskog proučavanja znanosti. Bez obzira na problematiku preciziranja značenja citata, citati su intuitivan mehanizam u formalnoj znanstvenoj komunikaciji. Kao takve možda ih najbolje zahvaća poznata metafora Blaizea Cronina (1984) koji kaže da su citati zamrznuti tragovi u krajobrazu znanstvenih postignuća. Kvantitativno proučavanje znanstvene literature koje se ne koristi informacijama o citiranosti, a takvi podaci su dostupni u zadovoljavajućem opsegu, propušta važnu informaciju o životu rada nakon objave i o odjeku znanja koje se tim radovima prenosi.

Citatna analiza, kako je trenutačno zastupljena u scientometrijskoj literaturi, ide ruku pod ruku s časopisima. Razlog nije samo dostupnost podataka već i radi toga što znanstveni časopisi tvore relativno uređen i zatvoren sustav objavljivanja, autora i publike što ih čini idealnim za kvantitativne studije znanosti (Yanovsky, 1981). Dapače, upravo su navedene specifičnosti ove literature prethodile izgradnji većine citatnih indeksa upravo na časopisima, te čine časopise međusobno usporedivim. Za uspjeh citatnih indeksa nije koliko zaslužno pronalaženje literature koliko mogućnost vrednovanja časopisa i autora kroz citate koje su primili (Garfield, 2010). Kasnije korištenje pokazatelja o citiranosti u ove svrhe pokazuje značajne razlike među područjima znanosti i otvara nova poglavlja u citatnim analizama. U ovom smislu, citatna analiza je primjerena za skupove radova u časopisima koji su konzistentno indeksirani u citatnom indeksu i koji su iz znanstvenih disciplina u kojima je članak u časopisu glavni medij komunikacije (Cronin, 1984).

Što se časopisa *Scientometrics* tiče, s obzirom na njegov međunarodan status, kontinuirano izlaženje i indeksiranost u citatnim indeksima od početka izlaženja do danas, citatna analiza je primjeren alat za proučavanje tijela literature koji je taj časopis objavio. Važna tema u

rezultatima je, dakle, citiranost radova objavljenih u časopisu *Scientometrics*. U ovom smislu se proučavaju svi radovi objavljeni 1978-2010 i svi citati indeksirani u bazi WoS 1978-2012. Kao što je u poglavlju o pripremi podataka opisano, samo WoS citati su odabrani budući da pokrivaju kompletno vremensko razdoblje, što ih čini mjerodavnijim za glavnu temu istraživanja, za razliku od te vrste podataka dostupnih u Scopusu koji pokrivaju pola promatranog razdoblja.

U sklopu ovog istraživanja, detaljnija citatna analiza časopisa *Scientometrics* se provodi samo na člancima kao nosiocima primarnog znanja u području. Razlog ovomu je umanjivanje šuma prisutnog u razlozima citiranja ili necitiranja, odnosno uključeni su oni radovi koji prema sadržaju imaju jednaku šansu biti citirani. Naravno, i kod članaka postoje varijante u razlozima citiranja. Mnoga istraživanja pokazuju kako su teorijski (posebno pregledni radovi) i metodološki radovi više citirani od primijenjenih (Peritz, 1983), što je provjereno i u ovom radu.

Citatne analize časopisa *Scientometrics* provedene u ovom istraživanju mogu se podijeliti na analize članaka objavljenih u *Scientometrics* prema citatima koje su primili i prema citatima koje pružaju. Budući da proučavanje članaka u časopisu *Scientometrics* kao citiranih radova i proučavanje istih članaka kao citirajućih radova promatra iste teme u različitom smjeru, rezultati su sakupljeni tematski, a manje po smjeru citiranja. Na primjer, citirajući časopisi i citirani časopisi obrađeni kao tematska cjelina koja pruža uvid u disciplinarno opredjeljenje i multidisciplinarnost časopisa scientometrics.

U ovom istraživanju, što je čest slučaj i u ostalim istraživanjima ove vrste, ulazni podaci pružaju više informacija za analizu citata koje radovi u *Scientometrics* primaju jer su po definiciji svi citirajući radovi indeksirani u citatnom indeksu (vidi opis citatnog indeksa u uvodu) pa su za njih dostupnu puni standardizirani bibliografski zapisi. Dodatno, u ovom istraživanju zapisi o citirajućim radovima su direktno povezani na šifre zapisa o radovima objavljenim u *Scientometrics* tj. nisu identificirani samo kroz identifikatore tipa "Garfield E., 1979, CITATION INDEXING IT" već kroz pune bibliografske zapise.

S druge strane, citirani radovi nisu svi indeksirani i puno je teže povezati citirane radove s kvalitetnim bibliografskim zapisima. Za razliku od citirajućih radova kod kojih su svi citatni navodi povezani s metapodacima, kod citiranih radova moralo se pristupiti analizi putem informacija koje se mogu dobiti ili iz navoda poput već prikazanog "Garfield E., 1979,

CITATION INDEXING IT" ili pak iz navoda iz punog teksta članaka. Navodi u popisu referenci u časopisu *Scientometrics* nisu uniformni, a početkom razdoblja uključuju i fusnote tj. tekstualne bilješke na popisima referenci. Nakon testiranja, procijenjeno je da se preciznije informacije ipak dobivaju iz WoS podataka.

Nedostatak WoS referenci je što je teško procijeniti o kojoj se točno publikaciji radi, kao i utvrditi točno autorstvo budući da je zapisana samo informacija o prvom autoru. Ipak, u slučaju popisa literature u časopisima kod WoS referenci je lako utvrditi relativno jednoznačno skraćeno ime časopisa kao i godinu izdavanja. Budući da su upravo ovi podaci procijenjeni kao zanimljivi za potrebe pokazatelja o multidisciplinarnosti časopisa i zastarijevanja literature, analizi citiranih radova u časopisu *Scientometrics* se pristupilo iz WoS podataka.

U nastavku slijede opisi tematskih cjelina opisanih kroz citatne analize.

Međunarodna citiranost časopisa Scientometrics

Pregled brojeva citata po vremenskim razdobljima, odnosno vezanih pokazatelja, pružaju uvid u međunarodnu relevantnost časopisa *Scientometrics* u odnosu na časopise indeksirane u WoS citatnim indeksima. Kod ovih podataka riječ je o relativno jednostavnim pokazateljima poput postotka citiranih članaka i deskriptivnih statistika o distribuciji broja citata po članku u časopisu *Scientometrics*. Radi asimetričnosti distribucije, kao i kod ostalih analiza, naglasak je stavljen na neparametrijsku statistiku tj. kvartile i vezane mjere koje omogućavaju usporedbu među vremenskim razdobljima s obzirom na različiti broj radova po razdoblju.

Samocitati

Izraz "samocitati" se najčešće upotrebljava u smislu samocitata autora koji se definiraju kao citati prema radovima među čijim autorima se pojavljuje barem jedan od autora citirajućeg rada. Izraz se katkad koristi i u smislu samocitata časopisa odnosno samocitata ustanova (Aksnes, 2003).

Samocitati autora su posebno značajni u korištenju citata kao aproksimacije odjeka (MacRoberts i MacRoberts, 1989), jer se u smislu "odjeka" istraživanja, u kontekstu širenja znanja, samocitatima ne može pridodati isti značaj kao i drugim citatima. U sklopu ovog

istraživanja provjereno je koliki je udio samocitata autora u različitim razdobljima te koliko oni utječu na sliku citiranosti časopisa *Scientometrics*.

Budući da su podaci o citatima prikupljeni iz istih izvora prema kojima je izrađen popis varijanti imena za svakog autora koji je objavljivao u časopisu *Scientometrics*, isti je ponovno iskorišten za detekciju samocitata autora. Budući da se podacima upravljalo algoritamski, detekcija samocitata je logično i bez puno dodatnog truda uslijedila iz postupaka razriješavanja imena autora i povezivanja citirajućih i citiranih radova. Nedostatak pristupa je što ukoliko se u skupu citirajućih radova pojavio autor s istim inicijalima (i bez punog imena autora) kao i neki autor u časopisu *Scientometrics* automatski je pretpostavljeno da se radi o istom autoru što ne mora biti slučaj. Razrješavanje ovakvih grešaka, međutim, vrlo je teško bez obzira na način detekcije samocitata. Ipak, za pretpostaviti je da je udio takvih slučajeva malen, pogotovo zato jer WoS selektivno uključuje i puna imena autora, a u procesu detekcije koristilo se najdulje dostupno ime.

Citiranost radova u odnosu na osnovnu tematiku radova

Kao što je već opisano, svi članci su podijeljeni na teorijske, metodološke i primijenjene. Teorijski i metodološki članci uobičajeno primaju više citata i dominiraju na popisima najcitiranijih radova. Zbog važnosti problematike i tematskog određenja radova objavljenih u časopisu *Scientometrics* na razvoj scientometrije kao discipline, analiziran je i profil citiranja članaka u *Scientometrics* prema njihovoj osnovnoj tematici.

Citirajući i citirani časopisi

Časopisi koji često citiraju časopis *Scientometrics* prikazuju skup časopisa, pa tako i područja, u kojima je časopis *Scientometrics* relevantna literatura. Kako bi se bolje prikazala zastupljenost citata u citirajućim časopisima uz brojeve citata prikazani su i brojevi citirajućih i citiranih radova te prosječan udio citata u popisu literature citirajućih radova. Zadnji navedeni pokazatelj odgovara na pitanje "Koliki udio literature u citirajućim radovima je na časopis *Scientometrics*?". Ovakva formulacija pokazatelja izbjegava varijacije u gustoći citiranja u citirajućim časopisima. Drugim riječima, razlikuje dva časopisa iz kojih sličan broj radova pruža sličan broj citata, ali gdje se u jednom od njih radovi koji citiraju časopis *Scientometrics* više baziraju na njima nego u drugom koji citira puno veći udio druge literature i drugih časopisa.

S druge strane, časopisi koji su citirani u člancima u časopisu *Scientometrics* pružaju uvid u časopise na kojima se bazira *Scientometricsu*. Osim samog pregleda časopisa na koje se scientometrijska literatura poziva, područja tih časopisa pozicioniraju *Scientometrics* u odnosu na ostale discipline. Na ovaj način dobiva se uvid i u multidisciplinarnost područja scientometrije.

Sam časopis *Scientometrics* nalazi se u i skupini citiranih i u skupini citirajućih časopisa jer članci u časopisu *Scientometrics* mogu citirati članke u istom časopisu. S obzirom na jedinstvenu specijalizaciju časopisa u promatranom periodu pretpostavka je da će prosječan udio citata na *Scientometrics* u popisima literature članaka objavljenih u istom časopisu biti velik.

Visoko citirani članci

Pregled visoko citiranih članaka objavljenih u časopisu *Scientometrics* pruža uvid u članke koji su najčešće korišteni kao podloga drugim člancima. Visoko citirani radovi su posebno zanimljivi radi potencijalne povezanosti s visoko-kvalitetnim istraživačkim radom (Levitt i Thelwall, 2009). Kao što je spomenuto u uvodu, odabir visokocitiranih radova zaobilazi šum prisutan u razlozima citiranja te odabire radove za koje je vrlo vjerojatno da su citati relativno dobra aproksimacija kvalitete.

Važna odluka prilikom analize visoko citiranih članaka je kriterij selekcije. Glänzel i Schubert (1992) definiraju dva pristupa odabiru: putem poretka ili putem definiranja minimalnog broja citata. Putem poretka možemo jednostavno odabrati prvih N radova po citiranosti. N može biti odabran i prema udjelu tj. može se definirati kao prvih 100 ili kao prvih 10%. U drugom slučaju odabire se neki minimalni N citata i svi radovi s brojem citata većim od N se smatraju visoko citiranima.

U ovom slučaju odlučilo se na potonju mogućnost kako bi se zahvatili svi radovi koji su značajno citiraniji od ostalih radova u skupini bez obzira na njihov broj. Na ovaj način umanjena je nasilnost "reza" jer prilikom odabira prvih N, rad koji je na poziciji N + 1 može imati isto citata kao i rad na poziciji N. Pri odabiru broja citata koji se koristio kao granica, odabrana je definicija ekstrema kroz interkvartilni raspon (IQR) odnosno prema istoj definiciji tzv. *outliera* kao što se definiraju na *box and whiskers* grafikonu. Konkretan izračun granične vrijednosti za gornje ekstreme (radi značajki distribucije to su i jedini mogući u ovim

podacima) jest Q3 + IQR * x. Oko odabira za vrijednosti x ne postoji opće prihvaćeno pravilo, ali često (na primjer, u statističkom softveru poput SPSS-a) se koriste vrijednosti 1,5 za blage ekstreme (eng. *mild outliers*) i 3 za snažne ekstreme (eng. *extreme outliers*). U ovom istraživanju korištena je vrijednost za blage ekstreme za odabir članaka i snažne ekstreme za odabir časopisa. Kao visoko citirani članci definirani su, dakle, svi radovi čija vrijednost je prelazila zbroj trećeg kvartila (Q3) i interkvartilnog raspona pomnoženog s 1,5. Isto vrijedi i za visoko citirane časopise osim što je x=3 kako bi se dobili snažni ekstremi.

Visoko citirani i seminalni autori

Nasljeđivanjem broja citata s radova na autore dobivamo brojeve citata koje su autori primili na radove koje su objavili u časopisu *Scientometrics*. Uz ranije predočenu produktivnost autora, citati pružaju uvid u odjek radova autora pa time i dodatan uvid u autore važne za područje.

Doprinos autora se može kvantitativno promatrati kroz broj radova i broj citata. Jorge Hirsch (Hirsch, 2005) razvija *h*-indeks, pokazatelj koji ova dva aspekta doprinosa prikazuje kao jednu brojčanu vrijednost. *h*-indeks nekog skupa radova je najveći broj radova *h* koji je primio barem *h* citata. Ovaj pokazatelj predložen je kao mjera doprinosa autora i u tom smislu se najčešće i koristi, iako ju je moguće izračunati za bilo koji skup radova za koje su poznati citati.

h-index postaje iznimno popularan (rad u kojem je indeks predložen nalazi se među najcitiranijim radovima u *Scientometrics* kao što je vidljivo iz rezultata), a ubrzo se pojavljuje velik broj alternativnih pokazatelja u istom stilu. U sklopu ovog istraživanja koristi se i *g*-indeks kojeg je razvio Leo Egghe (2006) kao osjetljiviju alternativu *h*-indeksu. Jedan od nedostataka *h*-indeksa naime je neosjetljivost na autore koji su objavili manji broj visoko citiranih radova. *g*-indeks je definiran kao najveći *g* tako da top *g* radova zajedno imaju najmanje g^2 citata. Ovakva definicija čini *g* osjetljivijim na velike brojeve citata kod manjeg broja radova nego *h*, ali i osjetljivijim na utjecaj samocitata (Schreiber, 2008) te manje robusnim.

g-indeks često pruža finiju sliku, ali ne zamjenjuje *h*-indeks, već su ova dva indeksa u nekim slučajevima komplementarni (Costas i Bordons, 2008). *g*-indeks je osjetljiv na jedan visokocitirani rad, potencijalno s velikim brojem autora, a *h*-indeks nije podoban za "selektivne" autore koji objavljuju mali broj potencijalno visokocitiranih autora.

Kako bi se detektirali seminalni autori u području scientometrije iskorišten je primarno *g*-indeks kao osjetljivija mjera. Kao dodatni validator rangiranja autora prikazan je i *h*-indeks, koji može poslužiti i za usporedbu pokazatelja budući da se radi o visoko popularnom pokazatelju kojeg počinju računati mnogi citatni indeksi.

Python funkcije za izračun *h* odnosno *g*-indeksa mogu se pronaći u prilozima ovog rada.

Ko-citatna analiza

Jednostavno rečeno, dva rada su kocitirana ako su zajedno citirani u nekoj publikaciji. Formalnije, Small (1973) daje sljedeću definiciju kocitiranosti:

Ako je A skup radova koji citiraju dokument a, B skup radova koji citiraju b, tada je A∩B skup radova koji citiraju i a i b. n(A∩B) je frekvencija ko-citiranja. Relativna frekvencija ko-citiranja može se definirati kao n(A∩B)÷n(A∪B).

Kao i kod ostatka bibliometrijskih postupaka, informacije vezane uz radove se mogu naslijediti i na vezane entitete poput autora. Ko-citatna analiza autora rezultira klasterima autora koji su međusobno povezani zajedničkim pojavljivanjem na popisima referenci.

Ko-citiranost autora ima značajno mjestu u ko-citatnoj analizi i često se naziva akronimom ACA (i.e. eng *author co-citation analysis*). Ova metoda stvara mape znanstvenih disciplina više u smislu klastera autora, nego samih područja i one se moraju temeljiti na već poznatim istraživačkim interesima tih autora.

Otkad se prvi put pojavila u radovima White i Griffith (1981a, 1981b, 1982), ko-citatna analiza autora (eng. *author cocitation analysis*, ACA) se prikazuje u dvodimenzionalnim mapama (White, 2003). Tradicija ko-citatne analize je korištenje analize klastera (ulavnom se koristi aglomerativno hijerarhijsko klasteriranje) i multidimenzionalno skaliranje. (Leydesdorff, 1987) Kasnije se ovome pridodaje i faktorska analiza (White i McCain, 1998).

S vremenom se počinje koristiti velik broj tehnika za analizu, prikaz i interpretaciju podataka dobivenih kroz kocitatne analize. Među tehnikama se s vremenom počinju koristiti i mreže koje zadržavaju potrebnu kompleksnost te pružaju nove mogućnosti (White 2003). Mreže su same po sebi idealne za prikaz odnosa između predmeta. U ovom smislu mreža ko-citiranosti nije izvedbeno različita od mreže ko-autorstva: obje prikazuju supojavnosti u nekom okviru. Ovako dobivene mreže sa svim ko-citatnim vezama, međutim, se teško prikazuju i interpretiraju jer sve veze povezuju velik broj predmeta, a ključ je u selekciji značajnih veza. U ovom smislu koriste se dva postupka: određivanje minimalne snage koju veza mora imati kako bi ju se promatralo i korištenje takozvanih *pathfinder networks* koje se uglavnom skraćeno nazivaju PFNET.

Radi jednostavnosti prikaza u ovom istraživanju se koristilo određivanje minimalne snage veze. Budući da je prikazan velik broj analiza, ovako dobivena mreža prikazuje relativno jednostavno interpretabilne podatke bez potrebe pojašnjavanja značenja veze kroz relativno kompleksnu metodologiju. Što se PFNET mreža tiče, jednostavno rečeno PFNET je način kako preuzeti samo neke od veza u mreži. Nakon primjene PFNET algoritama s određenim postavkama dobiva se acikličan graf. Drugim riječima, korištenje PFNET mreže s određenim postavkama može se usporediti s korištenjem dendograma za vizualizaciju hijerarhijskog klasteriranja. Korištenje PFNET pristupa smanjuje zahtjevnost ACA-e dok zadržava, ili čak unapređuje, dobiti ostalih stilova mapiranja (White 2003). Prema istom radu, ovakva mreža bi bila svrsishodnija za daljnju kvalitativnu analizu dobivenih podataka.

2.5.8 Pokazatelji dobiveni iz tekstova radova

S točke gledišta bibliometrijskih istraživanja znanosti, bibliografski zapis je standardizirani metapodatkovni surogat za same radove. Kontrolirane bibliografije radova omogućavaju bibliometrijska istraživanja jer je ponavljanje posla prikupljanja podataka iz samih izvora neefikasno za bavljenje scientometrijom na značajnoj razini. Nadalje, standardizirani podatkovni zapisi koji se mogu pohraniti u bazama podataka podobni su za daljnju računalnu obradu, a korištenje istih izvora podataka ima pozitivan utjecaj po usporedivost i ponovljivost istraživanja. Uz to, sami zapisi su kompleksni tj. kodirani na različite načina, bogati informacijama i bazirani na pobiranju tekstualnih vrijednosti što čini jednoznačnu identifikaciju unutar nekog skupa zapisa problematičnom. Bibliografski metapodaci, dakle, ne

samo da su nezamjenjivi za scientometrijska istraživanja već je i formalno baratanje njima (u svrhu znanstvenih istraživanja) dovoljno problematično.

Ipak, sve veći udio digitalnih publikacija u znanosti odnosno pristupa digitalnim inačicama tiskanih, donosi mogućnost uključivanja samih znanstvenih publikacija u istraživanja koja se tradicionalno koriste njihovim metapodatkovnim surogatima[11].

Podaci iz tekstova radova se upotrebom računalnih tehnologija, ukoliko su isti digitalizirani ili digitalno stvoreni (kao što je slučaj s novijom literaturom), mogu direktno uključiti u kvantitativna istraživanja znanstvene literature. Računalno procesiranje strukture znanstvenih tekstova kao i samog teksta na jezičnoj razini je, međutim, problematika druge vrste od istraživanja bibliografskih metapodataka. Dapače, sama priprema teksta je znatan problem kao što je već prikazano, a potencijalna jezična obrada će se koristiti metodama razvijenim u području procesiranja prirodnog jezika (eng. *natural language processing* ili NLP) s kojim scientometrija tradicionalno nema preklapanja. Ipak, mogućnost uključivanja samog predmeta promatranja u metodološki postupak djeluje dovoljno važnom za početak rada u ovom smjeru.

U ovom istraživanju pristupilo se eksperimentalnoj izradi korpusa znanstvenih tekstova i to iz formata u kojem su najčešće dostupni što predstavlja preduvjet za provedbu ovakvih istraživanja. Dodatno, kao mogućnosti koje ovakav korpus predstavlja postavljene su:

1. analiza strukture članka
2. analiza citata kako su pozicionirani u tekst kroz navode
3. ekstrakcija informacija (npr. ekstrakcija imenovanih entiteta; eng. *named entity extraction*) i jezična analiza

Analiza strukture članka uključuje prepoznavanje glavnih dijelova članka, poput uvoda, metode, rezultata i rasprave te zaključka. Uz to, uključuje prepoznavanje posebnih dijelova znanstvenih radova poput tablica i slika. Analiza citata koja ne koristi samo popis literature već i navode u tekstu se nameće kao jasan dodatak citatnim pokazateljima. Korištenje informacija iz samih teksta tiče se tzv. ekstrakcije informacija (eng. *information extraction*) ili

[11] Sam *online* pristup na razini članka radije nego brojevima časopisa donosi i nove pokazatelje korištenja publikacije nakon objave (od kojih su u tisku, na razini članka, dostupni samo citati). Riječ je o pokazateljima poput broja pristupa ili preuzimanja teksta. S obzirom da isti nisu dostupni za radove u časopisu *Scientometrics* i opseg problematike isti nisu prikazani u ovom radu. Zanimljivo ih je, međutim, spomenuti jer se radi o procjeni članaka na razini upravo članaka dok se u tradicionalnoj scientometriji radovi često, katkad implicitno, procjenjuju kroz časopis u kojem su objavljeni.

rudarenja podataka iz teksta (eng. *text mining*). U sklopu ovog istraživanja, u rezultatima su iskroštene prve dvije kategorije, a jezična analiza je ostavljena za buduća istraživanja.

Analiza strukture članka uključuje prepoznavanje glavnih dijelova članka, poput uvoda, metode, rezultata i rasprave te zaključka. Uz to, uključuje prepoznavanje posebnih dijelova znanstvenih radova poput tablica i slika. Analiza citata koja ne koristi samo popis literature već i navode u tekstu, nameće se kao jasan dodatak citatnim pokazateljima. Korištenje informacija iz samih tekstova ovisno je od tzv. ekstrakcije informacija (eng. *information extraction*) ili rudarenja podataka iz teksta (eng. *text mining*). U sklopu ovog istraživanja, u rezultatima su iskorištene prve dvije kategorije, a jezična analiza zbog kompleksnosti ostavljena je za buduća istraživanja.

3 REZULTATI I RASPRAVA

Za dobivanje što cjelovitije slike o značenju časopisa *Scientometrics* slijede rezultati provedenih scientometrijskih analiza časopisa od početka izlaženja 1978. godine do 2010. godine, uključujući citiranost i citatne analize praćene zaključno s 2012. godinom. Razlika u analiziranim godinama je obrazložena u poglavlju o metodologiji.

Većina analiza, rađena je na uzorku znanstvenih članka objavljenih u časopisu *Scientometrics* u 33-godišnjem razdoblju. Analiza radova koji su citirali radove u časopisu *Scientometrics* zahvaća 35 godina kako bi članci objavljeni u časopisu *Scientometrics* 2010. godine imali šansu biti citirani. S obzirom da je naglasak u analizama postavljen na longitudinalan opis razvoja područja, primarna podjela koja je odabrana za prikaz podataka bila je podjela na tri razdoblja. Ovakva podjela zahvaća otprilike tri desetljeća razvoja (točnije 11-godišnja razdoblja) scientometrije, što ih čini idealnim za interpretaciju i prikaz podataka. Dulja razdoblja manje su osjetljiva na specijalne slučajeve poput posebnih izdanja, što je primjereno za većinu analiza koje nastoje longitudinalno prikazati razvoj.

S obzirom da se istraživanje koristi većim brojem scientometrijskih postupaka, rezultati i rasprava su prikazani zajedno u tematskim cjelinama koje odgovaraju scientometrijskoj metodologiji.

3.1 Radovi u časopisu *Scientometrics* 1978-2012

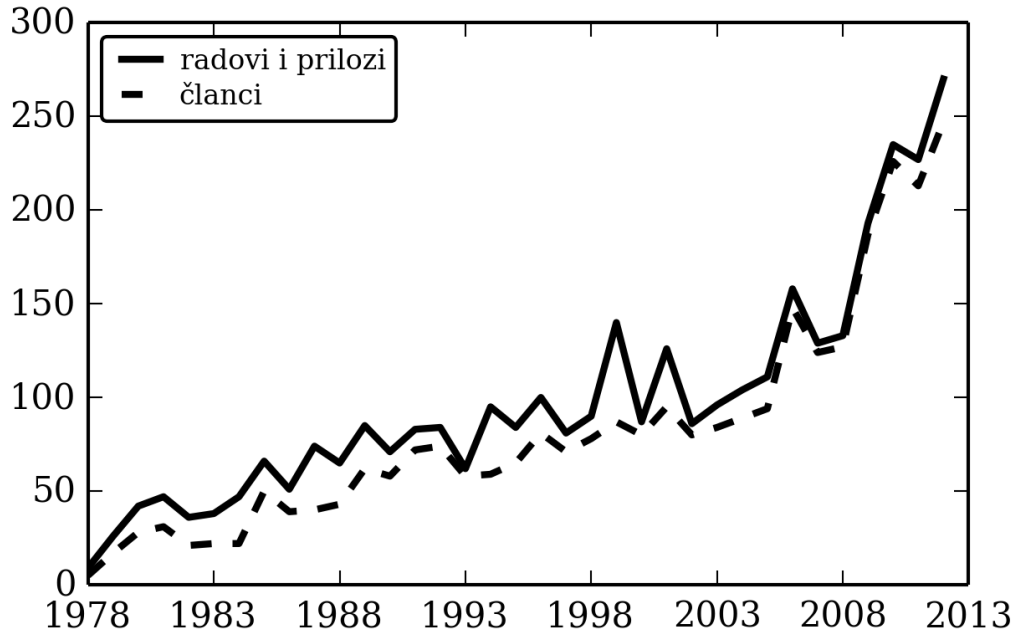
Časopis *Scientometrics* kontinuirano izlazi od 1978. godine, s tim da je od 2000-te godine zapažena intenzivnija produkcija, kako brojem volumena tako i porastom u broju radova. Od početka objavljivanja do 1984. godine časopis objavljuje jedan volumen godišnje, od 1985. do 1990. dva, od 1991. do 2004. tri, a od 2005. do danas četiri volumena godišnje. Do 1984. časopis *Scientometrics* objavljuje šest brojeva po volumenu, od 1985. do 1990. brojevi su konzistentno u formi dvobroja, a od 1990. nadalje, časopis objavljuje tri broja po volumenu. Pregled osnovnih brojeva vidljiv je iz tablice 3. Podaci o radovima iz 2011. i 2012. godine su prikazan kako bi se omogućio uvid u nastavak rasta broja radova u časopisu *Scientometrics* u te dvije godine, koje u scientometrijskim analizama radova, izuzev citatnih analiza, nisu korištene.

Tablica 3. Pregled objavljivanja svih radova u časopisu *Scientometrics* 1978-2012

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010	2011-2012
n radova	501	975	1458	2934	497
n članaka	325	791	1353	2469	460
n teorijskih članaka	80	141	158	379	0
n metodoloških članaka	82	211	409	702	0
n primijenjenih članaka	163	439	786	1388	0
n autora	394	861	1773	2690	888

Dinamika objavljivanja radova i priloga kao i broja članaka prema godinama objave vidljiva je na slici 13. Radi detaljnije slike, grafikoni koji prikazuju vremensku dinamiku prikazuju informacije za svaku godinu zasebno.

Slika 13. Broj svih radova i članaka po godinama objave



Kao što se iz prikazanih podataka može vidjeti, broj radova i priloga nastalih u časopisu *Scientometrics* kroz promatrane vremenske periode pokazuje izraziti trend rasta. Postotak rasta je u drugom razdoblju (1989-1999) u odnosu na prvo (1978-1988) bio 94.6%, a u trećem razdoblju (2000-2010) u odnosu na drugo 71%. Krivulja porasta broja članaka s vremenom postaje sličnija krivulji broja rasta svih radova i priloga. O kojim se sve vrstama radova radilo, vidljivo je u tablici 4. S obzirom na velik stupanj neslaganja između WoS i Scopus podataka, ulazni podaci za ovu klasifikaciju su dodijeljeni ručno i to uvidom u tekstove radova.

Dobiveni i prikazani rezultati upućuju na vrlo dinamičan razvoj znanstvene discipline praćene indikatorom broja radova u časopisu koji je najreprezentativniji predstavnik znanstvene discipline.

Tablica 4. Broj radova prema vrsti i godini objave

vrsta rada	1978-1988	1989-1999	2000-2010	1978-2010
članak	318	765	1334	2417
kratki članak	7	26	19	52
bibliografija	16	13	4	33
recenzija knjige	37	12	6	55
korespondencija	26	47	26	99
podatkovni izvještaj	8	16	3	27
uredničk tekst	10	18	27	55
novosti	63	18	11	92
ispravak	2	4	2	8
sažetak sastanka	0	3	2	5
ostalo	14	53	24	91

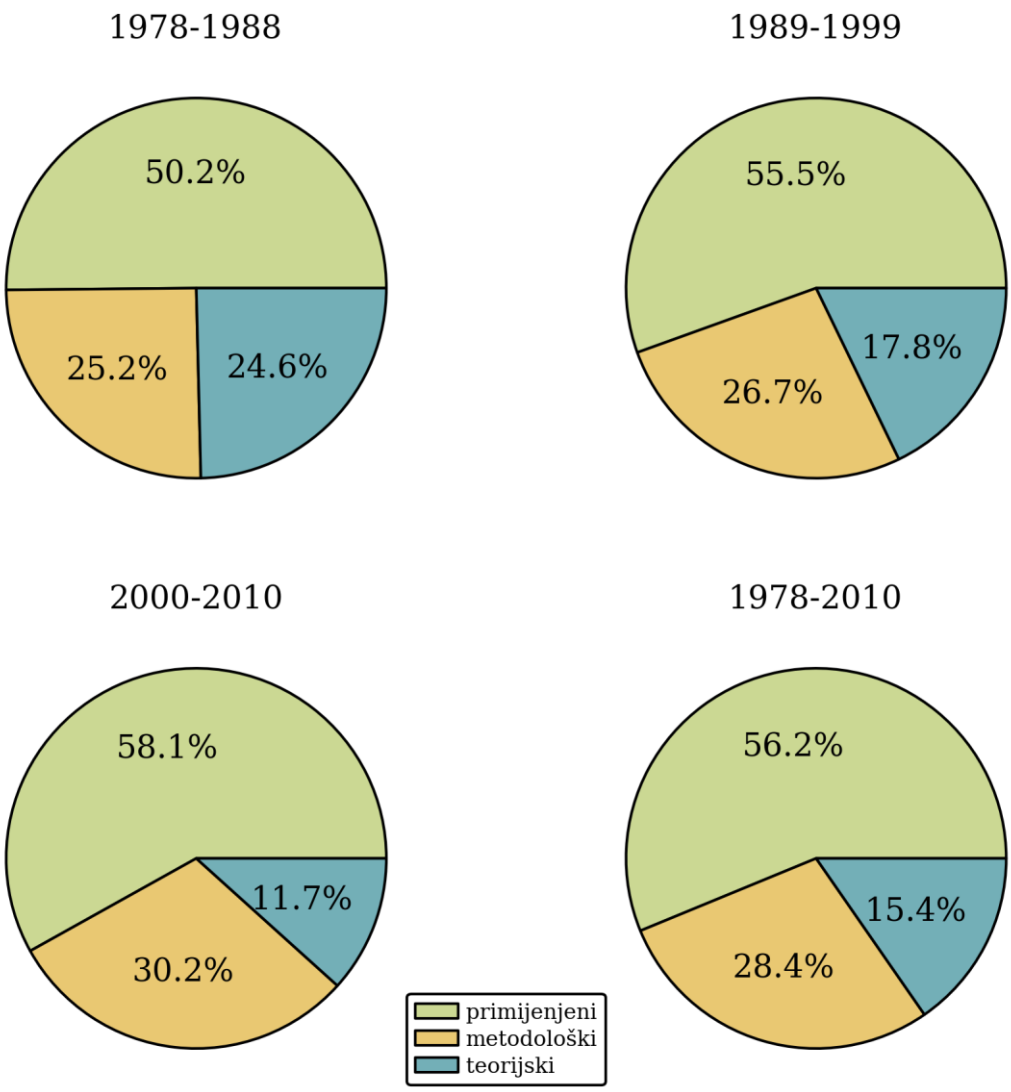
Kao što vidimo iz tablice postoji velik broj relativno slabo zastupljenih kategorija. Neke od ovih kategorija nestaju s vremenom, što je također jedan od pokazatelja razvoja discipline. Podatkovni izvještaji, na primjer, informatizacijom gube na važnosti kako ovakvi agregatni podaci postaju dostupniji kroz *online* baze. Slično vrijedi i za bibliografije, koje svoju funkciju gube razvojem tehnologije i dostupnosti i primarnih i sekundarnih izvora znanstvenih informacija. Slična se primjedba odnosi na priloge tipa novosti. Prije pojave weba, novosti u znanstvenim časopisima bile su nezaobilazan oblik informiranja znanstvene zajednice.

Kao šira kategorija "članaka" preuzeti su svi članci i kratki članci. Razlog uključivanju kratkih članaka uz članke je bio što je uvidom u tekst radova zaključeno da prenose istu vrstu sadržaja kao i tekstovi kategorizirani kao članci, te da je "kratkoća" kratkih članaka u odnosu na članke nekonzistentna. Postoje i takvi kratki članci koji su dulji od članaka, ali koji nisu takvima označeni. Uz uvid u informaciju da je kratkih članaka u časopisu *Scientometrics* relativno malo, odlučeno je i ove radove tretirati kao članke. Iz tablice 4 razvidno je da se broj članaka praćen kroz tri razdoblja približno udvostručuje u odnosu na prethodno razdoblje.

Za razvoj znanstvene discipline osim brojčanih pokazatelja porasta broja i vrste radova, vrlo značajan pokazatelj je i vrsta radova s obzirom na aspekt istraživanja te discipline. Ono što neku disciplinu čini novom znanstvenom disciplinom jest metodološki pristup i instrumentarij. Egzaktnom je čini instrumentarij za mjerenje. Stoga porast broja metodoloških radova ima indikativu ulogu. Naši rezultati (tablica 3) govore u prilog porastu broja metodoloških radova, što je u korelaciji s ukupnim porastom broja radova u časopisu *Scientometrics*. Porast je približno dvostruko veći za svako od naredna tri razdoblja.

Ove rezultate potvrđuju rezultati prikazani na slici 14, pri čemu je udio metodoloških radova u odnosu na ostale radove u ukupnom uzorku 28,3%. Pored metodoloških radova za razvoj znanstvene discipline nezaobilazni su i teorijski radovi, koji izravno ili neizravno mogu utjecati i na nastanak metodoloških radova. U dobivenim rezultatima broj teorijskih radova ne prati dinamiku pojave metodoloških i empirijskih radova. Značajniji rast je zabilježen između prvog i drugog razdoblja života časopisa *Scientometrics*, od 80 na 141 rada. Međutim u posljednjem razdoblju, 2000-2010, taj broj je bio 158 (tablica 1). Ukupnom uzorku (slika 14) udio teorijskih radova bio je 15,4%. Na određeni način logičnim se može opravdati da je u prvom životnom razdoblju časopisa Scientomerics bilo proporcionalno najviše teorijskih radova, 24,6% u odnosu na sve radove (slika 14), jer je to vrijeme osnutka discipline. Osim toga u tom vremenu je bila i najveća koncentracija teoretičara scientometrije.

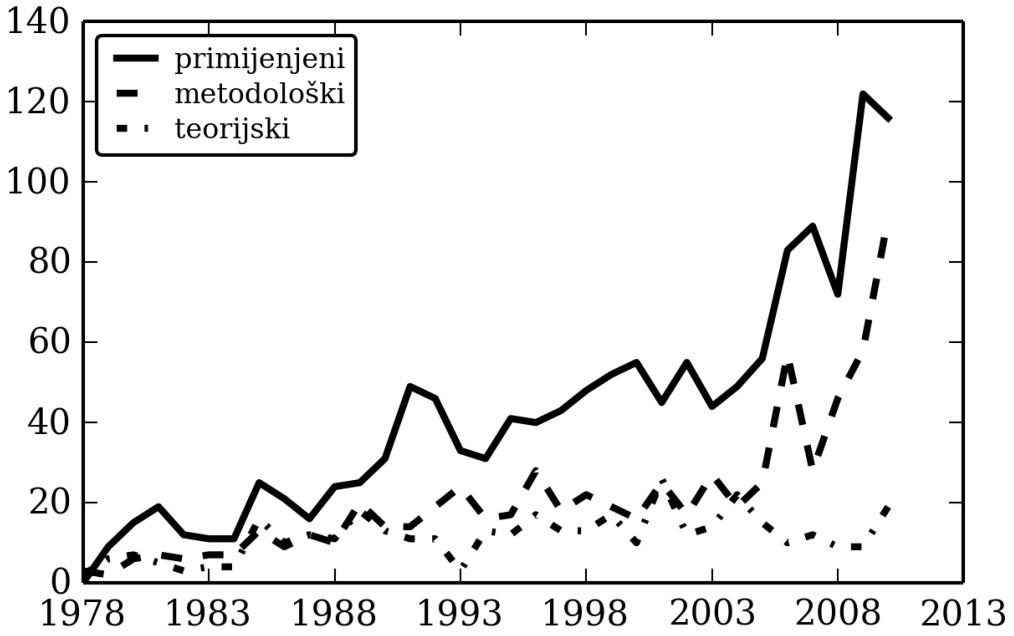
Slika 14. Udijeli radova po tematici za sva vremenska razdoblja



Kao što je sa slike vidljivo, broj primijenjenih radova pokazuje konzistentan udio u svim vremenskim periodima i iznosi oko 50% članaka objavljenih u časopisu *Scientometrics*. Ovaj udio, uz informaciju o porastu broja radova, ukazuje na kontinuiranu i sve češću aplikaciju scientometrijskih metoda u svrhu istraživanja znanosti. S druge strane ukazuje i na važan status časopisa *Scientometrics* u razvoju scientometrije kao aplikativne discipline. Ta aplikativna strana scientometrije izuzetno je važna naročito s aspekta znanstvene politike, ali i praćenja razvoja drugih znanstvenih polja i znanosti općenito.

S obzirom na visoku stopu rasta broja radova u časopisu *Scientometrics*, zanimljivo je vidjeti i dinamiku objavljivanja članaka prema broju primijenjenih, metodoloških i teorijskih radova. Dinamika objavljivanja članaka u odnosu na tematiku prikazana je na slici 15.

Slika 15. Tematika članaka po godinama objave



Kao što vidimo, visoka stopa porasta u broju radova se odnosi na primijenjene i metodološke članke. Prikazom rezultata u kraćim razdobljima, petogodišnjim, dobiva se malo sofisticiranija slika dinamike porasta pojedinih vrsta radova. Uočljivo je da je broj teorijskih članaka, više-manje konzistentan u cijelom promatranom periodu. Porast u broju metodoloških članaka kao i primijenjenih istraživanja gotovo se sigurno u velikoj mjeri može pripisati informatizaciji odnosno sve dostupnijim podacima kao i sve većim mogućnostima računalne obrade podataka.

Opisani odnos teorijskog i metodološkog razvoja scientometrije, kao što je na slikama gore prikazan kroz radove, nije prošao neprimijećen. 1994 godine Glänzel i Schoepflin (1994) u radu naslovljenom "Little scientometrics, big scientometrics ... and beyond?" navode da je područje scientometrije u krizi iako rapidno raste zajedno s interesom za scientometrijske pokazatelje. Među razlozima za "krizu" navode pomak u literaturi s temeljnih istraživanja na primijenjena. Ovaj rad je dobio najviše replika relevantne znanstvene zajednice od svih radova u časopisu *Scientometrics*. Replike ukazuju na činjenice koje ublažavaju ideju "krize",

poput sve sofisticiranijih postupaka u novijim istraživanjima, ali se u načelu slažu oko potrebe rješavanja temeljnih metodoloških pitanja poput, u uvodu opisane, teorije citiranja.

Deset godina kasnije, van Raan (2005) opisuje razvoj scientometrije sa *plus ça change, plus c'est la même chose*[12]. Prema istom radu, ono što se promijenilo od 70-ih, odnosno od formalizacije scientometrije kao zasebne tematike, jest napredak u primijenjenim indikatorima, posebno onima koji se tiču vrednovanja znanstvenog rada. Također, u odnosu na 70-e, dostupnost podataka i mogućnosti obrade su daleko veće. Neka od osnovnih pitanja u scientometriji, međutim, ostaju prisutna. Pitanje primjerenost i točne upotrebe citatnih analiza u svrhu vrednovanja znanstvenog rada, na primjer, ostaje otvoreno i kontinuirano polemizirano pitanje. 2012. godine, Francis Narin (2012) iskazuje slično mišljenje.

S druge strane, bez obzira na otvorena pitanja ne treba zaboraviti da sve sofisticiraniji postupci kao i sve dostupniji podaci, kontinuirano otvaraju nove mogućnosti za razborito korištenje scientometrije u evaluativne ili druge svrhe. Također, kao što van Raan spominje (2005), otvorena pitanja su manje opasna od jedne pomodne teorije na koju se svi pretplaćuju.

3.2 Autorstvo, autori i suradnja u časopisu *Scientometrics* 1978-2010

Produktivnost autora kao i suradništvo kroz koautorstvo mogu se pronaći u istraživanjima koja se sporadično provode od početka 20-og stoljeća. Neka od ovih istraživanja su već opisana u uvodu. Što se scientometrije i časopisa *Scientometrics* tiče, istraživanja produktivnosti autora i suradnje putem koautorstva su zastupljena od samog početka i objavljuju se u članku od tri dijela s glavnim naslovom "*Studies in scientific collaboration*" u prva tri broja časopisa (Beaver i Rosen, 1978; Beaver i Rosen, 1979; Beaver i Rosen, 1979).

Nakon kratkog opisa problematike, ovo poglavlje prikazuje podatke o autorstvu radova, produktivnosti autora te suradnji autora s osvrtom i na međunarodnu suradnju među ustanovama autora.

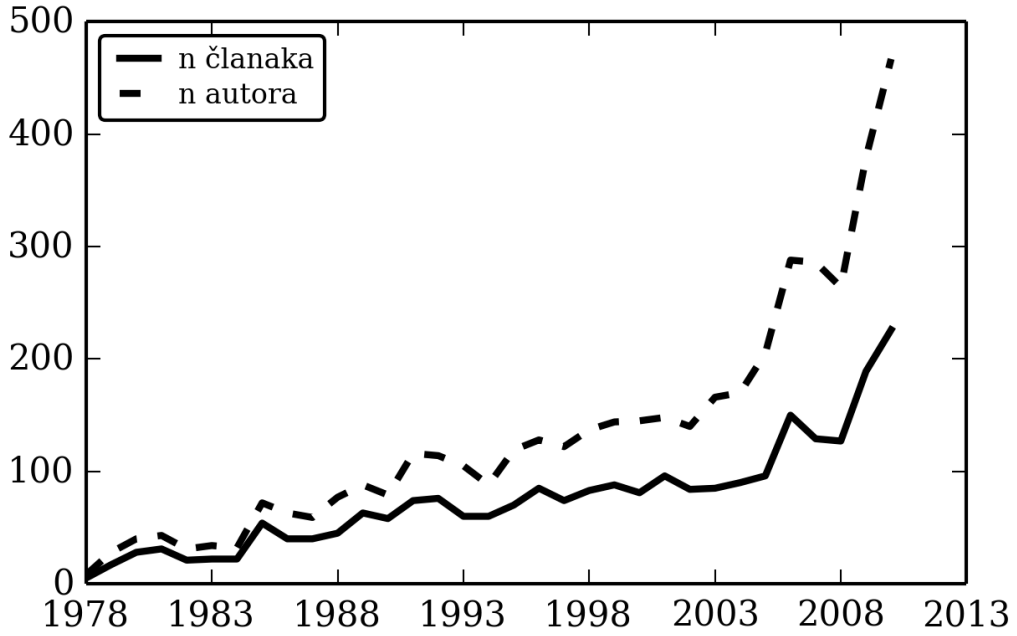
Informacije o autorima su, u bibliometrijskom smislu, agregacije informacija o radovima. Ove informacije poput onih o dobi, spolu i području rada znanstvenika se u istraživanjima o znanosti, naravno, često nadopunjuju informacijama iz drugih izvora poput upitnika ili

[12] u slobodnom prijevodu: čim se više mijenja tim je više ista stvar

intervjua. U ovom istraživanju, informacije o autorima su poznate iz publikacija te ih se promatra isključivo iz konteksta proučavanja scientometrijske literature.

Kako bi se umanjio šum u prebrojavanju i mjerenju suradnje i kako bi podaci kasnije bili iskoristivi za analizu citiranosti, podaci preneseni u ovom poglavlju odnose se samo na članke objavljene u časopisu *Scientometrics* za period od 1978. do 2010. godine. Broj autora koji su objavljivali članke u časopisu scientometrics kao i broj objavljenih članaka prikazani su na slici 16.

Slika 16. Broj članaka i autora članaka po godinama objave



Kao što vidimo, broj autora u časopisu *Scientometrics* u zadnjem razdoblju pokazuje veću stopu rasta od broja članaka. Porast broja članaka u trećem razdoblju u odnosu na drugo bio je 71%, a porast broja autora 118%. Navedeno nam govori da dolazi do porasta broja autora na člancima koji se objavljuju u časopisu *Scientometrics*.

3.2.1 Autorstvo članaka u časopisu *Scientometrics*

Prije no što se posvetimo samim autorima koji su objavljivali u časopisu *Scientometrics*, korisno je proučiti radove u odnosu na broj autora zbog razumijevanja trendova višeautorstva scientometrijskih radova. Za potrebe dizajna pokazatelja i interpretacije rezultata, područje u kojem je uobičajeno da radovi imaju mali broj autora treba tretirati drugačije od područja za

koje je tipično da radove potpisuje velik broj autora. Dapače, broj autora može rasti do razina na kojim je upitno da li se još uvijek radi o autorstvu.

Na primjer, u nekim područjima društvenih znanosti poput sociologije tipično je autorstvo do tri autora, s velikim udjelom jednoautorskih radova (De Haan, 1997). Na člancima koje su hrvatski informacijski znanstvenici objavili u časopisima indeksiranim u WoS-u 1991-2005, uobičajen broj autora na radovima je bio 3 (Jokić, 2005). U psihologiji, pak, nije neuobičajeno pronaći više od deset autora na radu i to s porastom broja autora kako se tematika bliži području medicine (Letina et al., 2012). Očekivano, u medicini je tipičan broj autora u prosjeku znatno veći nego u većini područja društvenih znanosti, a jednoautorski radovi su rjeđi (Shaban i Aw, 2009). Dapače, u medicinskim istraživanjima nalazimo i radove s više stotina autora. Situacija kulminira u nekim područjima fizike gdje se pronalaze radovi s više stotina autora koji su često potpisani abecedno i s dodatnom naznakom autora s kojim se komunicira i koji se može smatrati na određeni način glavnim autorom.

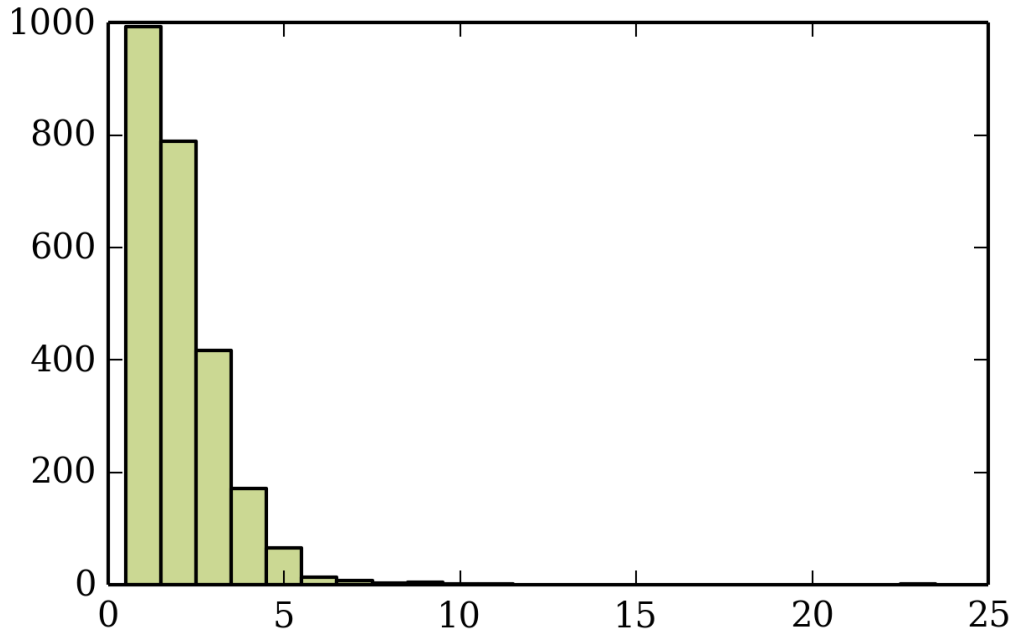
Informacija o broju autora u nekom skupu radova je, dakle, vrlo relevantna za interpretaciju podataka o produktivnosti autora kao i odabir i dizajn strategija brojanja (što je detaljnije opisano u metodologiji). Dobiveni rezultati o autorstvu članaka u časopisu *Scientometrics* prema razdobljima, prikazani su u tablici 5.

Tablica 5. Autorstvo članaka objavljenih u časopisu *Scientometrics* 1978-2010 u odnosu na razdoblje

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
n radova	325	791	1353	2469
n jednoautorskih radova	183	382	428	993
n višeautorskh radova	142	409	925	1476
n dvoautorskih radova	99	258	432	789
n troautorskih radova	29	95	293	417
n radova s 4 do 10 autora	14	56	198	268
n radova s preko 10 autora	0	0	2	2
prosječno autora po radu	1.63	1.81	2.31	2.06
medijan autora po radu	1	2	2	2
max autora po radu	7	10	23	23

Iako prosječan broj autora po radu raste, članci u *Scientometrics* pokazuju značajke "normalnog" autorstva (za razliku od "super-autorstva" kako je opisano u metodologiji). Radi dobivanja detaljnijeg uvida, distribucija broju autora po radovima je prikazana na slici 17.

Slika 17. Distribucija radova po broju autora

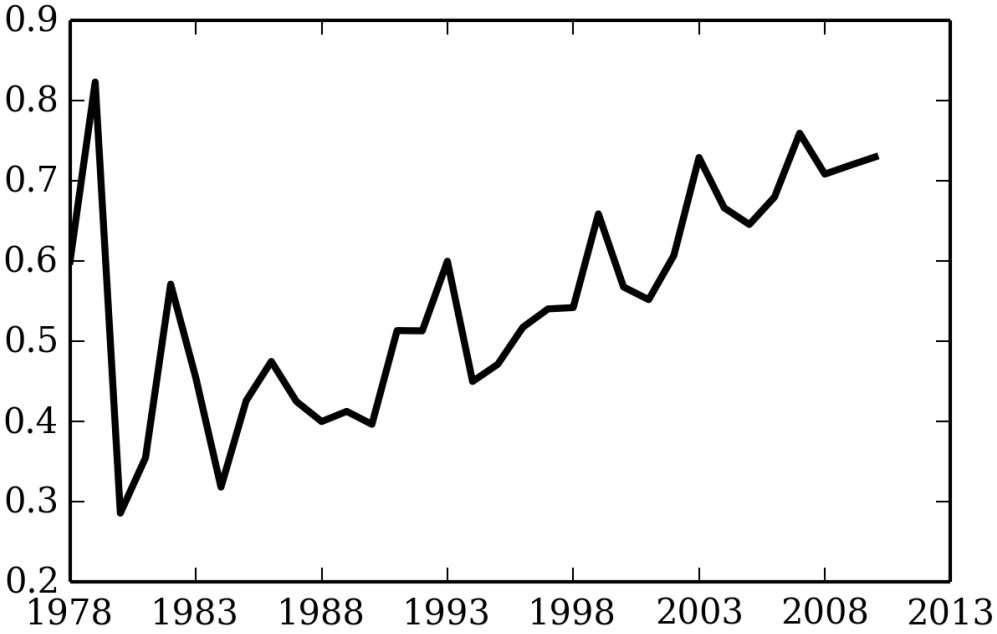


Prema slici 17 i tablici 5, gotovo svi članci (98,7%) imali su pet ili manje autora, a u cijelom promatranom razdoblju postoji samo jedan rad koji se približava super-autorstvu, i to s 23 autora. Navedeno prikazuje obrasce autorstva sličnije društvenim znanostima (vidi opise

područja sociologije, informacijskih znanosti i psihologije gore) nego većini STM područja. Kao što je opisano u metodologiji, ova informacija je provjerena na početku istraživanja i korištena je pri odabiru metode brojanja. Za razliku od produktivnosti autora, podatke o broju autora po radu lako je provjeriti već u postupku pripreme podataka jer izračun ne zahtijeva pripremu imena autora.

Kao što je vidljivo iz tablice 5, apsolutni broj jednoautorskih i višeautorskih radova je rastao kroz vrijeme, ali je trend rasta višeautorskih radova bio izraženiji. Rast udjela višeautorskih radova kroz godine je vidljiv na slici 18.

Slika 18. Proporcija višeautorskih članaka po godinama objave



Na početku razdoblja izlaženja časopisa udio višeautorskih radova bio je oko 40%. Uočljiv visoki postotak (82%) višeautorskih radova u prvoj godini objavljivanja može se pripisati činjenici da se radi o udjelu od vrlo malog broja članaka (5). Broj članaka u ostalim godinama je višestruko veći pa se može smatrati mjerodavnijim. Na kraju praćenog razdoblja oko 70% članaka u časopisu *Scientometrics* su višeautorski.

Navedeno ukazuje na očekivan porast u prosječnom broju autora po radu kroz vrijeme. Broj višeautorskih radova kao i prosječan broj autora po radu se, naime, povećava u većini znanstvenih disciplina (prema Beaver i Rosen, 1979). Iako udio višeautorskih radova kontinuirano raste, Abt (2007) predviđa da trend rasta nije takav kako bi ukazivao da će jednoautorski radovi nestati.

Osim po razdobljima, zanimljivo je i pogledati da li postoje razlike u autorstvu u odnosu na tematiku članka. Podaci iz kojih je navedeno vidljivo, prikazani su u tablici 6.

Tablica 6. Autorstvo članaka objavljenih u časopisu *Scientometrics* 1978-2010 u odnosu na tematiku

pokazatelj	primijenjeni	metodološki	teorijski
n radova	1388	702	379
n jednoautorskih radova	501	279	213
n višeautorskh radova	887	423	166
n dvoautorskih radova	466	223	100
n troautorskih radova	249	124	44
n radova s 4 do 10 autora	170	76	22
n radova s preko 10 autora	2	0	0
prosječno autora po radu	2.16	2.07	1.69
medijan autora po radu	2.0	2.0	1
max autora po radu	23	9	6

Podaci da li postoje razlike u autorstvu među klasifikacijom članaka s obzirom na tematiku, odnosno da li su teorijski, metodološki ili empirijska istraživanja, prilično su važni za razumijevanje razvoja nove znanstvene discipline. Teorijski članci, očekivano, imali su prosječan i maksimalan broj autora nešto manji u odnosu na ostale vrste članaka. Naime, teorije i u ostalim znanstvenim poljima najčešće kreiraju pojedinci ili manji broj znanstvenika. U ovom istraživanju jedino su teorijski radovi imali medijan 1 autora po radu. Metodološka, a pogotovo primijenjena istraživanja, po svojoj naravi imaju veću potrebu za ekstenzivnim radom na podacima od teorijskih, što može objasniti češću suradnju u navedenim istraživanjima u odnosu na teorijski rad. Na temelju iskustva u radu na scientiometrijskim istraživanjima, opsežnija ne samo makro nego i mikro istraživanja, zahtijevaju timski rad jer se radi o velikim skupovima različitih podataka koji zahtijevaju specifična i nova znanja u baratanju podacima, njihovoj obradi i interpretaciji.

3.2.2 Produktivnost autora članaka u časopisu *Scientometrics*

Nakon pregleda osnovnih rezultata o autorstvu i vrsti brojanja možemo pristupiti proučavanju broja članaka u časopisu *Scientometrics* po autorima, odnosno proučavanju produktivnosti autora u ovom časopisu. Pregled podataka o produktivnosti prikazan je u tablici 7.

Tablica 7. Pregled produktivnosti autora u časopisu *Scientometrics* 1978-2010

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
n radova	325	791	1353	2469
n autora	350	807	1755	2595
prosječno članaka po autoru	1.51	1.77	1.78	1.96
medijan članaka po autoru	1.0	1	1	1
maks članaka po autoru	11	16	48	67

Kao što vidimo iz tablice 7, najveći broj autora objavio je samo jedan rad u 33-godišnjem razdoblju, za sva razdoblja medijan je bio 1, a u cijelom promatranom skupu udio autora koji su objavili samo jedan članak je 69,9%. Drugim riječima, prosjek produktivnosti je nizak, a varijanca je visoka (1 autor je u čitavom periodu objavio 67 članaka). Dobiveni podaci, iako vrlo izraženi čak i u odnosu na očekivanu krivulju produktivnosti, nisu posve iznenađujući radi specifičnosti područja istraživanja scientometrije. Na temelju podataka o broju radova i autora koji su se bavili empirijskim istraživanjima, za očekivati je da će u časopisu *Scientometrics* objavljivati znanstvenici koji se primarno bave nekim drugim područjem, no koji su istražili literaturu ili aktivnosti vlastite discipline kvantitativnim metodama. Ukoliko, međutim, želimo sagledati scientometriju kao disciplinu, za analizu iste nam je najvažnija skupina autora koja se bavi scientometrijom kao primarnim područjem.

S obzirom na značajke dobivene distribucije produktivnosti autora u časopisu *Scientometrics*, vrlo ju je teško bilo pregledno grafički prikazati. Distribucija je stoga prikazana tablicom 8.

Tablica 8. Produktivnost autora s obzirom na broj članaka objavljenih u časopisu *Scientometrics* 1978-2010

n članaka po autoru	n autora	n članaka
1-1	1813	1080
2-5	661	1273
6-10	75	473
11-20	30	395
21-30	10	233
31-5000	6	243

Prikazana distribucija produktivnosti ima značajke koje se konzistentno dobivaju u brojnim istraživanjima provedenim u nekoliko zadnjih desetljeća u različitim poljima znanosti. Spomenut fenomen je već opisan u uvodu kao Lotkin zakon, a značajke distribucije možemo sumirati kao: a) niska prosječna produktivnost i b) velike varijacije u produktivnosti među znanstvenicima.

Prema navedenom, za većinu radova je odgovoran manji broj znanstvenika. Prema tablici 8, postoji 6 autora s više od 30 članaka i sveukupno su objavili 243 članaka, odnosno tih 6 autora bilo odgovorno za 6,6% članaka. U skupu je prisutno 52 autora (2% od svih autora članaka) koji su objavili više od deset članaka i ovi autori su zaslužni za 1004 članaka odnosno 40% svih članaka u promatranom skupu. Da bi ilustrirao učinak znanstvenika s drugog kraja distribucije, odnosno "elitističku" distribuciju produktivnosti u znanosti, Simonton (2004) navodi empirijske podatke Dennisa (1955) i Kyvika (1989): 10% najproduktivnijih je bilo odgovorno za gotovo 50% publikacija, te procjenjuje da se u nekom slučaju publikacije znanstvenika iz donje polovice distribucije po produktivnosti zauvijek izgube, svaka disciplina bi i dalje zadržala 82% svojih publikacija.

Te karakteristike znanstvene produktivnosti se konzistentno dobivaju u istraživanjima. Navedeno stoji bez obzira da li se promatra samo objavljivanje unutar jedne godine, period od pet ili više godina, ili čak čitavi profesionalni vijek; bez obzira na heterogenost uzorka po dobi, disciplini, vrst istraživačkog nacrta (krossekcijski ili longitudinalni), mjeri produktivnosti koja se koristi, itd. Drugim riječima, neovisno o uzorku, operacionalizaciji i nacrtu istraživanja, tipična asimetrična krivulja (u obliku slova L) i visoka varijanca mogu biti u različitoj mjeri izraženi, ali u pravilu će opisati dobivenu distribuciju.

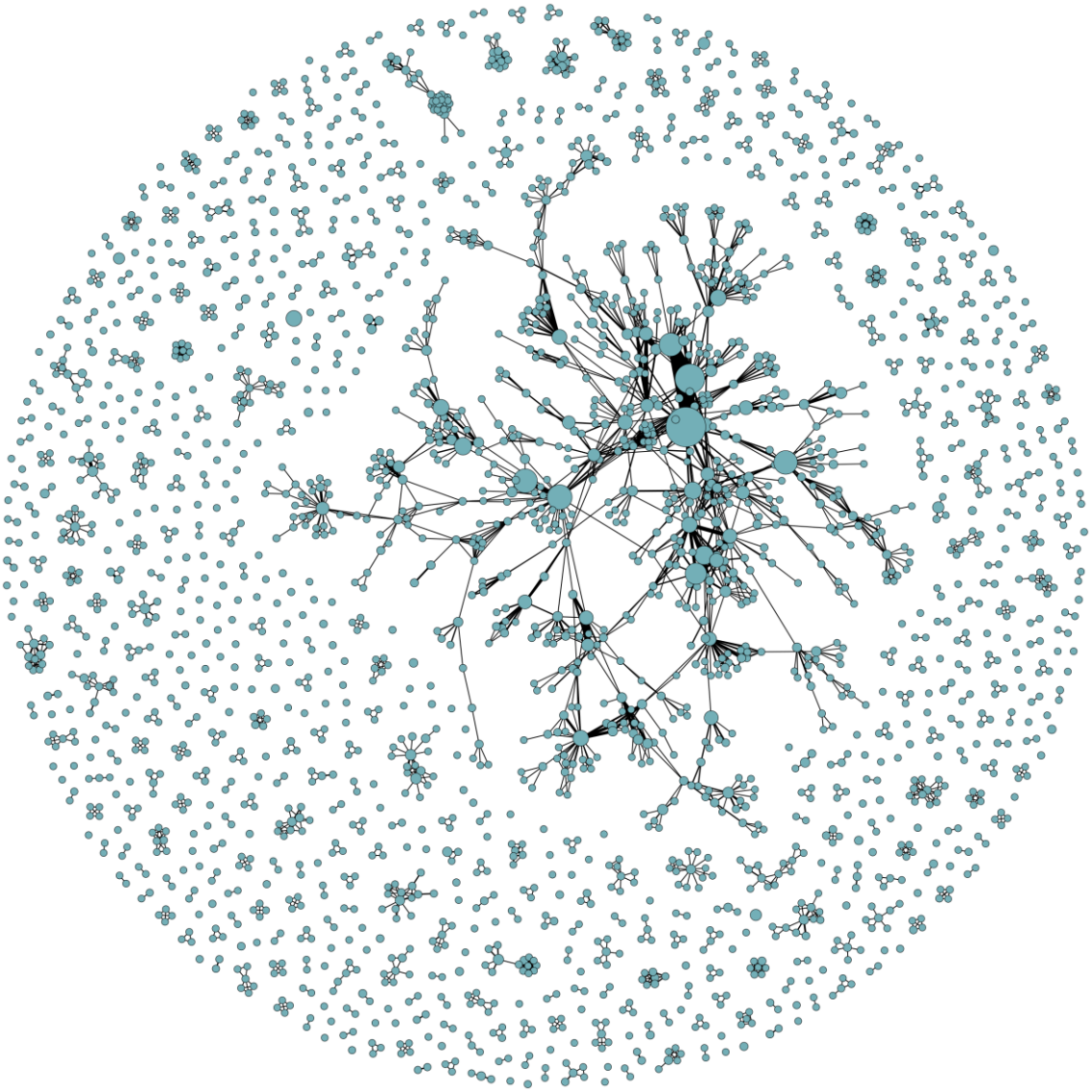
Kao što je već u uvodu napomenuto, u većini istraživanja koja se bave produktivnošću pojedinog znanstvenika, kao mjera produktivnosti koristi se ukupan broj objavljenih radova. Takva mjera ne uključuje druge aktivnosti u kojima znanstvenik može biti produktivan (poput nastavne aktivnosti, aktivnosti vezane uz diseminaciju rezultata i popularizaciju znanosti, itd). U kontekstu istraživanja autorstva, međutim, produktivnost je moguće nadopuniti informacijama o suradnji autora kako bi se upotpunila slika o trendovima autorstva nekog skupa radova i autora.

3.2.3 Suradnja među znanstvenicima na člancima u časopisu *Scientometrics*

Mrežne analize su važna metoda za svaki pokušaj razumijevanja procesa stvaranja i kolanja znanja (Mali et al., 2012). Mreže koautorstva su česta i preporučena metodologija za promatranje suradnje među znanstvenicima kroz objavljene publikacije. Struktura mreže idealna je za obradu informacija koje uključuju definirane relacije između pojedinih članova. U ovom slučaju ta relacija je uniformna (i.e. svaka veza ima isto značenje) i neusmjerena (suradnja između a i b se tretira isto kao suradnja između b i a).

Za potrebe ovog istraživanja izrađena je mreža svih autora članaka. Prikaz cjelokupne mreže autora i koautorstava među njima vidljiv je na slici 19. Slika pokazuje jednu veliku komponentu i mnogo malih komponenti koje se sastoje od manjeg broja ili samo jednog čvora.

Slika 19. Mreža suradnje autora na člancima u časopisu *Scientometrics* 1978-2010



Podaci o koautorstvu na člancima objavljenim u časopisu *Scientometricsu* su iskorišteni kako be se ispitala i relacijska strukturu koautorstva. Da bi ustanovili značajke mreže koautorstva, izračunate su vrijednosti osnovnih globalnih svojstava mreže za čitavi vremenski period. Osnovne mjere i pokazatelji analiza društvenih mreža su prikazani na tablici 9. Spomenute mjere detaljno su opisane u metodologiji.

Tablica 9. Pregled mreže koautorstva na člancima objavljenim u časopisu *Scientometrics* 1978-2010

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
n radova	325	791	1353	2469
u najvećoj komponenti	12 (4%)	108 (14%)	530 (39%)	1015 (41%)
n čvorova	350	807	1755	2595
u najvećoj komponenti	10 (3%)	72 (9%)	467 (27%)	720 (28%)
n veza	258	787	2760	3713
u najvećoj komponenti	20 (8%)	143 (18%)	1015 (37%)	1499 (40%)

Cjelokupna mreža se sastoji od 2 595 autora (čvorova u mreži), među kojima je koautorstvom ostvareno 3 713 veza. Kako je prosječan broj veza po čvoru veći od jedan, teorijski se može očekivati pojavljivanje velike glavne komponente. Najveća grupa povezanih čvorova sadrži 721 autora, odnosno 28% svih autora prisutnih u mreži. Mreža u cjelokupnom promatranom razdoblju ne pokazuje fragmentaciju u više značajnih grupacija što ukazuje na razvoj scientometrije kao zasebnog područja s centralnom skupinom autora.

Postoji veliki broj manjih komponenti različitih veličina. One predstavljaju manje zajednice koautora koje nisu surađivale s autorima u glavnoj komponenti. Većinom se radi o dijadama (n = 207), ali postoji i veliki broj komponenti s više od tri članova (n = 125).

Stupanj centralnosti svakog čvora u mreži jednak je broju direktnih veza koje taj čvor ima s drugima tj. broj prvih susjeda. U mreži koautorstva, ta mjera prikazuje ukupan broj suradnika s kojima je neki autor surađivao. Svaka od veza se temelji na jednom ili više radova, a broj radova se koristi kao težina veza. Težina neke veze u mreži koautorstava pokazuje, dakle, ponovljenu suradnju s nekim autorom što omogućuje detekciju koautorstava značajnih za razvoj literature u području.

3.2.3.1 Indikatori preferencijalnog povezivanja i strukture maloga svijeta

Mreža koautorstva radova objavljenih u *Scientometrics* istovremeno pokazuje svojstva malog svijeta i djelovanje mehanizma preferencijalnog povezivanja. Navedena tvrdnja se temelji na sljedećim pokazateljima:

- *Veličina najveće komponente* - mreža se sastoji od nekoliko komponenti različite veličine, od kojih su neke povezane sa drugima, a neke su izolirane. Budući da s vremenom relativna veličina glavne komponente pokazuje trend rasta, može se

zaključiti da djeluje mehanizam preferencijalnog povezivanja. Međutim, kako je rast najveće komponente kroz vrijeme sporiji i raste broj komponenti, znači da mreža ima i strukturu malog svijeta .

- *Power-law distribucija broja suradnji* - u mreži s takvom distribucijom koja slijedi zakon potencije postoji mali broj čvorova čiji su stupnjevi nekoliko redova veličine veći od prosjeka. Iako prosječni stupanj centralnosti s vremenom raste, njegova distribucija postaje sve više nerazmjerna. Takvi rezultati ukazuju na postojanje mehanizma preferencijalnog povezivanja.
- *Prosječna udaljenost najkraćeg puta* - ukazuje na unutarnju povezanost mreže. Manja udaljenost znači bolju povezanost, i većina velikih mreža ima iznenađujuće nisku udaljenost, otkud i potječe koncept malog svijeta. Unatoč zamjetno velikom rastu mreže, ovaj indikator s vremenom pokazuje vrlo slabi trend rasta, što govori u prilog postojanju strukture malog svijeta.
- *Koeficijent grupiranja* - daje informaciju o tome kolika je vjerojatnost da će suradnici bilo kojeg autora u mreži i sami međusobno surađivati. Visoki koeficijent, kao i njegov rast (ali i sporo opadanje kroz vrijeme) sugerira da se promatrana mreža formira po principu malog svijeta.

Značajke mreže suradnje autora članaka u časopisu *Scientometrics* 1978-2010. prema razdobljima objavljivanja prikazane su u tablici 10.

Tablica 10. Značajke mreže koautorstva na člancima objavljenim u časopisu *Scientometrics* 1978-2010

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
n radova	325	791	1353	2469
n čvorova	350	807	1755	2595
n veza	258	787	2760	3713
gustoća	0.004	0.002	0.002	0.001
dijametar	4	9	14	16
asortativnost	0.542	0.335	0.656	0.503
n artikulacijskih čvorova	32	62	151	247
globalni koeficijent grupiranja	0.32	0.39	0.6	0.53
slučajni globalni koeficijent grupiranja	0.0	0.0	0.001	0.001
prosječna najkraća duljina puta	1.301	1.745	2.705	2.929
slučajna prosječna najkraća duljina puta	9.985	7.537	6.595	7.382

Mjera gustoće za čitavu mrežu je vrlo niska (0,001), što ukazuje da je mreža raspršena. Dijametar za čitavu mrežu je 16, što je relativno niska vrijednost. To je u skladu s fenomenom malog svijeta, prema kojem, unatoč velikom broju čvorova u mreži (odnosno njenoj glavnoj komponenti), prosječna udaljenost između bilo koja dva čvora je očekivano relativno niska. U našem slučaju, prosječna udaljenost iznosi samo oko tri koraka (2,933), što, kao što je rečeno ide u prilog tezi da mreža koautorstva pokazuje strukturu malog svijeta. Usporedbom s rezultatima dobivenim u drugim istraživanjima mreža koautorstva (na drugim uzorcima znanstvenika, vremenskim periodima različite duljine i drugačijim načinom definiranja granica uzorka), dobivena prosječna udaljenost u našem uzorku pripada u raspon nižih vrijednosti. Usporedimo li je s vrijednostima dobivenim na slučajnim mrežama, možemo zaključiti da više nego dvostruko niža, što opet ukazuje na visoku tendenciju povezivanja.

Da bi utvrdili da se mreža razlikuje od slučajne i zaista ima strukturu malog svijeta, dobiveni koeficijent grupiranja treba biti veći od onog koji se dobiva na slučajnim mrežama. Kao što se može vidjeti iz rezultata u tablici 10, koeficijent grupiranja je nekoliko stotina puta veći od očekivanog koeficijenta grupiranja za slučajne mreže.

Ove informacije, međutim, potrebno je interpretirati u skladu s komponentama grafa. Komponente pokazuju važne grupacije i komunikaciju u cijelom promatranom grafu, velik broj nepovezanih komponenti može ostvarivati snažnu komunikaciju unutar svake od komponenti, ali cjelokupno područje može biti rasuto u više komponenti. Ovakva situacija

tipična je za pre-paradigmatsko stanje neke discipline (CITAT). Pojava snažne glavne komponente ukazuje, pak, na povećanje kohezije i konsenzusa unutar područja.

Vrlo visok koeficijent asortativnosti za sve periode pokazuje da postoji tendencija zajedničkog koautorstva autora s obzirom na broj koautora s kojima su surađivali. Drugim riječima, koautori među kojima nastaje veza su vrlo slični s obzirom na ukupni broj svih veza koje imaju. Sve veći broj artikulacijskih čvorova ukazuje na manju strukturalnu kohezivnost mreže (Moody, 2004), ali budući da je broj članova u mreži također rastao, takav rezultat ne znači da su koautori članaka u časopisu *Scientometrics* bili manje povezani.

Detaljniji uvid u koautorske trendove u časopisu *Scientometrics* vidljiv je iz poglavlja u nastavku, a pregled važnih karakteristika vidljiv je iz tablice 11.

Tablica 11. Suradnja na člancima objavljenim u časopisu *Scientometrics* 1978-2010

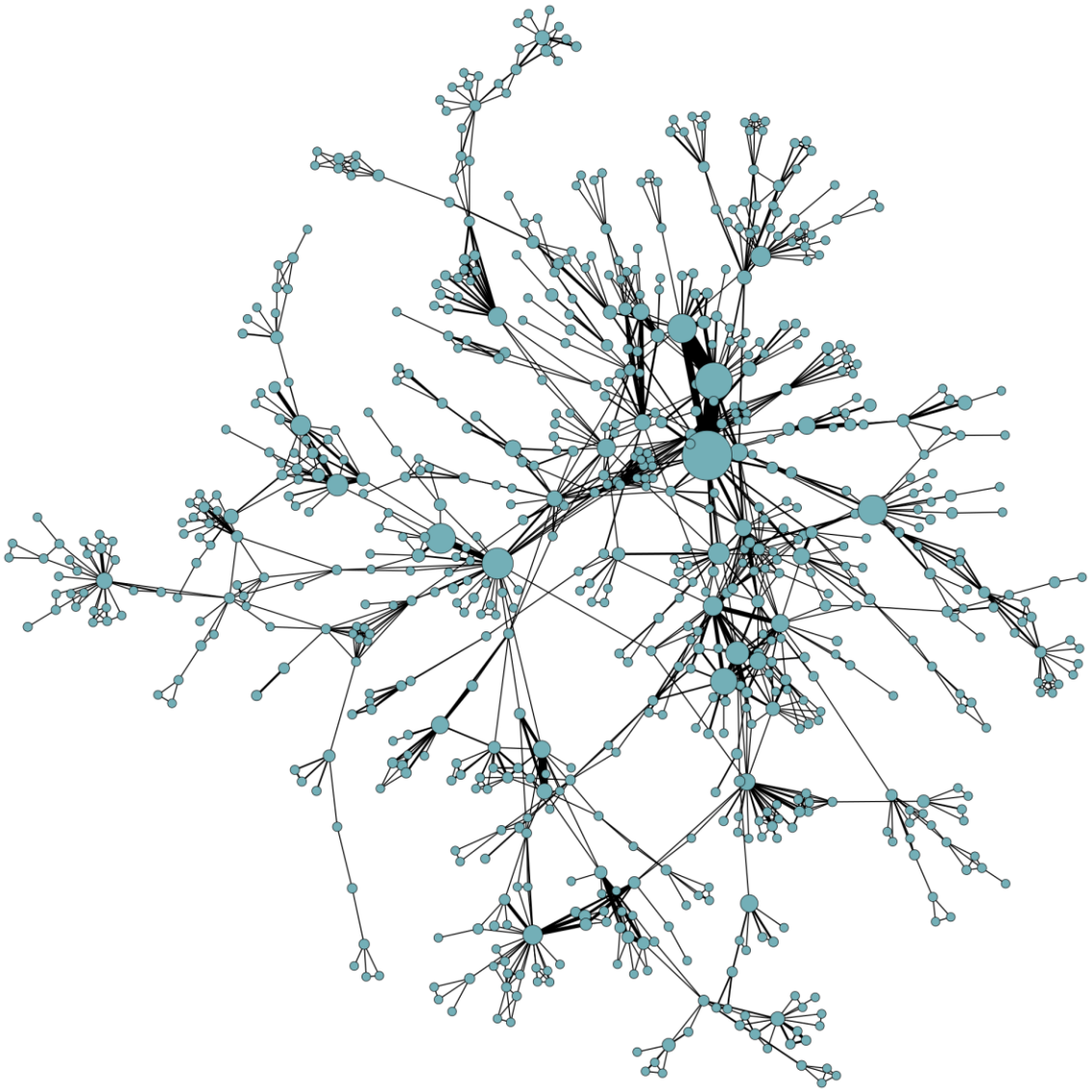
pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
prosječan stupanj centralnosti	1.474	1.95	3.145	2.862
StD stupnja centralnosti	1.513	2.126	3.488	3.391
medijan stupnja centralnosti	1.0	1	2	2
max stupanj centralnosti	7	19	29	37
n izoliranih čvorova	106	170	153	325
n izoliranih dijada	34	97	134	207
n izoliranih trijada	15	32	77	104
n ostalih komponenti (n>3)	23	40	104	127

Prosječan broj veza (suradnika) je u analiziranim desetljećima bio u stalnom porastu, što je osobito izraženo u posljednjem razdoblju (tablica 11). S obzirom na porast standardne devijacije stupnja centralnosti, može se tvrditi da su razlike među autorima prema broju koautora koje su imali bile sve veće. To je očito iz maksimalnih vrijednosti broja koautora: u prvom analiziranom desetljeću je najpovezaniji čvor imao 7 veza, a u zadnjem desetljeću je najpovezaniji čvor imao 37 veza. Takav obrazac ukazuje na pojavljivanje tzv. "zvijezda", tj. na djelovanje mehanizma preferencijalnog povezivanja.

3.2.3.2 Najveća komponenta

Važan koncept u proučavanju razvoja područja putem mreža koautorstva je broj zasebnih nepovezanih komponenti te analiza glavne komponente. Glavna komponenta je prikazana na slici 20.

Slika 20. Glavna komponenta u mreži suradnje autora na člancima u časopisu *Scientometrics* 1978-2010



Značajke glavne komponente suradnje autora članaka objavljenih u časopisu *Scientometrics* kao izdvojene mreže prikazane su u tablici 12.

Tablica 12. Značajke glavne komponente mreže koautorstva na člancima objavljenim u časopisu *Scientometrics* 1978-2010

pokazatelj	n
n radova	1015
n čvorova	720
n veza	1499
gustoća	0.006
dijametar	16
asortativnost	-0.055
n artikulacijskih čvorova	137
prosječan stupanj centralnosti	4.164
globalni koeficijent grupiranja	0.62
slučajni globalni koeficijent grupiranja	0.006
prosječna najkraća duljina puta	6.531
slučajna prosječna najkraća duljina puta	4.74

Ako se analizira samo najveća glavna komponenta, prosječna duljina puta iznosi 6,6, i veća je od očekivane za slučajne mreže (4,7). Takav rezultat ukazuje da kod glavne komponente nije izražena struktura malog svijeta. Naime, kod izračuna prosječne duljine puta na cjelovitoj mreži (sa svim komponentama), zbog većeg broja manjih komponenti (malih izoliranih i unutar sebe relativno dobro povezanih grupa) nužno dolazi do manjih vrijednosti prosječne udaljenosti. Drugim riječima, veliki broj malih nepovezanih komponenti dovodi do vrijednosti prosječne udaljenosti koja je niska te daje iskrivljenu sliku o dobroj povezanosti mreže. Također, mjera asortativnosti pokazuje da ne postoji korelacija u stupnjevima centralnosti između bilo koja dva povezana čvora što govori o nepostojanju linearnog odnosa između broja veza koje imaju bilo koja dva susjedna čvora.

3.2.3.3 Distribucija stupnja centralnosti

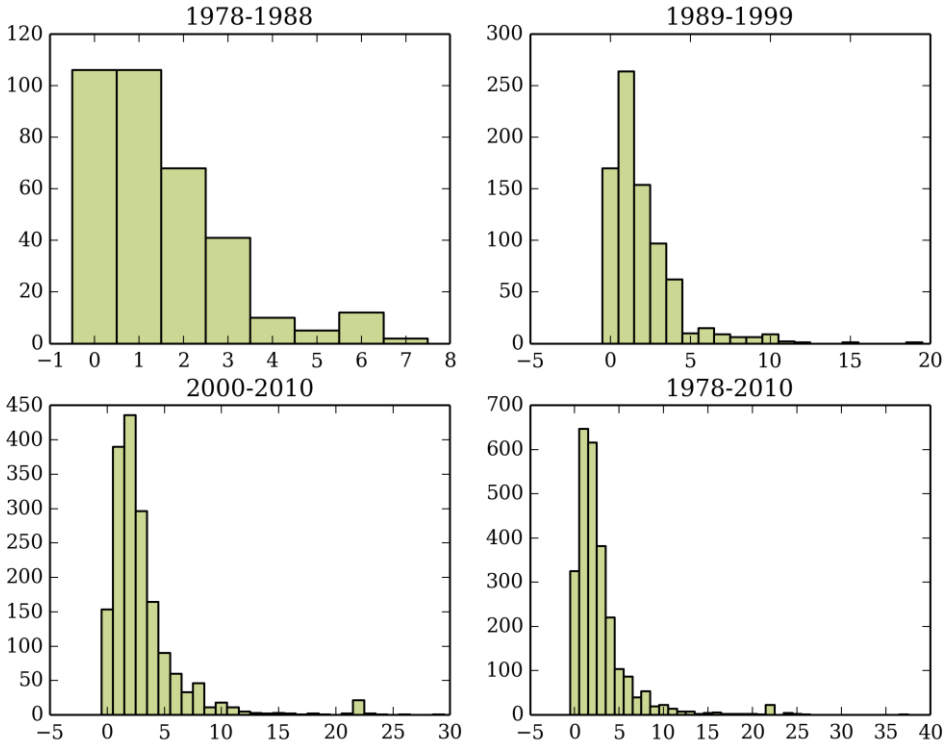
Distribucija broja suradnika je važan pokazatelj strukture mreže koautorstva i procesa koji dovode do te strukture (Milojević, 2010). Budući da je broj suradnika svakog autora u mreži zapravo stupanj centralnosti čvora (autora), distribucija broja suradnika je jednaka distribuciji stupnjeva centralnosti. U prosjeku je svaki autor surađivao s gotovo tri autora (2,9).

Glavni pokazatelj strukture koja nastaje kao rezultat preferencijalnog povezivanja je nerazmjerna distribucija stupnjeva (Kronegger et al., 2011). Mreže kod kojih je distribucija stupnjeva takva da većina čvorova ima manji broj veza, a mali postotak čvorova (tzv. zvijezde

ili koncentratori, poveznici; prema Barabási et al, 2002) imaju nekoliko puta više veza od prosjeka, smatraju se nerazmjernim mrežama i slijede zakon potencije. Prema tome se zaključuje da djeluje mehanizam preferencijalnog povezivanja.

Dobivena distribucija stupnjeva pokazuje tipična odstupanja od zakona potencije (Milojević, 2010). Budući da većina autora ima 2 suradnika, dolazi se do zaključka da distribucija ipak nije nerazmjerna. Usto, broj autora s većim brojem suradnika pokazuje odstupanje od nerazmjerne distribucije. Dakle, iz navedenog proizlazi da se distribucija ne može opisati zakonom potencije, odnosno da se mreža ne može objasniti samo procesom preferencijalnog povezivanja. Slična odstupanja su dobivena i opisana u prijašnjim istraživanjima (Wagner i Leydesdorff, 2005).

Slika 21. Distribucija stupnja centralnosti po autoru



Distribucija stupnja centralnosti u svim promatranim periodima ima tipično asimetričnu krivulju, koja pokazuje da manji broj autora ima velik broj suradnika, dok većina autora ima jednog ili dvoje suradnika. U sukcesivnim vremenskim periodima, distribucija postaje sve više nalik tzv. Pareto ili Bradfordovoj distribuciji (slika 21). Prema tome, iako mehanizam

preferencijalnog povezivanja nije jedini kolji djeluje u nastanku veza, njegovo djelovanje je s vremenom sve izraženije.

3.2.3.4 Vremenska dinamika mreže koautorstva

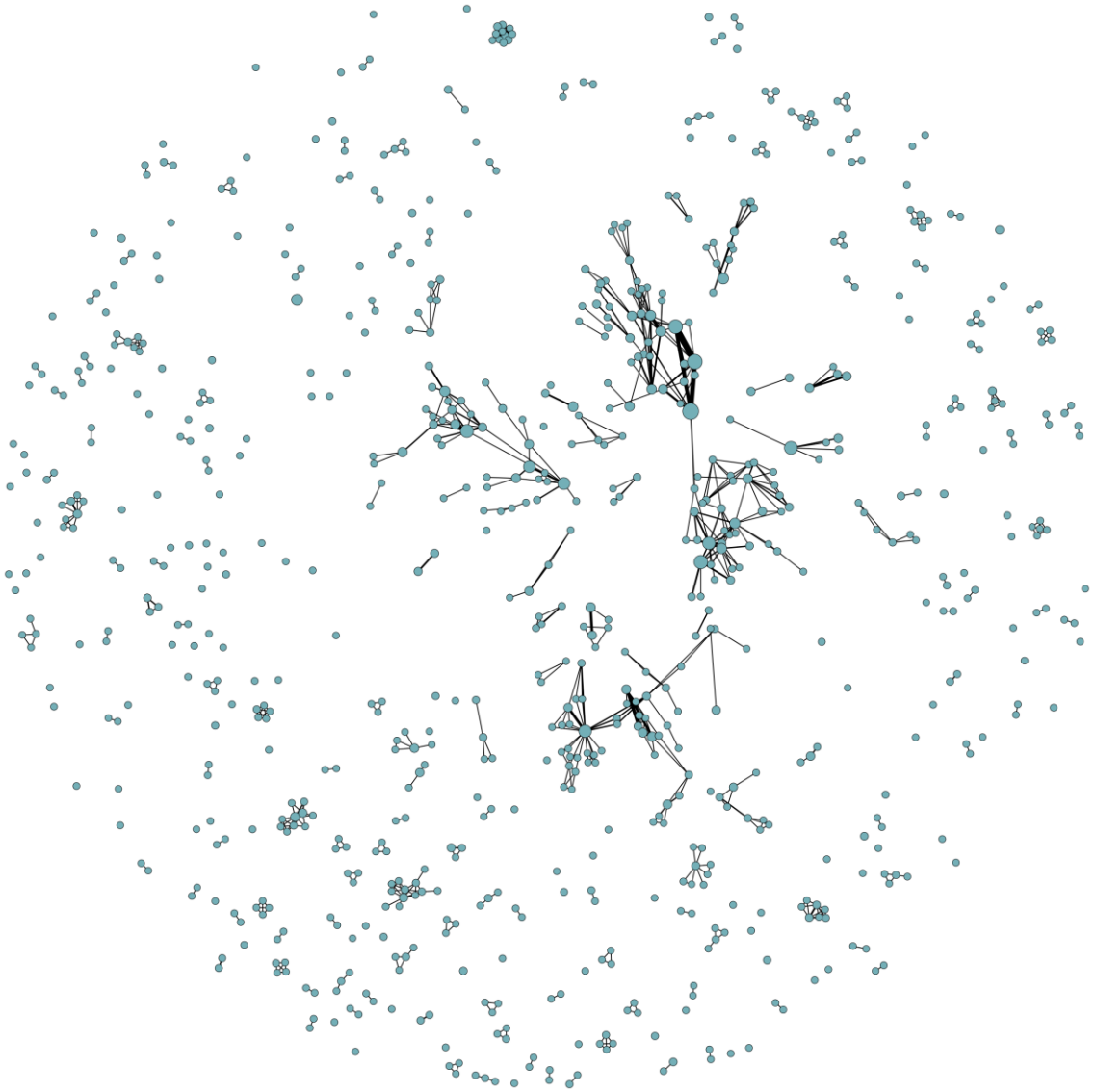
U svrhu razumijevanja dinamike mreže, iste analize su provedene zasebno za tri vremenska perioda (1978-1988, 1989-1999, 2000-2010), koje demonstriraju kako se značajke mreže mijenjaju kroz vrijeme.

Slika 22. Mreža suradnje autora na člancima u časopisu *Scientometrics* 1978-1988

U prvom periodu (1978 - 1988), ne postoji kohezivna struktura mreže: mreža se sastoji od većeg broja nepovezanih manjih komponenti, grupica autora koji međusobno ne surađuju. Najveća komponenta uključuje samo 4% autora. 30,2% autora su izolirani čvorovi, odnosno

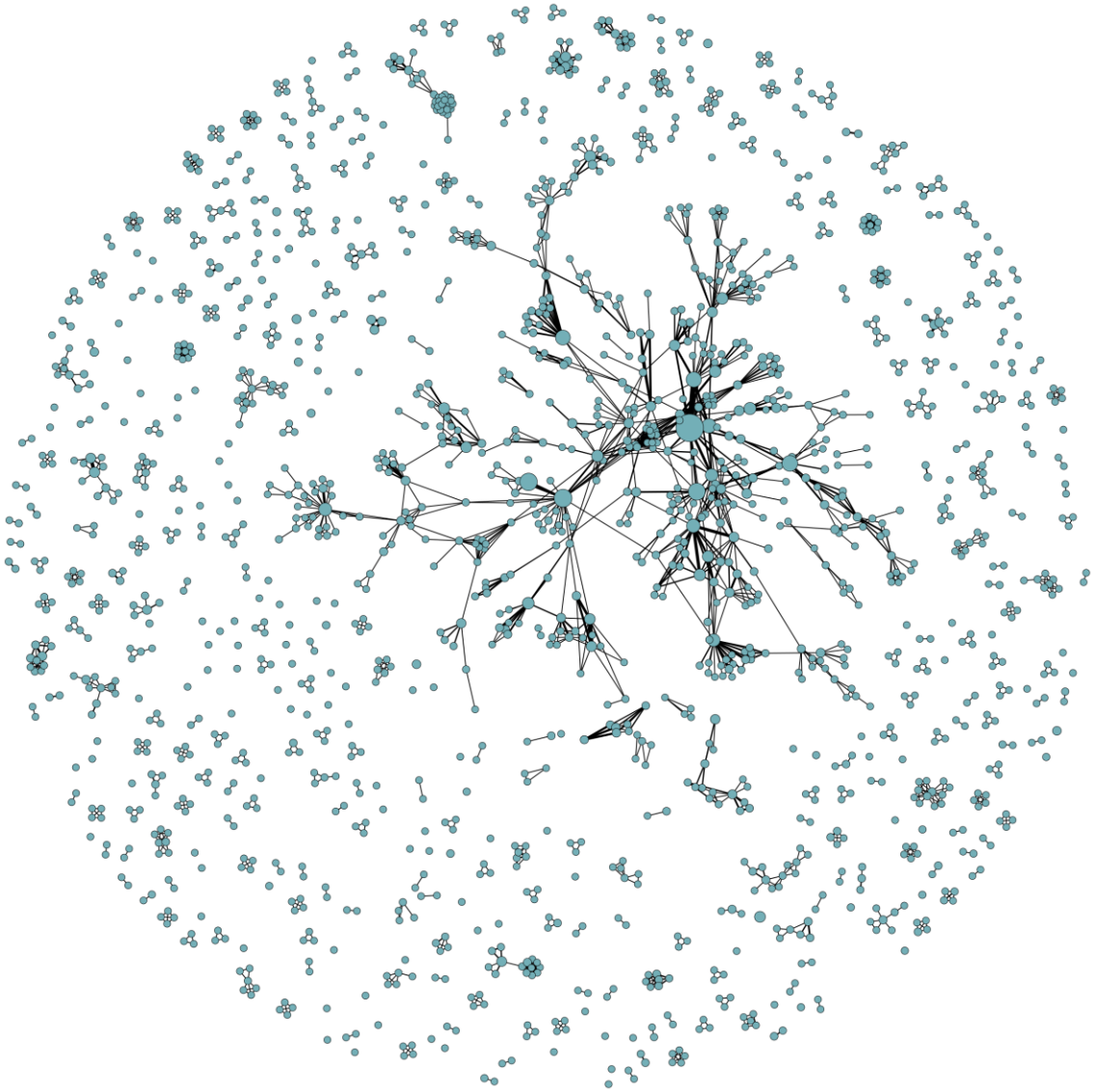
objavljuju samo jednoautorske radove, i to jedan ili više. Budući da se radi o mladoj znanstvenoj disciplini, a naročito u prvom praćenom razdoblju koje karakterizira uspostava discipline i u kojoj dominiraju teorijski i metodološki radovi, ovakva stuktura mreže je očekivana.

Slika 23. Mreža suradnje autora na člancima u časopisu *Scientometrics* 1989-1999



U drugom periodu (1989 - 1999), glavna komponenta u apsolutnom i relativnom smislu postaje veća – uključuje 10% čvorova u mreži, a relativni udio izoliranih autora se smanjuje na 11,9%. Upravo i ovi podaci tj. drugačija struktura mreže u odnosu na prvo analizirano razdoblje, potvrđuju tezu o razvoju scientometrije kao discipline. To se može obrazložiti porastom empirijskih istraživanja, intenzivnijim radom i suradnjom autora koji se primarno bave scientometrijom te razvojem metodoloških instrumenata.

Slika 24. Mreža suradnje autora na člancima u časopisu *Scientometrics* 2000-2010



U trećem periodu (2000-2010) glavna komponenta raste i uključuje 26% čvorova. Samo je 8,7% izoliranih čvorova, a zanimljivo je primijetiti da oko 15% autora pripadaju u izolirane dijade, a oko 13% u izolirane trijade.

Vremenom gustoća pokazuje trend opadanja, što je u skladu s empirijskim podacima koji se dobivaju na mrežama koautorstva. Naime, uključivanjem sve većeg broja autora u mrežu, broj mogućih veza među autorima eksponencijalno raste, pa je gustoća u pravilu niža kako mreža postaje veća. Iz istog razloga je dijametar, duljina puta između dva najudaljenija čvora u mreži, očekivano vremenom rastao. Ipak, ne dolazi do njegovog dramatičnog povećavanja, već raste umjereno i konstantno u promatranim periodima (4, 9, 14). Prosječna duljina puta, također, vremenom pokazuje tendenciju rasta, budući da raste i broj čvorova u glavnoj

komponenti. Međutim, prosječna udaljenost je vrlo niska i njen rast kroz vrijeme je vrlo spor, što ukazuje na dobru povezanost glavne komponente. To posredno govori da postoji mali broj artikulacijskih čvorova, odnosno osoba koje povezuju inače nepovezane dijelove mreže.

Prosječni stupanj centralnosti u glavnoj komponenti pokazuje da su autori radova u *Scientometricsu*, surađivali sa samo jednom do dvije osobe u prva dva vremenska perioda, dok u zadnjem periodu prosječni broj suradnika (veza, koautora) po čvoru raste na tri (3,5) suradnika. Za taj porast su odgovorni višeautorski radovi koji se pojavljuju u većem broju u zadnjem vremenskom periodu, pa na taj način dolazi do porasta u prosjeku veza, dok je medijan - dva suradnika po čvoru. To nije u kontradikciji sa sve manjom gustoćom, budući da čak i kad suradnja postaje sve učestalija, broj mogućih veza u sve većoj mreži raste eksponencijalnom brzinom.

U analizi obrazaca suradnje, korisnom se pokazala kategorizacija autora koju su predložili Price i Gürsey (prema Glänzel i Schubert, 2005). Autore svrstavaju u jednu od četiri kategorija s obzirom na njihovo objavljivanje prije i poslije promatranog vremenskog perioda:

- **kontinuirani** – autori koji su objavljivali u promatranom vremenskom periodu, te ili prije ili poslije njega
- **privremeni** – autori koji objavljuju u određenom vremenskom periodu, ali ne prije i ne poslije
- **pridošli** – autori koji objavljuju u promatranom vremenskom periodu i poslije njega, ali nisu objavljivali prije
- **terminatori** – autori koji su objavljivali prije i za vrijeme promatranog perioda, ali ne poslije.

Promatranje načina kako su na ovaj način kategorizirani autori međusobno surađivali, dovelo je do bitnih zapažanja o nekim aspektima koautorstva. Naime, u većini koautorskih radova su sudjelovali kontinuirani autori, pa se smatra da su u podlozi sve veće suradnje u znanosti zapravo stabilni timovi. U našem uzorku je udio kontinuiranih autora najmanji (tablica 13). S druge strane, broj privremenih autora raste. Takvi rezultati ukazuju na prije opisani trend da se scientometrijom bave mnogi autori samo povremeno.

Tablica 13. Promjene u broju autora po razdoblju

pokazatelj	1978-1988	1989-1999	2000-2010
n autora	350	807	1755
n privremenih	243	540	1533
n pridošlih	350	712	1533
n terminatora	243	597	1755
n kontinuiranih	107	267	222

3.2.4 Zemlje ustanova autora i međunarodna suradnja

Informacija vezana uz autore, ali posebno bilježena u metapodacima, su podaci o ustanovama autora. Rad s podacima o ustanovama autora slična je problematika kao i imenima autora s dodatnim problemom granularnosti i prijevoda naziva na engleski jezik. U slučaju, na primjer, Sveučilišta u Zagrebu, na različite radove i ovisno o vremenskom razdoblju potpisani su sveučilište, fakultet, odjel, katedra i slične jedinice i to neujednačeno. Kod nekih radova postoji samo informacija o sveučilištu dok je kod drugih poznata informacija samo na razini katedre. Kada se navedenom pridodaju inačice pisanja kao i prijevodi i promjene naziva ili prijevoda kroz vrijeme, razrješavanje imena ustanova uz tehnološki problem postaje i konceptualan.

U ovom istraživanju, radi opsega, ustanovama nije pridodana posebna pažnja s obzirom na vrlo velik broj različitih autora (pa, u nešto manjoj mjeri, i vezanih ustanova) koji su objavili samo jednom što, pored te informacije, rezultate čini teško prikazivaim i interpretabilnim, što je pojašnjeno kasnije u tekstu. Iskorištene su međutim informacije o zemljama ustanova koje su prisutne u samoj adresi i koje je jednostavnije automatski jednoznačno identificirati. Navedeno pruža uvid u svjetsku relevantnost i geografsku poziciju aktera, odnosno zemalja koji razvijaju scientometriju kroz objavljivanje u časopisu *Scientometrics*.

Radi uvida u međunarodnu suradnju i strukturu doprinosa razvoju scientometrije, iz adresa autora preuzeti su nazivi zemalja ustanova. Ovi podaci mogu se shvatiti grupirajućima za autore u nekoj točci u vremenu (autor je mogao promijeniti ustanovu ili djelovati u više njih). Analiza je bila provedena na svim člancima za koje su bile dostupne informacije o adresi. Od ukupno 2 469 radova koji su promatrani kao članci, ova analiza zahvaća njih 1 828 tj. 74%. I ovaj pokazatelj ima svoju vrijednost u praćenju razvoja znanstvene discipline, ali i razvoja forme članka. Podaci o zemljama za sva promatrana razdoblja vidljivi su iz tablice 14.

Tablica 14. Pregled zemalja koje su zastupljene u časopisu *Scientometrics* s više od 19 članaka

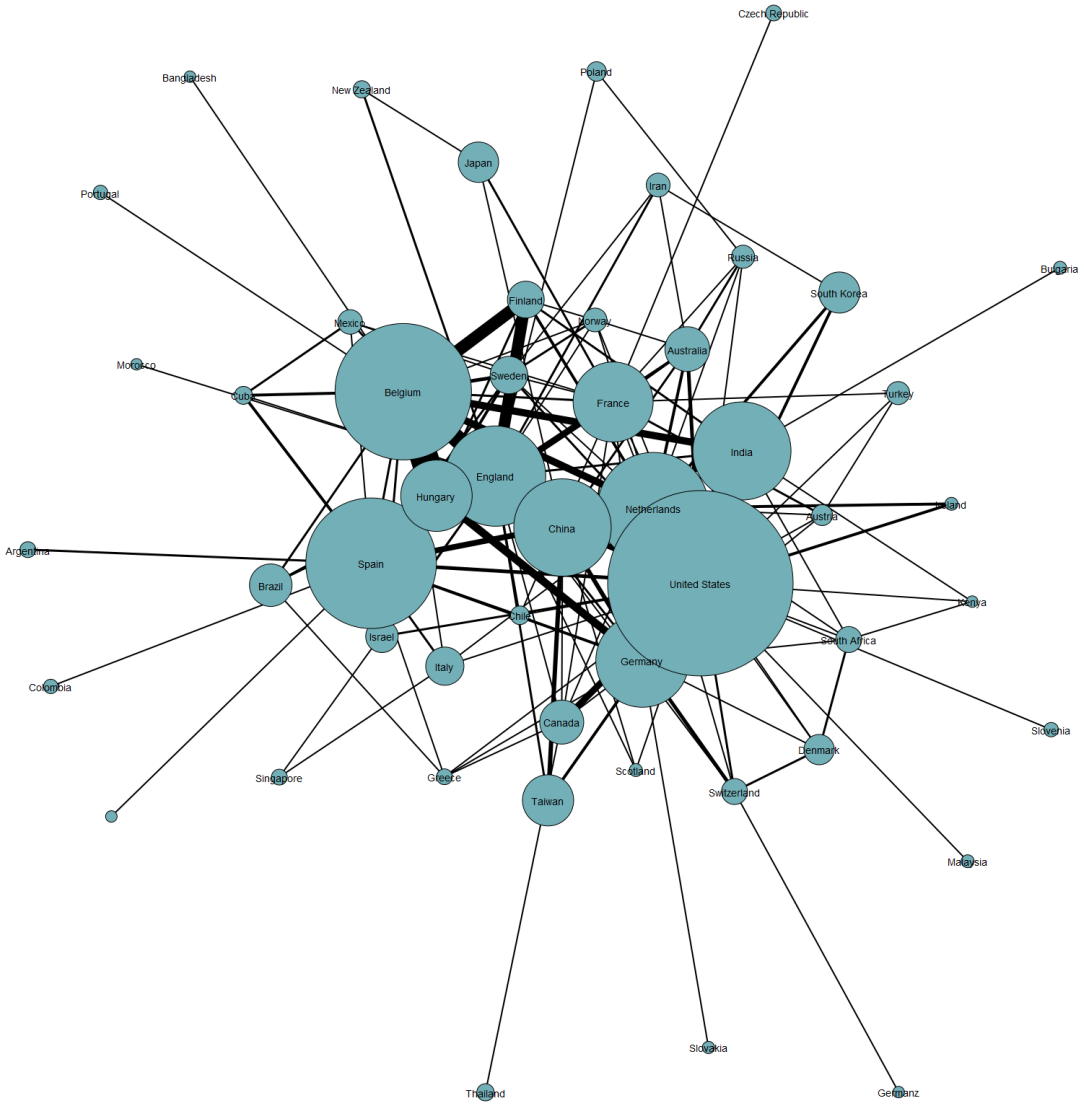
država	broj radova	min-max godina objave
United States	244	1979-2010
Belgium	176	1986-2010
Spain	167	1986-2010
Netherlands	140	1985-2010
India	127	1985-2010
England	124	1980-2010
China	124	1985-2010
Germany	116	1990-2010
France	97	1982-2010
Hungary	93	1981-2010
Taiwan	56	2000-2010
Australia	47	1982-2010
Canada	46	1978-2010
Brazil	44	1993-2010
South Korea	42	1999-2010
Japan	41	1983-2010
Italy	38	1995-2010
Sweden	37	1993-2010
Finland	36	1984-2010
Israel	30	1985-2010
Denmark	27	1996-2009
Switzerland	24	1988-2010
South Africa	21	2000-2010

Kao što vidimo, u časopisu *Scientometrics* objavljuju podjednako autori iz više zemalja s različitih dijelova svijeta. Prema ovim podacima *Scientometrics* možemo uistinu nazvati časopisom globalne relevantnosti. Na popisu zemalja s više od 100 radova zastupljena su tri kontinenta, na popisu zemalja s više od 40, 5 kontinenata, a na popisu s više od 20, 6.

U interpretaciji produktivnosti zemalja svakako treba uzeti u obzir duljinu aktivnog perioda. U sklopu analize časopisa *Scientometrics* s aspekta razvoja discipline važna je i produktivnost i duljina perioda, visoko produktivne zemlje u vrlo kratkom razdoblju manje su longitudinalno važne za razvoj scientometrije od zemalja sa sličnim ili manjim brojem radova u dužem periodu čije djelovanje je kontinuirano.

Glavna komponenta suradnje među autorima koji su zaposleni u ustanovama različitih zemalja vidljiva je iz prikaza mreže na slici 25.

Slika 25. Mreža međunarodne suradnje na radovima u časopisu *Scientometrics*



Slika 25 pokazuje da su u mreži međunarodne suradnje centralni akteri ujedno i najproduktivniji: SAD, Belgija, Španjolska, Nizozemska, Engleska, Indija, Francuska, Kina i Njemačka. Mnogo su u surađivali i autori iz Kanade, Švedske, Finske i Mađarske, iako su bili relativno manje zastupljeni prema broju radova. Pozicija Mađarske pokazuje visoku povezanost na manjem broju radova no ostale zemlje što se može objasniti činjenicom da je jedan od izdavača časopisa mađarski izdavač Akadémiai Kiadó, a dugodišnji urednik, Tibor Braun, kao i njegovi suradnici, Wolfgang Glänzel i Andras Schubert, visoko su produktivni i sva trojica su zaposleni u Mađarskoj akademiji znanosti i umjetnosti u Budimpešti.

Prema obrascu veza, može se uočiti da manje produktivne zemlje (na periferiji slike) su većinom surađivale s manje centralnim i produktivnim zemljama, dok su visoko produktivne zemlje međusobno intenzivnije surađivale. To je u skladu s prije navedenim rezultatima o visokoj asortativnosti – tendenciji povezivanja aktera koji su slični po ukupnom broju koautora. Takvo asortativno biranje (biranje po sličnosti) ukazuje na postojanje zemalja koje su centri te čiji autori u najvećoj mjeri oblikuju polje scientometrije.

3.3 Citatne analize članaka u časopisu *Scientometrics*

U citatne analize uključeni su svi članci objavljeni u časopisu Scientometrics 1978-2010 i radovi koji su ih citirali prema WoS citatnim indeksima, koji su objavljeni 1978-2012, uključujući i radove u časopisu Scientometrics koje možemo smatrati samocitatima časopisa. Radovi objavljeni u časopisu Scientometrics 2011. i 2012. godine uključeni su, dakle, u analize samo kao citirajući radovi jer nije prošlo dovoljno vremena u trenutku analize kako bi isti prikupili dovoljno citata da ih se može promatrati kroz citate koje su primili.

Analiza citata koje su članci u časopisu Scientometrics primili pruža informaciju o odjeku časopisa Scientometrics u međunarodnim časopisima indeksiranim u bazi WoS. Proučavanje radova koji su te citate pružili tj. njihovih časopisa, daje uvid u područja koja konzultiraju i koriste rezultate scientometrijskih istraživanja. S druge strane, analiza citata koje su članci u časopisu Scientometrics pružili prema drugim radovima, odnosno proučavanje literature navođene u člancima u Scientometricsu daje pak uvid u područja na čije se radove članci u časopisu Scientometrics pozivaju.

3.3.1 Citiranost članaka u časopisu *Scientometrics*

Ovo poglavlje se koncentrira na odjek članaka objavljenih u časopisu Scientometrics. Radovi koji citiraju, odnosno uključuju članke objavljene u časopisu Scientometrics u popisima literature na koju se pozivaju, u ovom su kontekstu citirajući radovi. Jedan citirajući rad pruža više citata i to, u standardnoj operacionalizaciji, onoliko koliko ima navoda u popisu literature. Članci objavljeni u časopisu Scientometrics koji su primili barem jedan citat su u ovom kontekstu citirani radovi, odnosno članci, budući da se analiza ograničila samo na njih.

Citirajuće radove treba interpretirati kao sve radove i priloge koji uključuju popis literature kao sastavan dio rada odnosno sve radove i priloge unutar kojih WoS prati citate. S obzirom na usku specijalizaciju časopisa, očekivano je da će se članci u Scientometricsu često pozivati na druge radove objavljene u tom časopisu. Takvi citati se nazivaju samocitatima časopisa (Aksnes, 2003). Neki članak u časopisu Scientometrics može citirati radove iz časopisa Scientometrics, što ga smješta u skup citirajućih radova, može biti citiran bilo u časopisu Scientometrics ili nekom drugom časopisu indeksiranom u bazi WoS, što ga smješta u skup citiranih radova, a može pripadati i u obje skupine.

Kao što je već opisano u metodologiji, podaci o citiranosti preuzeti su iz WoS citatnih indeksa koji se mogu smatrati mjerodavnim i relevantnim za časopis Scientometrics. Pregled citata te citiranih i citirajućih radova prikazan je u tablici 15.

Tablica 15. Pregled citata svih radova objavljenih u časopisu *Scientometrics* 1978-2010 koje su dobili iz časopisa indeksiranih u WoS-u 1978-2012

pokazatelj	svi radovi	članci 2010	ostalo
n citirajućih radova	9050	8914	1176
n citirajućih časopisa	1818	1792	268
n citata	31427	29482	1945
maks. citata po citiranom radu	280	280	182
n citiranih radova	2705	2298	407
n Scientometrics radova	3431	2469	962

Očekivano, najveći broj citata bio je na članke objavljene u razdoblju 1978-2010 (93.8%). U citatnoj analizi uobičajen je postupak ostaviti barem dvije godine razmaka od objave do analize citiranosti (Jokić, 2005). Primjerenost tog razdoblja odmaka ovisi o trendovima navođenja literature u raznim disciplinama. Pri interpretaciji rezultata, bez obzira na koje je razdoblje odmaka odabrano, treba držati na umu da su najnoviji ili noviji radovi imali manje vremena prikupiti citate, za razliku od potencijalne mogućnosti radova objavljenih ranije, odnosno kojima je prošlo više od tri godine od objavljivanja. Naravno, točnost ove tvrdnje prvenstveno se odnosi na radove iz problematike kojom se bavi ovaj doktorski rad. Za radove iz vrlo dinamičnih područja, ako nisu citirani unutar prvih godina od izlaženja, velika je šansa da nikada neće biti citirani. Gupta (1990) je proučavao 15 vodećih časopisa iz fizike i pronašao je da gustoća citata pada *eksponencijalno* nakon pet godina.

Kao što vidimo, postotak članaka objavljenih u časopisu Scientometrics koji su primili barem jedan citat u časopisima indeksiranim u WoS–u, izuzetno je visok (93,1%) što govori u prilog svjetskoj relevantnosti ovog časopisa. Za usporedbu, prema istraživanju koje je proveo Gisvold (1999), veliki postotak članka u većini časopisa indeksiranih u WoS-u ne bude nikada citiran. Dok najveći udio citata dobivaju članci (93,8%), i ostale vrste radova imaju šansu biti visoko citiranim s maksimalnim brojem od 182 citata po radu. Pregled citata na ove vrste radova prikazan je u tablici 16.

Tablica 16. Pregled citata prema vrstama radova i priloga u časopisu *Scienotmetrics*

vrsta rada	1978-1988	1989-1999	2000-2010	1978-2010
sažetak sastanka	0	1	0	1
ispravak	1	0	0	1
bibliografija	72	14	3	89
novosti	1	4	4	9
recenzija knjige	14	5	5	24
uredničk tekst	23	26	8	57
ostalo	20	71	19	110
podatkovni izvještaj	206	457	26	689
korespondencija	210	136	63	409

Najcitiraniji prilog objavljen u časopisu Scientometrics je podatkovni izvještaj koji prenosi usporedne podatke o časopisima i zemljama[13]. Iz tablice 16 je vidljivo da su većina visoko citiranih radova i priloga koji nisu članci, izvještaji slične vrste i korespondencija uz pokoji uredički prilog i bibliografiju.

Ostatak citatnih analiza prikazanih u ovom radu proveden je na svim člancima objavljenim u časopisu Scientometrics 1978-2010 i citiranim zaključno s 2012-om godinom. Pregled citata na članke kroz razdoblja prikazan je u tablici 17.

[13] Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981–1985

Tablica 17. Pregled citata članaka objavljenih u časopisu *Scientometrics* 1978-2010 i citiranih 1978-2012 u odnosu na razdoblja

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
n Scientometrics radova	325	791	1353	2469
n citiranih radova	309	744	1245	2298
n citata	5118	9008	15356	29482
n citirajućih radova	2726	4067	5428	8914
n citirajućih časopisa	544	840	1306	1792
prosječno citata po citiranom radu	16.56	12.11	12.33	12.83
medijan citata po citiranom radu	9	7.0	7	7.0
iqr citata po citiranom radu	15	12.0	10	11.0
maks. citata po citiranom radu	162	178	280	280

Kao što vidimo iz tablice, u svakom od razdoblja preko 90% članaka je dobilo barem jedan citat što govori o visokoj relevantnosti objavljenih članaka. Obzirom da se radi o gotovo jedinom specijaliziranom časopisu za tematiku, navedeno ukazuje i na kontinuiranu prihvaćenost scientometrijskih tema. Također, medijan broja citata po citiranom radu ukazuje na relativno visoku prosječnu citiranost individualnih članaka u svim razdobljima.

Zanimljivo je da je prosječna citiranost po radu kroz godine kao i za cjelokupno razdoblje gotovo identična, što ukazuje na sličan status scientometrijskih članaka od početka objavljivanja časopisa. Jedino odstupanje je vidljivo za početno razdoblje što nije neobično. Pored toga što su ovi radovi imali najviše vremena prikupiti citate, za očekivati je da se u prvim godinama časopisa objavljivalo više teorijskih radova s ciljem definiranja područja koji ostaju kontinuirano relevantni.

Na popisima najcitiranijih radova često nalazimo metodološke i teorijske radove (Peritz, 1983) te su ovi radovi u prosjeku više citirani. Citiranost radova prema njihovoj tematici je vidljiva iz tablice 18.

Tablica 18. Pregled citata članaka objavljenih u časopisu *Scientometrics* 1978-2010 i citiranih 1978-2012 u odnosu na njihovu tematiku

pokazatelj	primijenjeni	metodološki	teorijski
n Scientometrics radova	1388	702	379
n citiranih radova	1288	661	349
n citata	13289	10354	5839
n citirajućih radova	5773	4830	3562
n citirajućih časopisa	1266	953	761
prosječno citata po citiranom radu	10.32	15.66	16.73
medijan citata po citiranom radu	6.0	8	8
iqr citata po citiranom radu	9.0	15	16
maks. citata po citiranom radu	280	272	151

Prema prikazanim podacima može se zaključiti da su metodološki i teorijski radovi u prosjeku citiraniji od primijenjenih s medijanom 8 u odnosu na 6 citata. Veći interkvartilni raspon ukazuje i na veće raspršenje citata kod metodoloških i teorijskih članaka: među tim radovima je više onih koji su malo, odnosno vrlo visoko citirani.

Najcitiraniji članak, međutim, je empirijsko istraživanje i to s gotovo dvostruko više citata od najcitiranijeg teorijskog rada. Riječ je o članku "Citation review of Lagergren kinetic rate equation on adsorption reactions" (Yuh-Shan, 2004) koji prenosi rezultate istraživanja citata koje je primio jedan seminalan rad i primjer je "scientometrije za znanstvene discipline". Velik broj citata koji ovakvi članci katkad dobivaju mogu proizlaziti i iz trendova citiranja u znanstvenoj disciplini koja je primarno namijenjena publici članka. Preciznije rečeno, ovaj rad se ne može interpretirati kao važan za *razvoj* scientometrije, iako je dobio najviše citata.

Što se najcitiranijeg metodološkog rada tiče, riječ je o "Theory and practice of the g-index" (Egghe, 2006) odnosno članku koji je predložio *g*-indeks kao pokazatelj i koji je već spomenut u metodologiji. Najcitiraniji teorijski rad je "Is citation analysis a legitimate evaluation tool?" (Garfield, 1979), koji je također spomenut, ali u uvodu, s naglaskom da se radi o važnom radu za današnji status citatnih analiza kao važnog i prihvaćenog dijela scientometrijskog instrumentarija. Spomenuti primjeri govore o svrsishodnosti podjele na primijenjene, metodološke i teorijske radove za potrebu istraživanja razvoja neke discipline.

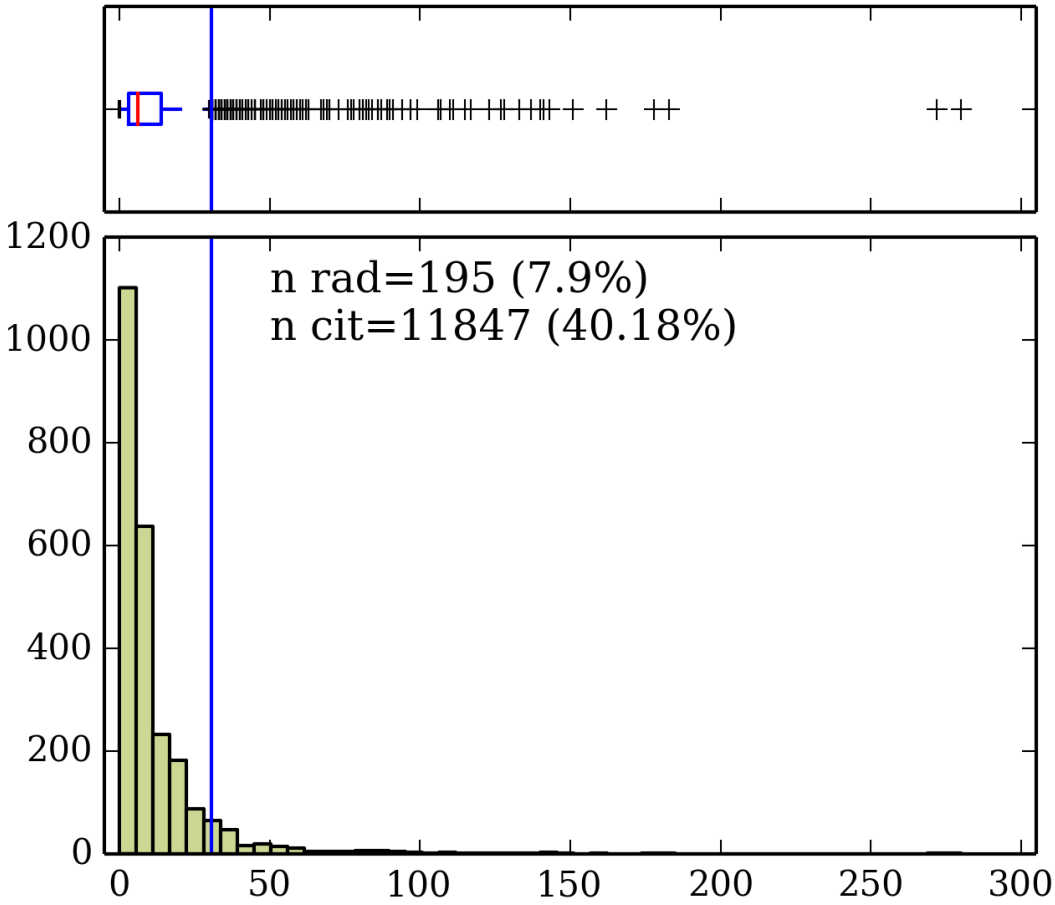
Podaci o radovima koji su u svojim skupinama dobili maksimalan broj citata ne govore, međutim, puno o sastavu skupa najcitiranijih radova u časopisu *Scientometrics*. Ovaj skup radova je prikazan u sljedećem poglavlju.

3.3.2 Najcitiraniji članci objavljeni u časopisu *Scientometrics*

Najcitiraniji radovi u nekoj skupini radova su oni koji su najviše odjeknuli tj. koji su najčešće korišteni kao podloga novim radovima. Uključivanjem ovih radova u analize promatra se podskup radova za koje je moguće pretpostaviti da su imali značajan utjecaj na daljnja istraživanja, u odnosu na ostale radove u istom skupu. Slično vrijedi i za najcitiranije autore s tim da najcitiraniji autori ne moraju biti autorima najcitiranijih radova, a autori najcitiranijih radova ne moraju biti najcitiraniji autori. Drugim riječima, mogu postojati takvi autori koji su visoko citirani, ali na većem broju citiranih radova i obratno.

U operacionalizaciji odabira najcitiranijih radova tj. autora iskorištena je definicija "gornjih" ekstremnih vrijednosti prema interkvartilnom rasponu. Kao visoko citirani članci odabrani su takozvani "blagi ekstremi" tj. svi radovi čiji broj citata prelazi Q3 + IQR * 1,5. Distribucija citata po radovima koja prikazuje i ekstreme vidljiva je na slici 26.

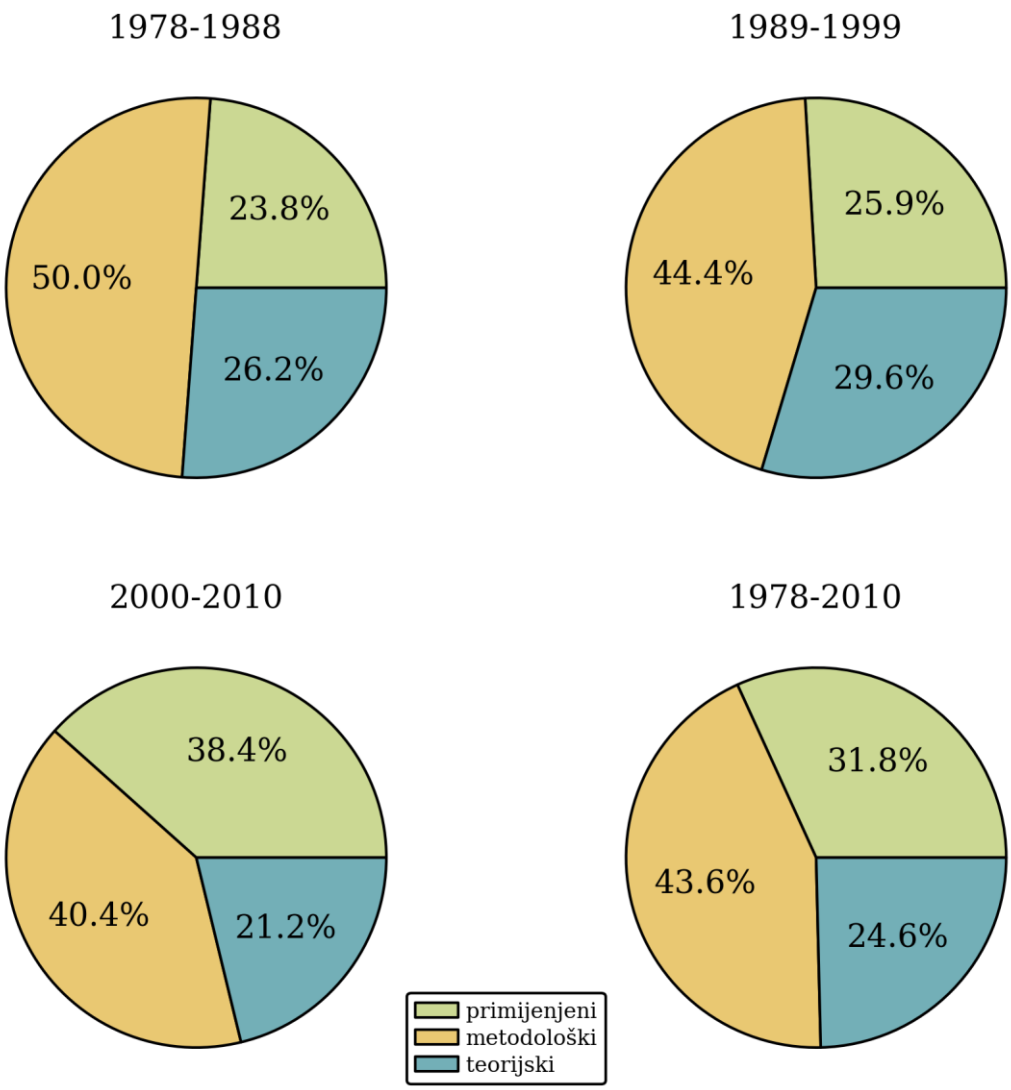
Slika 26. Distribucija citata po citiranim člancima i visiko citirani radovi



U ovom skupu članaka, visoko citirani članci su svi članci koji su dobili 31 ili više citata. Ovih članaka je 8% od ukupnog skupa članaka i primili su 40% od svih citata koje su primili članci objavljeni u časopisu *Scientometrics*. Za usporedbu, Gisvold (1999) je na svom uzorku radova iz prestižnog časopisa Nature, dobio podatak da je 20% radova dobilo oko 80% svih citata, što je gotovo identično Paretovom pravilu 80/20. Zanimljiv je i podatak da je 16% najcitiranijih radova dobilo više od polovice svih citata.

Prilikom proučavanja skupa visoko citiranih članaka, zanimljivo je provjeriti udjele članaka u odnosu na tematiku. Navedene informacije prikazane su na slici 27.

Slika 27. Udjeli teorijskih, metodoloških i primijenjenih visokocitiranih članaka



Kao što je sa slike vidljivo, među najcitiranijim radovima i dalje su najčešći rezultati empirijskih istraživanja, ali je to manje izraženo u odnosu na skup svih radova. U cijelom razdoblju udio svih primijenjenih članaka je 56,2%, a u skupu visoko citiranih radova udio primijenjenih članaka je 43,6%. Dodatno, udio ovakvih članaka u zadnjem razdoblju raste kod svih članaka i opada kod visoko citiranih članaka.

Kao i kod svih radova, broj metodoloških radova pokazuje povećanje u posljednjem razdoblju, ali je zapaženi porast udjela visoko citiranih metodoloških članaka izraženiji. Kod svih radova udio metodoloških članaka u trećem razdoblju bio je 30,2% s porastom od 3,5% u odnosu na prethodno razdoblje. Udio visoko citiranih metodoloških članaka u trećem razdoblju bio je 38,4% s porastom od 14,5%, što je blizu udjela primijenjenih članaka u tom

razdoblju koji iznosi 40,4% i koji se smanjio za 4% u odnosu na prošlo razdoblje. Dobiveni rezultati ukazuju na važnost razvoja scientometrijske metodologije, što je posebno opisano i naglašeno u drugom poglavlju ovog rada.

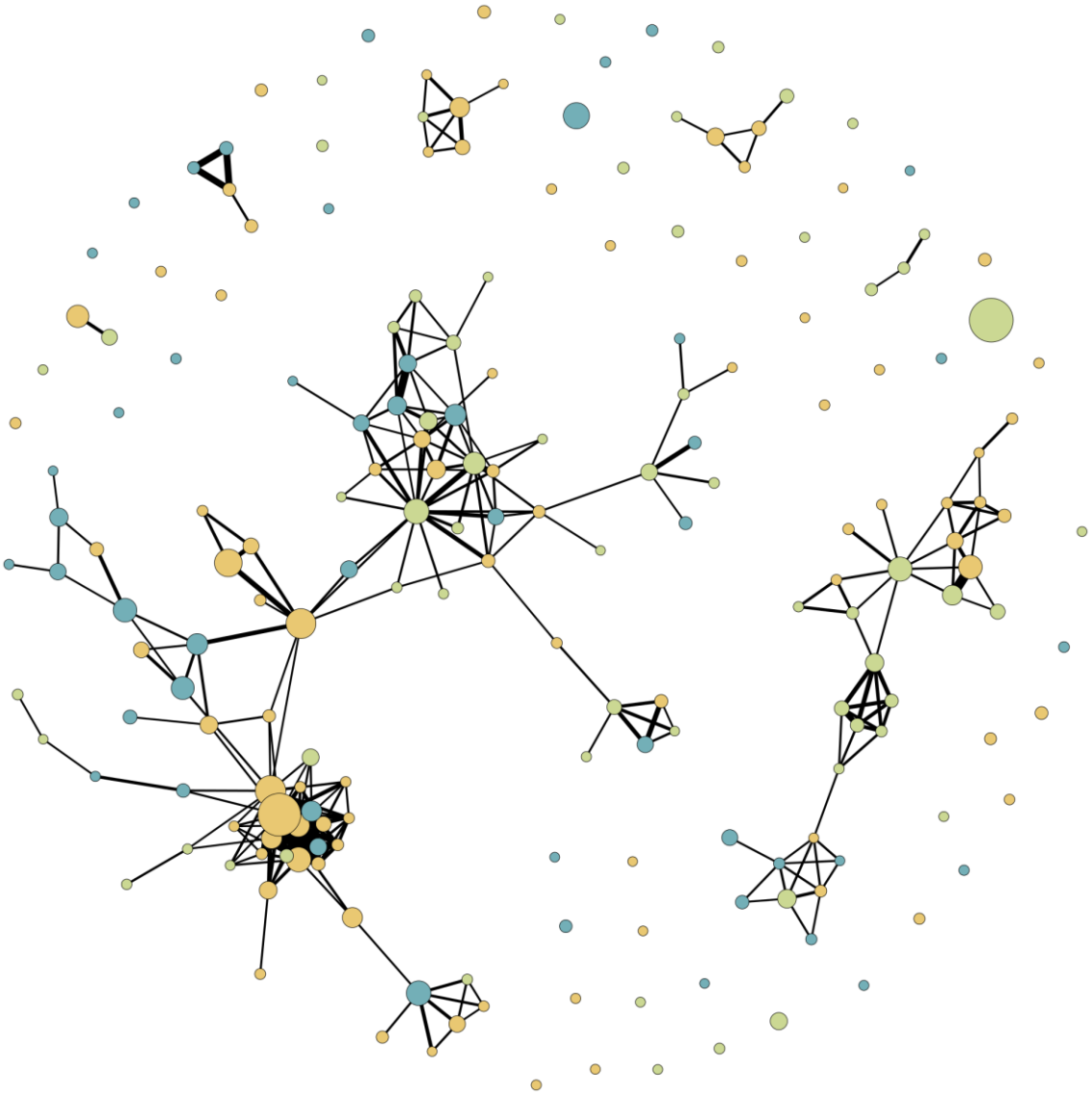
Udio visoko citiranih teorijskih članaka kroz sva tri razdoblja je postojaniji nego što je to slučaj za korpus svih članaka. Kod korpusa svih članaka, udio teorijskih članaka kontinuirano pada s približno četvrtine u prvom razdoblju (24,6%) do približno desetine svih članaka u posljednjem razdoblju (11,7%). Kod visoko citiranih članaka udio teorijskih članaka raste u drugom razdoblju u odnosu na prvo, zatim opada u odnosu na treće, ali taj pad je s četvrtine (26,2%) na petinu (21,2%) što odgovara poimanju teorijskih radova kao radova od kontinuirane važnosti za razvoj discipline.

3.3.2.1 Kocitiranost najcitiranijih članaka

Kocitatna analiza može otkriti međupovezanost znanstvenih publikacija definiranu kroz njihovo "su-korištenje" u drugim znanstvenim publikacijama. S obzirom da se u ovom slučaju povezuju visoko citirani članci, ovako dobivene grupe radova su značajne grupe radova za razvoj scientometrije. Ako je taj skup radova reprezentativan za neku disciplinu, što je u ovom uzorku razumno za pretpostaviti, tada ove grupe prikazuju tematske cjeline ili specijalizacije radova te discipline prema informacijama o njihovom korištenju u drugim znanstvenim radovima.

Mreža kocitata visoko citiranih članaka je prikazana na slici 28. Budući da se radi o visoko citiranim člancima, velik broj ih je kocitirano barem jedan put. Radi lakše interpretacije, prikazane su samo snažne veze za što je odabrana težina veze od 10 kocitata, jer značajke distribucije stupnjeva centralnosti (izrazito pozitivno asimetrična) nisu pogodovale odabiru putem ekstrema ili sličnog formalnog pravila. Veličina čvorova je u skladu s ukupnim brojem citata koji su primili. U mreži su zelenom bojom označeni primijenjeni članci, žutom metodološki, a plavom bojom teorijski članci.

Slika 28. Mreža kocitata visoko citiranih članaka objavljenih u časopisu *Scientometrics* koji su kocitirani barem 10 puta



Na grafu je jasno vidljivo nekoliko komponenata. Najveća komponenta sadrži 90 odnosno 46% visoko citiranih članaka. Pregledavanjem radova utvrđeno je da je riječ o radovima koji se primarno bave citatnim analizama i to najčešće u svrhu vrednovanja rezultata znanstvenog rada često u sklopu s kombinacijom odjeka i produktivnosti. Glavna tema mnogim od ovih članaka je *h*-indeks.

Tri najcitiranija rada u ovoj grupi su već spomenuti "Theory and practise of the *g*-index" (Egghe, 2006), a zatim i "Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups" (van Raan, 2006) te

"New bibliometric tools for the assessment of national research performance" (Moed et al., 1995).

Druga najveća komponenta sastoji se od 29 članaka koji se tematski mogu vezati uz ko-citatne analize i proučavanje odnosa između znanosti i tehnologije, odnosno uz šire područje mapiranja znanosti. Mnogi od ovih članaka vezani su uz patente kao glavnu jedinicu obrade, ali veći broj citata primaju kocitatne analize. Tri najcitiranija članka u ovoj grupi su "Mapping the backbone of science" (Boyack et al., 2005) te "Clustering the Science Citation Index using co-citations Part I: A Comparison of methods" (Small i Sweeney, 1985) kao i drugi dio ovog članka s podnaslovom "Mapping science" (Small et al., 1985).

Među ostalim komponentama s 5 ili više radova nalazimo dvije komponente. Jedna šest članaka i glavna tema svih šest je analiza supojavnosti riječi. Dva najčešće citirana članka iz ove komponente su "Co-word maps of biotechnology: An example of cognitive scientometrics" (Rip i Courtial, 1984) i "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry" (Callon et al., 1991).

Kao što je iz naslova navedenih radova vidljivo, od njih osam, tri su vezana uz područja prirodnih znanosti, a dodatna dva uz podatke iz SCI-a, koji je reprezentativan za STM područja. Preostala tri rada su metodološka i nisu vezana uz niti jednu disciplinu. Navedeno potkrepljuje spomenutu činjenicu da je tradicionalno scientometrija vezanija uz STM područja.

U ovom kontekstu, zadnja opisana komponenta se sastoji od pet članaka koji prikazuju prodor scientometrije i u društvene znanosti i humanistiku što je i tema svih pet članaka. Dva najcitiranija članka u ovoj komponenti su "Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review" (Nederhof, 2006) te "The difficulty of achieving full coverage of international social science literature and the bibliometric consequences" (Hicks, 1999).

Zanimljivo je i primijetiti da najcitiraniji članak, koji je najveći čvor u mreži vidljiv u gornjem desnom uglu slike 28, prema podacima o kocitiranosti nije snažno povezan s niti jednim drugim radom. Navedeno je u skladu s interpretacijom tog rada kao "scientometrije za znanstvene discipline" odnosno interpretacijom da primarna publika tog rada nisu scientometričari te da citati na taj rad nisu toliko značajni iz pogleda razvoja scientometrije kao discipline kao što je to, na primjer, najcitiraniji metodološki rad koji je ukupno dobio manje citata, ali je povezan u glavnu komponetu prikazane mreže.

Mreža prikazana na slici 28, odnosno njena interpretacija u skladu je sa slikom scientometrije kao područja kakvim je prikazano u uvodu i metodologiji ovog rada.

3.3.3 Visoko produktivni i citirani autori u časopisu *Scientometrics*

O produktivnosti autora na razini cijelog uzorka više je bilo govora u poglavlju o autorstvu. Budući da je zanimljiv odnos između produktivnosti i citiranosti autora, u ovom poglavlju dat je naglasak na produktivnosti individualnih autora uz podatke o odjeku odnosno citiranosti tih radova. Preko podataka o broju radova dobivamo informacije o autorima koji su bili najproduktivniju u području, a preko podataka o citiranim radovima možemo promotriti i citiranost tih autora kako bi dobili uvid u autore za koje je razumno pretpostaviti da su objavljivali utjecajne radove. Naravno, ako pod pojmom najcitiranijih radova podrazumijevamo i najutjecajnije radove.

Top autori kao i pokazatelji o njihovim publikacijama su prikazani u tablici 19. Radi značajki distribucije produktivnosti autora, što je prikazana u poglavlju o autorstvu, koja otežava odabir skupa outliera, koji je smislen za prikaz i interpretaciju, za prikaz u tablici 19 odabrana je minimalna vrijednost od 20 radova.

Tablica 19. Pregled autora koji su u časopisu *Scientometrics* 1978-2010 objavili više od 19 članaka

autor	n radova	n citata	*h*-index	*g*-index	najčešća suradnja	min-max godina objave
Glänzel, Wolfgang	67	1961	25	43	21:Schubert, András	1983-2010
Schubert, András	49	1438	21	38	22:Braun, Tibor	1981-2010
Van Raan, Anthony F J	35	1274	20	35	7:Moed, Henk F	1985-2010
Braun, Tibor	29	873	14	29	22:Schubert, András	1981-2007
Leydesdorff, Loet	39	866	17	29	3:Meyer, Martin S;Park, Hyun Woo;Zhou, Ping	1981-2010
Moed, Henk F	28	1042	19	28	7:Van Leeuwen, Thed N;Van Raan, Anthony F J	1985-2009
Egghe, LEO.	40	693	11	27	9:Rousseau, Ronald	1986-2010
Meyer, Martin S	25	596	16	25	3:Bhattacharya, Sujit;Glänzel, Wolfgang;Leydesdorff, Loet	1998-2010
Vinkler, Péter	27	459	13	22	0:bez suradnje	1986-2010
Rousseau, Ronald	43	552	12	22	9:Egghe, LEO.	1987-2010
Van Leeuwen, Thed N	21	698	13	21	8:Tijssen, Robert J W	1993-2010
Courtial, Jean Pierre	21	392	10	20	6:Bailón-Moreno, Rafael;Ruiz-Baños, Rosario	1982-2007
Lewison, Grant	22	279	13	17	2:Grant, Jonathan	1991-2010
Kretschmer, Hildrun	21	211	9	15	3:Kundra, Rameshi;Li-Ming, Liang	1983-2008
Garg, KC.	22	221	9	14	6:Padhi, P.	1985-2006
Gupta, BM.	28	137	6	10	10:Karisiddappa, CR.	1990-2010

Uz podatke o produktivnosti i citiranosti radova, iz tablice su vidljivi i pokazatelji koji kombiniraju produktivnost i citiranost: *h*-indeks i *g*-indeks.

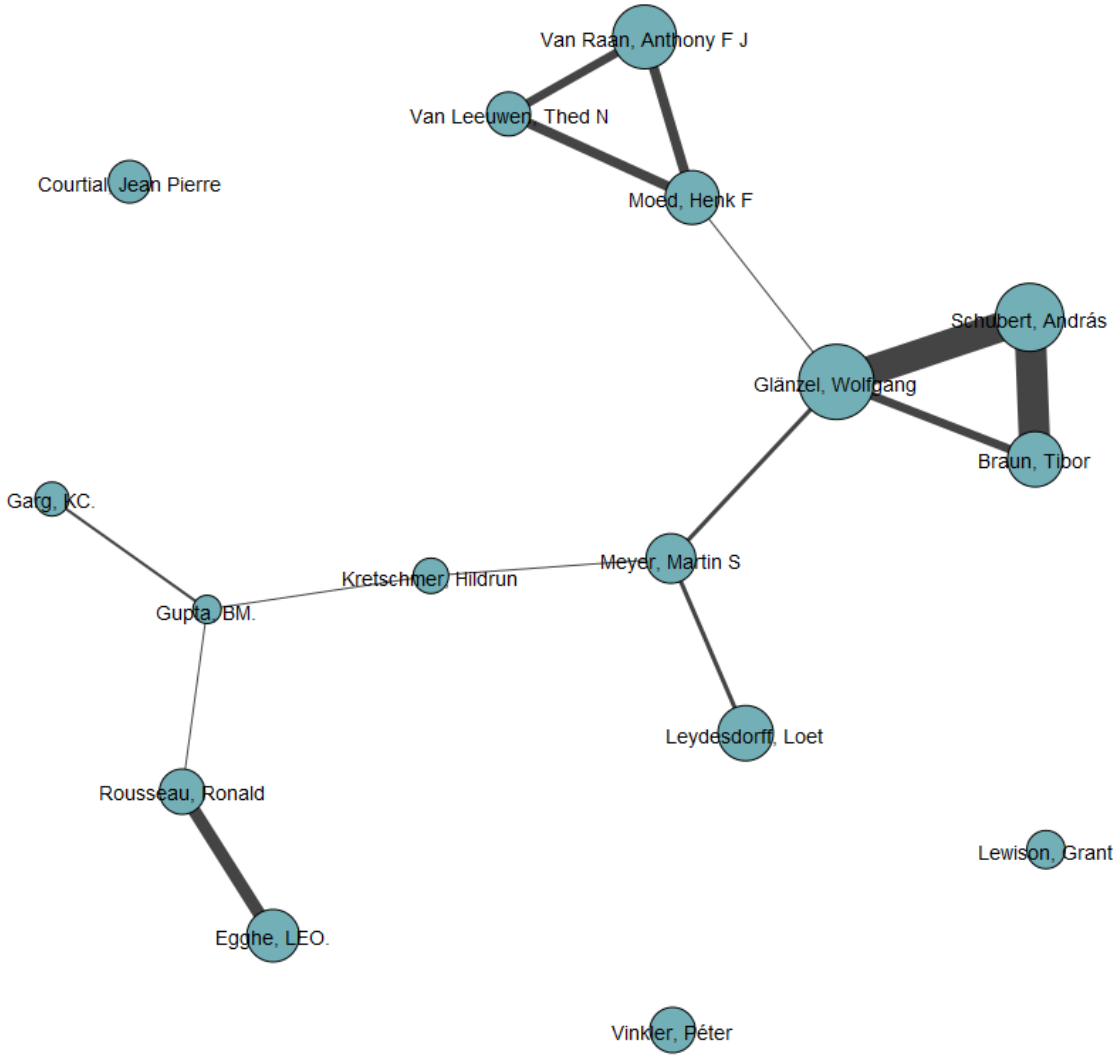
Umjesto po produktivnosti ili odjeku, tablica je poredana silazno po *g*-indeksu. *g*-indeks je odabran kao osjetljivija mjera, a *h*-indeks je uključen radi popularnosti tog pokazatelja, ali i kao nadopuna *g*-indeksu budući da je robusniji, odnosno manje osjetljiv na jedan supercitirani rad (Costas i Bordons, 2008).

Među najproduktivnijim autorima svakako treba istaknuti suradnju između Wolfganga Glänzela, Tibora Brauna i Andrása Schuberta koji su u top pet autora prema *g*-indeksu i

zajedno pokazuju najčešću suradnju u ovom časopisu, u promatranom razdoblju. Kao što je već u uvodu opisano, Braun je pokretač i urednik časopisa u cijelom razdoblju, Glänzel postaje urednikom 2014. godine, a Schubert je uz članke autor svih bibliografija objavljenih u časopisu *Scientometrics*. Oni su kao tim radili u Mađarskoj akademiji znanosti i umjetnosti u Budimpešti, gdje je bilo sjedište uredništva ovog časopisa i isključivo su se bavili istraživanjem znanosti, odnosno razvijali su scientometriju.

Suradnja među svim autorima iz tablice 21 je prikazana na slici 29. U prikazanoj vizualizaciji mreže, veličina čvora je linearno povezana uz *g*-indeks, a debljina veze uz broj članaka na kojem su potpisana oba autora koje veza povezuje.

Slika 29. Mreža suradnje autora koji su objavili više od 20 članaka u časopisu *Scientometrics*



I na ovoj slici jasno je vidljiva uloga trojice spomenutih autora s naglaskom da je ta veza podjednake jačine između Glänzela i Schuberta te Schuberta i Brauna. Zanimljivo je da je među najproduktivnijim autorima Peter Vinkler, čiji su svi članci u časopisu *Scientometrics* u promatranom razdoblju jednoautorski, a koji kao i spomenuta trojica autora radi u Mađarskoj akademiji znanosti i umjetnosti u Budimpešti. Rezultati, dakle, upućuju da su u tom dijelu Europe djelovali važni autori za ovo područje.

Prema podacima sa slike 29, druga najvažnija trojka najproduktivnijih autora koji međusobno najviše surađuju je leidenska grupa iz CWTS-a (The Centre for Science and Technology Studies, Leiden) Moed, van Raan i van Leeuwen. Njihova suradnja vidljiva kroz debljinu veze upućuje na čvrstu međusobnu povezanost ove trojice znanstvenika naročito značajnih za razvoj scientometrije, kao u metodološkom tako u empirijskom, ali i teoretskom diskursu.

Osim jasno vidljive povezanosti dviju ključnih i dominantnih grupa autora koji su najviše doprinijeli razvoju scientometrije, kao nezaobilazna ističe se i veza dvojice metodičara Lea Eggea i Ronalda Rousseaua.

Na ostatku slike uočljiva je suradnja svih visoko produktivnih i visoko citiranih kao i relativno slaba težina tih veza, odnosno vrlo rijetka suradnja, budući da se radi o visoko produktivnim autorima. Slične značajka suradnje među najproduktivnijim autorima u nekom promatranom skupu česti su rezultati analiza koautorstva (Newman, 2004).

3.3.4 Radovi i časopisi koji su citirali časopis *Scientometrics*

Kako bi se dobila cjelovitija slika znanstvene komunikacije unutar scienotmetrije praćene kroz časopis *Scientometrics* kao njenog ključnog reprezentanta, posebno je zanimljivo promotriti koji su to radovi i časopisi citirali radove iz Scientiometricsa. Kako bi se prikazao utjecaj časopisa *Scientometrics* na druge znanstvene publikacije, ovo poglavlje prikazuje skup radova koji su citirali *Scientometrics*. Kao i u prethodno opisanim obradama, radi opsega citatnih indeksa, uključeni su samo radovi iz časopisa. Pregled radova koji su citirali časopis *Scientometrics* od 1978. do 2012. godine vidljiv je iz tablice 20.

Tablica 20. Pregled radova objavljenih u časopisima u WoS citatnim indeksima 1978-2012 koji su citirali članke objavljene u časopisu *Scientometrics* 1978-2010

pokazatelj	1978-1992	1993-2002	2003-2012	1978-2012
n citirajućih časopisa	208	358	1512	1792
n citirajućih radova	1038	1638	6238	8914
n citiranih radova	405	992	2095	2298
n citata	2537	5001	21944	29482
prosječno citata po citirajućem radu	2.68	3.32	3.69	3.5
medijan citata po citirajućem radu	2.0	2.0	2.0	2.0
IQR citata po citirajućem radu	24.0	25.0	31.0	31.0
n referenci svih citirajućih radova	37797	54283	263130	355210
mean referenci po radu	36.41	33.14	42.18	39.85
medijan referenci po radu	21.0	23.0	33.0	30.0
prosječan udio citata na Sci. u referencama	0.14	0.16	0.13	0.14

U razdoblju od 1978. do 2012. godine ukupno 8 914 radova objavljenih u 1 792 časopisa indeksiranih u WoS-u, citiralo je članke objavljene u časopisu *Scientometrics*. Na određeni način, dobiveni rezultat je očekivan. Razlog ovoj tvrdnji je u samoj prirodi scientometrije kao discipline odnosno u već više puta spominjanoj "scientometriji za znanstvene discipline". Gotovo da i nema znanstvenog polja ili discipline koja se s određenim brojem radova ne bavi barem nekim aspektom scientometrije, od vrednovanja znanstvenog rada tog područja do specifičnosti razvoja određene znanstvene discipline.

U ovom smislu, korisno je vidjeti poimence koji su to konkretno časopisi koji su često, odnosno najčešće citirali časopis *Scientometrics*. U tablici 21 je prikazan popis od top 20 časopisa koji su najučestalije citirali članke iz časopisa *Scientometrics*.

Tablica 21. Top 20 časopisa indeksiranih u WoS-u 1978-2012 koji su objavljivali radove koji citiraju časopis *Scientometrics*

časopis	n citata	n citirajućih radova	n citiranih radova	prosječno udjela citata u referencama
Scientometrics	10783	2374	1995	0.23
Journal Of The American Society For Information Science And Technology	2903	637	969	0.14
Journal Of Informetrics	1517	268	644	0.18
Research Evaluation	701	178	388	0.17
Research Policy	685	267	335	0.07
Journal Of Information Science	474	144	333	0.14
Annual Review Of Information Science And Technology	462	35	364	0.05
Information Processing & Management	449	131	305	0.13
Journal Of Documentation	330	91	248	0.11
Plos One	247	67	177	0.12
Revista Espanola De Documentacion Cientifica	224	52	189	0.16
Technological Forecasting And Social Change	209	72	139	0.07
Aslib Proceedings	202	43	169	0.15
Czechoslovak Journal Of Physics	190	56	81	0.11
Online Information Review	182	45	109	0.12
Malaysian Journal Of Library & Information Science	181	41	153	0.19
Library & Information Science Research	148	32	121	0.1
Current Science	138	55	114	0.23
Social Studies Of Science	115	51	86	0.07
Interciencia	113	37	80	0.15

Kao što je vidljivo iz tablice 21, nešto više od trećine (36, 6%) svih citata na članke u časopisu *Scientometrics* je iz istog časopisa, tj. samocitata. S obzirom na status časopisa u području scientometrije, navedena činjenica nije neobična budući da se radi o mladoj disciplini i o specijaliziranom i ključnom časopisu koji promovira ovu disciplinu. Potvrda ovoj tezi i podataka da 96% posto članaka u časopisu *Scientometrics* citira radove iz časopisa Scienotmetrics, što je u prosjeku otprilike četvrtina (23%) svih citiranih publikacija po članku.

Udio samocitata časopisa, iako na prvi pogled djeluje relativno velik, zapravo i nije toliko velik s obzirom na upravo spomenute podatke, a naročito ne bi stajala tvrdnja ako bi ga se

htjelo interpretirati da priječi razmjenu ideja s drugim časopisima i područjima. Za usporedbu s područjem s većim brojem časopisa može poslužiti istraživanje (Krauss, 2007) u kojem se proučilo 107 časopisa iz područja ekologije i pronašlo prosječno 12% samocitata časopisa. Iz istih razloga iz kojih velik broj časopisa citira časopis *Scientometrics*, za očekivati je da će i broj časopisa citiranih u časopisu *Scientometrics* biti također velik, što su rezultati istraživanja i pokazali te je to kasnije u tekstu obrazloženo.

Pokazatelj "prosjek udjela citata u referencama" je izračunat tako što je za svaki citirajući rad izračunat udio citata koje pruža prema časopisu *Scientometrics* u odnosu na ukupan broj pruženih citata i zatim je izračunat prosjek tih udjela. Taj pokazatelj informira o tome kolika je "gustoća" citiranja časopisa *Scientometrics* u radovima u nekom časopisu bez obzira na broj citata koji članci u tom časopis uobičajeno pružaju, što varira između različitih područja i polja. Kroz ovaj pokazatelj moguće je utvrditi, na primjer, da radovi objavljeni u Research Evaluation koji citiraju *Scientometrics* uključuju veći udio referenci prema *Scientometricsu* u odnosu na radove objavljene u časopisu Research Policy, koji pruža sličan broj citata, ali raspršeniji po više radova. Navedeno ukazuje na drugačiji odnos ovih dvaju časopisa prema časopisu *Scientometrics* i možemo interpretirati da se radovi objavljeni u Research Evaluation više temelje na njemu nego radovi objavljeni u Research Policy, iako je i za njih časopis *Scientometrics* relevantan.

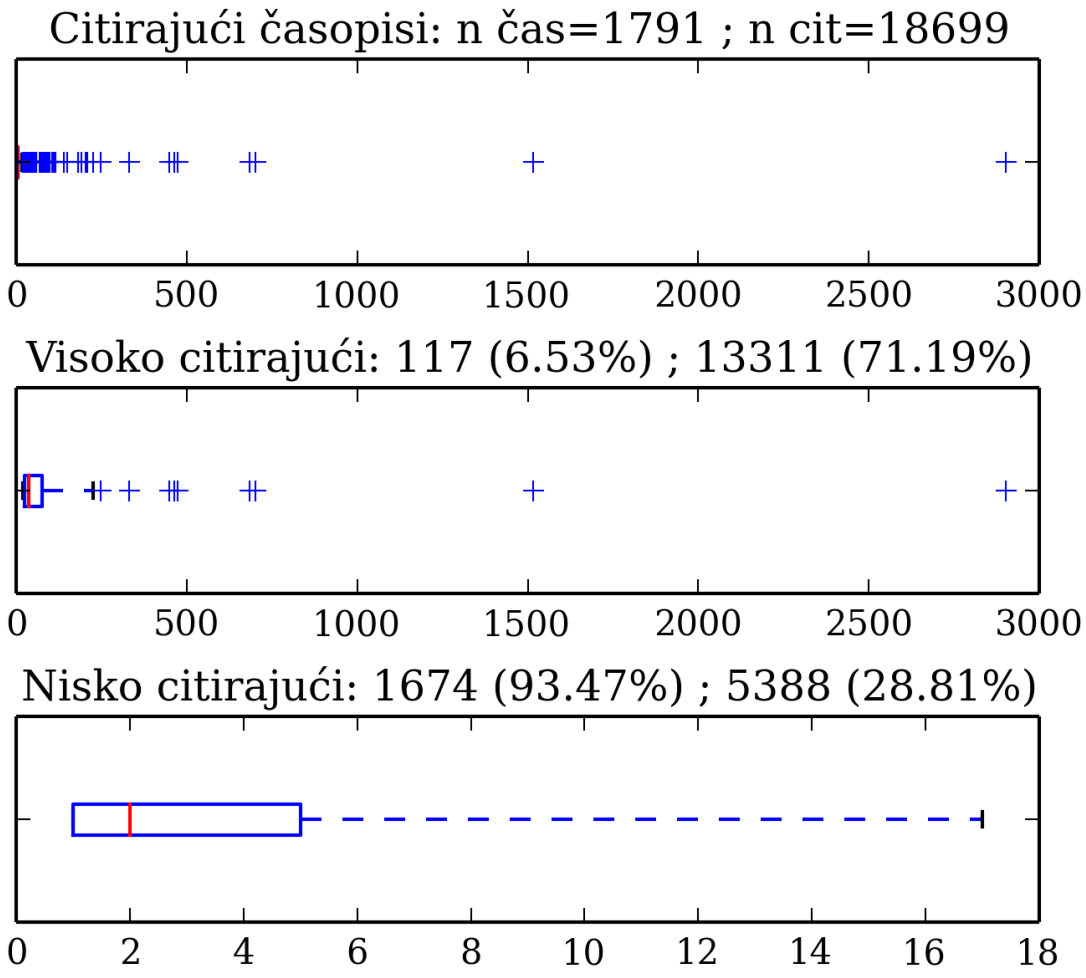
Među časopisima koji najučestalije citiraju Scienotmetrics (tablica 21) nema puno iznenađenja i oni odražavaju časopise koji uz *Scientometrics* prate scientometriju, što je u uvodnom dijelu doktorata i spomenuto. JASIST i *Journal of Informetrics* zajedno pružaju 15% svih citata što zajedno s časopisom *Scienometrics* odnosi oko 50% dobivenih citata i govori o povezanosti ovih časopisa. Tu su i drugi časopisi iz područja informacijskih znanosti poput *Journal of Documentation* ili *Online Information Review*, ali nailazimo i na časopise iz tzv. STM područja poput PLoS One i *Czechoslovak Journal of Physics*. Zanimljivo je da su prirodnjaci, posebno fizičari, kemičari, matematičari i biolozi svojim radovima iz scientometrije doprinijeli njenom razvoju.

S obzirom na velik broj časopisa koji se pozivaju na radove iz *Scientometricsa*, korisno je promotriti koliko časopisa često citira časopis *Scientometrics* i za koji postotak ili udio citata su zaslužni. Visoko citirajući časopisi određeni su putem interkvartilnog raspona, kao i kod odabira visoko citiranih članaka, ali za časopise su radi navedenog, odabrani snažni ekstremi

odnosno Q3 + IQR * 3 je uzeta kao granična vrijednost. Navedena vrijednost za ovaj skup radova iznosi 17 citata.

Distribucija citata po citirajućim časopisima kao i podjela na visoko citirajuće i ostale, vidljiva je iz slike 30. S obzirom na velik broj samocitata časopisa, koje je najbolje interpretirati zasebno, časopis *Scientometrics* je isključen iz ovog prikaza.

Slika 30. Distribucija citata na članke u časopisu *Scientometrics* po citirajućim časopisima i visoko citirajući časopisi



Veliki ekstremi kod časopisa su očekivani budući da je razumno pretpostaviti da će radovi koji često navode radove iz časopisa *Scientometrics* biti grupirani u časopisima srodnih područja. Kao što vidimo, može se identificirati 117 časopisa (6,5%) koji pružaju 71,2% od svih citata koje je časopis *Scientometrics* primio, a da nisu samocitati časopisa.

Navedeno pruža širu sliku o "susjedstvu" časopisa *Scientometrics* izvan nekoliko najvezanijih časopisa koji su prethodno već spomenuti i za koje se može pretpostaviti da su najvažniji za razvoj područja scientometrije pored samog časopisa *Scientometrics*. Ako ignoriramo samocitate časopisa koji odnose oko trećinu citata i promotrimo druge dvije trećine, odnosno samo citate iz ostalih časopisa indeksiranih u WoS, možemo identificirati 117 časopisa koji odnose većinu citata prema časopis *Scientometrics* što nije malen broj časopisa i što ide u prilog tvrdnji o širini interesa za scientometriju.

3.3.5 Radovi i časopisi citirani u časopisu *Scientometrics*

Do sada smo promatrali članke u časopisu *Scientometrics* kao citirane članke, a kao citirajuće članke samo u kontekstu samocitata časopisa. U ovom poglavlju govori se o člancima objavljenim u *Scientometricsu* kao citirajućim člancima, a o publikacijama koje se nalaze na popisima literature u tim člancima kao citiranim publikacijama. Prilikom istraživanja nekog tijela znanstvene literature, naime, vrlo je informativno proučiti citatne navode na kojima počiva, jer podaci o radovima s popisa literature pružaju uvid u specifičnosti i razvoj područja.

Pregled članaka u časopisu *Scientometrics* kao citirajućih radova prema razdobljima prikazan je u tablici 22.

Tablica 22. Pregled podataka o citiranim publikacijama u člancima objavljenim u časopisu *Scientometrics* 1978-2010. u odnosu na razdoblja

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
n citirajućih članaka	317	775	1342	2434
n citata	6373	14479	35440	56292
avg citata po citirajućem članku	20.1	18.7	26.4	23.1
n citiranih publikacija	4865	9853	20782	32555
n citiranih časopisa	743	1340	2921	4004
avg citiranih časopisa po citirajućem članku	6.7	6.1	9.9	8.3

Prema WoS podacima, pronađeno je 56 292 citata na oko 32 555 različitih publikacija i to iz 4 004 časopisa što ide u prilog ranijim tvrdnjama o širini i interdisciplinarnosti područja. Kao što je u metodologiji opisano, pokazatelje vezane uz publikacije citirane u časopisu *Scientometrics* treba smatrati manje preciznima od pokazatelja vezanih za citiranost časopisa *Scientometrics*. Broj citiranih časopisa raste s 743 u prvom razdoblju do 2 921 u posljednjem

razdoblju, a iz velike brojke časopisa u svim razdobljima, vidljivo je da se mnogi časopisi ne pojavljuju u više razdoblja. Situacija, je slična i s citirajućim časopisima, što je prethodno već i spomenuto.

Tablica 22 prikazuje i porast u broju referenci kroz vrijeme, tj. u broju različitih citiranih časopisa iz prva dva razdoblja (oko 20 citatnih navoda u popisu literature i 6 časopisa po članku) u odnosu na posljednje razdoblje (oko 26 citatnih navoda u popisu literature i 10 časopisa po članku).

U tablici 23. navedeni su podaci o citiranim radovima prema vrstama radova po tematici iz časopisa *Scientometrics*.

Tablica 23. Pregled podataka o citiranim publikacijama u člancima objavljenim u časopisu *Scientometrics* 1978-2010. u odnosu na razdoblja

pokazatelj	primijenjeni	metodološki	teorijski
n citirajućih članaka	1367	694	373
n citata	30056	16297	9939
avg citata po citirajućem članku	22.0	23.5	26.6
n citiranih publikacija	19944	10758	7420
n citiranih časopisa	2900	1540	1045
avg citiranih časopisa po citirajućem članku	8.1	8.4	8.6

Iz tablice je vidljivo da članci različitih tematika u časopisu *Scientometrics* pokazuju vrlo slične, gotovo iste, prosječne trendove navođenja literature, s nešto većim prosječnim brojem citata po citirajućem članku kod teorijskih radova. Sve vrste članaka citiraju sličan broj časopisa, ali je broj citiranih časopisa kod primijenjenih istraživanja ukupno oko tri puta veći od skupa časopisa citiranih u teorijskim člancima odnosno oko dva puta veći u odnosu na metodološke članke.

3.3.6 Citirani časopisi u člancima u časopisu *Scientometrics*

Da bi se dobila cjelovitija slika komuniciranja u scientometriji, uz već prikazane časopise koji su citirali časopis *Scientometrics*, u ovom poglavlju bit će govora o tome koji su to časopisi na kojima počivaju radovi objavljeni u časopisu *Scientometrics*.

Same informacije o brojevima časopisa iznesene u tablicama 22 i 23 ne prikazuju temeljne časopise u člancima objavljenim u časopisu *Scientometrics* tj. temeljne časopise za razvoj

scientometrije. Časopisi koji su najčešće citirani u promatranom skupu članaka su vidljivi iz tablice 24.

Tablica 24. Top 20 najčešće citiranih časopisa u člancima objavljenim u časopisu *Scientometrics*

časopis	n citata	n citiranih radova	% citirajućih radova	avg citata iz Scientometrics
J AM SOC INF SCI TEC*	2447	693	39.2	2.5
RES POLICY	1641	509	24.2	2.7
SCIENCE	940	395	23.7	1.6
SOC STUD SCI	783	213	16.8	1.9
NATURE	719	323	18.7	1.6
J DOC	706	230	15.2	1.9
J INFORM SCI	499	141	14.6	1.4
INFORM PROCESS MANAG	420	143	12.5	1.4
AM SOCIOL REV	320	101	7.7	1.7
RES EVALUAT	276	125	8.4	1.3
P NATL ACAD SCI USA	262	73	7.8	1.4
AM PSYCHOL	236	107	6.1	1.6
SCI PUBL POLICY	211	79	6.0	1.4
SCI STUD	160	34	4.4	1.5
AM J SOCIOL	159	64	4.8	1.3
MINERVA	149	80	4.8	1.3
AM ECON REV	143	69	4.0	1.5
J INFORMETR	142	60	3.5	1.7
BRIT MED J	141	69	4.3	1.3
ANNU REV INFORM SCI	130	35	4.3	1.2

*J AM SOC INFORM SCI (JASIS) i J AM SOC INF SCI TEC (JASIST) su tretirani kao isti časopis.

Iz ove tablice ispuštene su informacije o samocitatima časopisa *Scientometrics*, budući da su one već opisane prilikom opisa časopisa koji su citirali časopis *Scientometrics*. Kod ostalih časopisa citiranih u časopisu *Scientometrics*, očekivano nalazimo časopise koji ga i citiraju s nekim razlikama. JASIST je i dalje na vrhu popisa, 40% svih članaka u časopisu *Scientometrics* citira ovaj časopis što uz informaciju o citiranosti *Scientometricsa* u JASIST-u ukazuje na snažnu povezanost između ovih časopisa.

Vrlo je važna i pozicija časopisa *Research Policy*, ne samo kao časopisa koji citira radove iz *Scientometricsa* nego i kao časopisa kojega citiraju autori koji objavljuju članke u *Scientometricsu*. Ova činjenica govori upućuje na važnost scientometrijskih istraživanja u svrhe znanstvene politike. Iako se ne nalaze među prvih pet časopisa koje najčešće citiraju autori koji pišu u *Scientometricsu*, tri značajna časopisa iz područja informacijskih znanosti, kao i u uzorku časopisa koji najčešće citiraju radove iz *Scientometricsa* nalaze se *Journal of Documentation*, *Journal of Information Science* i *Information Processing & Management*.

Časopisi čiji su radovi citirani kao literatura u člancima objavljenim u časopisu *Scientometrics*, međutim, pokazuju veću tematsku raznolikost od citirajućih časopisa. Zanimljivo je da je od pet najčešće citiranih časopisa u radovima u *Scientometricsu*, jedino tj. JASIST iz područja informacijskih znanosti i to na prvom mjestu. Ostali časopisi prikazuju i sociološke odnosno organizacijske aspekte scientometrije pa kao što je i u uvodu spomenuto, naglasak u istraživanjima na prirodnim i primijenjenim znanostima, vidljiva kroz časopise poput *Science* ili *Nature*, kao najprestižnije časopise za ta područja znanosti.

U tablici 24, pokazatelj "prosjek citata iz Scientometrics" je namijenjen kao dodatna informacija za interpretaciju i sadrži prosječan broj citata koje su radovi i prilozi u nekom časopisu primili iz članaka objavljenih u časopisu *Scientometrics*. Čim je ovaj pokazatelj veći, tim su individualni članci objavljeni u nekom časopisu češće citirani u časopisu *Scientometrics*. Visoki "prosjek citata u Sci" i niski broj citiranih radova ukazuje da je časopis objavio mali broj visoko relevantnih radova za članke objavljene u *Scientometricsu*, a nizak "prosjek citata u Sci" uz veliki broj radova ukazuje da časopis češće objavljuje radove citirane u časopisu *Scientometrics*, ali da se mali broj tih radova citira više puta.

Rangiranje citiranih časopisa na ovaj način ima važan nedostatak: časopisi koji su objavili manje radova (često zato jer kraće izlaze od ostalih) niže su rangirani. Popis nam dakle precizno govori na kojim časopisima scientometrijski radovi trenutačno počivaju, ali za odgovor na pitanje koji su časopisi najvezaniji uz scientometriju potreban bi nam bio i udio citiranih radova iz svakog od citiranih časopisa. Na ovaj način bi se učinila razlika između časopisa koji su dobili više bruto citata u časopisu *Scientometrics*, ali je citiran vrlo mali udio svih objavljenih radova u tim časopisima od časopisa koji su dobili manje citata u *Scientometrics*, ali je citiran veći udio svih radova u tim časopisima. Taj pokazatelj je, međutim, izvan dosega ovog istraživanja jer zahtijeva podatke o kumulativnom broju

objavljenih radova u svakom od časopisa za svaku od godina u promatranom razdoblju odnosno 1978-2010.

3.3.7 Starost citata prema i iz časopisa *Scientometrics*

Starost citata odnosno navođene literature ovisno je o području i vrsti rada. S obzirom na dinamičnu prirodu citatnih analiza, informacije o starosti citirane literature informiraju o primjerenosti vremenskog okvira u kojem se provodi analiza, što je detaljnije opisano u poglavlju metodologije. Analiza starosti citata pruža, dakle, važne informacije za dizajn primijenjenih bibliometrijskih istraživanja ili za korištenje indikatora u praksi.

Važna tema u citatnim analizama je tema zastarijevanja literature. Kao i kod ostalih scientometrijskih tema, od samih početaka autori upozoravaju da je tema zastarijevanja i posebno razloga zastarijevanja kompleksna da bi je se olako interpretiralo s nekoliko jednostavnih pokazatelja (Line i Sandison, 1974). U ovom istraživanju podaci o starosti literature opisani su kako bi obogatili sliku o trendovima citiranja časopisa *Scientometrics* kao i trendove citiranja (tj. navođenja literature) u časopisu *Scientometrics*. Za precizniju interpretaciju ovih podataka u smislu zastarijevanja scientometrijske literature, bilo bi potrebno provesti i dodatne analize, a poželjno ih i proširiti kvalitativnim upitima što je izvan dosega ovog istraživanja.

Kod područja u brzom, posebno tehnološkom, razvoju literatura brže zastarijeva nego kod ostalih područja (Line, 1993). Situacija, kulminira u područjima u kojima je komunikacija toliko brza da je objavljivanje na konferencijama primaran oblik komunikacije, poput nekih područja računalnih znanosti (npr. područje računalne sigurnosti). Promatranje zastarijevanja literature u nekom području informira i o samom kolanju znanja unutar područja kao i o vremenskom okviru koji je primjeren za scientometrijske pokazatelje o citiranosti (pogotovo za potrebe vrednovanja znanstvenog rada).

Kao i kod ostatka citatnih analiza, pitanja o zastarijevanju scientometrijske literature možemo postaviti iz dva smjera:

- Koliko dugo radovi objavljeni u časopisu *Scientometrics* dobivaju citatate?

- Koja je starost literature koja se citira u člancima objavljenim u časopisu *Scientometrics*?

Na prvo pitanje se može odgovoriti tablicom 25, koja prikazuje starost citata na članke objavljene u časopisu *Scientometrics* 1978-2010.

Tablica 25. Starost citata članaka u časopisu Scientometrics u odnosu na razdoblja

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
medijan starost citata	13.0	10	4.0	5.0
u Scientometrics	12	8	3.5	5.0
u ostalim časopisima	14.0	11.0	4	5.5
maks. starost citata	34	23	12	34
u Scientometrics	33	23	12	33
u ostalim časopisima	34	23	12	34

U tablici 25, s obzirom da velik udio citata (36,6%) otpada na samocitate časopisa, informacije su podijeljene i na citate iz ostalih časopisa te citate iz časopisa *Scientometrics* kako bi se utvrdilo postoji li razlika u starosti citiranja iz matičnog časopisa u odnosu na ostale časopise.

Kao što vidimo, medijan starost citata tj. 50% svih citata koje članci primaju u razdoblju 1978-2012, najveći je u prvom razdoblju i iznosi 13 s minimalnim razlikama u samocitatima časopisa od citata u ostalim časopisima. U drugom razdoblju, medijan je 8 što ide u prilog većem broju važnih teorijskih radova u prvom razdoblju. Obzirom na značajke citiranja u prva dva razdoblja, razlog malom medijanu u zadnjem vremenskom razdoblju je što nije prošlo dovoljno vremena za povećanje. Navedeno ide u prilog većem potrebnom odmaku za citatne analize od dvije godine za scientometrijske radove i to pogotovo u evaluativne svrhe. Za usporebu, drugi autori prijavljuju prosječnu starost citata od 16 u antropologiji, 9 u ekonomiji i 7 u sociologiji (Sangam, 1999), odnosno 3 u računalnim znanostima početkom 1990-ih (Cunningham i Bocock, 1995) te 10 godina u matematici, 8 u kemiji i 4 u fizici (Gupta, 1997). Zanimljiva činjenica je da se u svakom razdoblju maksimalna starost citata proširuje kako bi zahvatila početak objavljivanja časopisa *Scientometrics*, što ide u prilog već spominjanoj ulozi teorijskih radova u formiranju discipline.

Iste pokazatelji u odnosu na tematiku radova vidljivi su u tablici 26.

Tablica 26. Starost citata članaka u časopisu Scientometrics u odnosu na tematiku

pokazatelj	primijenjeni	metodološki	teorijski
medijan starost citata	5	5.0	7.0
u Scientometrics	5.0	5	6.0
u ostalim časopisima	5.0	5	7
maks. starost citata	33	33	34
u Scientometrics	32	31	33
u ostalim časopisima	33	33	34

Uvid u ove podatke nastavlja pokazivati konzistentost u starosti citata časopisa *Scientometrics*, ali s nešto većom prosječnom starosti citata na teorijske radove, što je i očekivano.

Ukoliko se pak usredotočimo na starost citata koje časopis *Scientometrics* pruža, podaci su opet konzistentni kroz razdoblja kao što je vidljivo na tablici 27.

Tablica 27. Medijan starosti citiranih publikaciju u časopisu *Scientometrics* prema godinama objave

razdoblje	članci u časopisima	ostale publikacije
1978-1988	7.0	8
1989-1999	7.0	6.0
2000-2010	7	7
1978-2010	7	7

Prema tablici 27, prosječna starost citirane literature u časopisu *Scientometrics* je 7 godina i to bez bitnih razlika između članaka u časopisima i ostalih publikacija.

3.3.8 Samocitati

Samocitati su važna tema za razumijevanje citatnih analiza. Podaci o samocitiranosti preneseni su u tablici 28.

Tablica 28. Samocitati autora i časopisa

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
n citata	5118	9008	15356	29482
n samocitata časopisa	1978	3795	5010	10783
n samocitata autora	600	1341	2461	4402
u Scientometrics	288	698	905	1891
u ostalim časopisima	312	643	1556	2511

U ovoj tablici su ponovljene informacije o samocitatima časopisa s obzirom da se radi trećini citata koje je časopis *Scientometrics* primio. Samocitati autora pokazuju očekivano manji udio od ukupnog broja samocitata časopisa, odnosno 15% od svih citata koje s. Kod samocitata autora je vidljivo da ih se više pruža iz ostalih časopisa nego časopisa *Scientometrics* što opet odražava već opisanu prirodu scientometrije u kojoj često objavljuju autori iz raznih područja, a možda i "scientometriju za znanstvene discipline". Budući da je tema samocitata prilično zahtjevna za detaljna istraživanja, u kojem smislu se često koriste i kvalitativne metode, u ovom radu nije detaljnije razrađena.

4 ZAKLJUČAK

Cilj ovoga rada bio je istražiti, analizirati, sintetizirati i interpretirati razvoj scienotmetrije kao mlade znanstvene discipline unutar informacijskih znanosti. Da bi se postigao željeni cilj, bilo je nužno razviti i primijeniti pouzdan metodološki instrumentarij i pristup. Kako je časopis Scientometrics ključni nositelj informacija o razvoju scientometrije, kao korpus ovog istraživanja korišteni su svi radovi i prilozi objavljeni u časopisu Scientometrics, od početka izlaženja 1978. godine, do zaključno s 2010-om. godinom. Većina analiza provedena je na člancima kao primarnim tekstovima scientometrije i homogenom (u smislu vrste odnosno uloge rada) skupinom radova.

Kako bi se dobila što cjelovitija slika stanja i prikazali razni metodološki postupci korisni u scientometriji, u ovom su radu prikazani rezultati relativno velikog broja scientometrijskih analiza. S obzirom da tematske podjele rezultata analiza uvjetuju fragmentiranost prikaza pokazatelja relevantnih za opis razvoja discipline, u ovom poglavlju je prikazan odabir najrelevantnijih pokazatelja. Radi preglednosti i lakšeg dobivanja uvida u prikaz najrelevantnijih rezultata, izrađene su tablice 29 i 30. Ove tablice u osnovi predstavljaju skraćeni scientometrijski profil časopisa kroz koji se može interpretirati razvoj znanstvene discipline. Prva prikazuje pokazatelje vezane uz radove, a druga pokazatelje vezane uz autore.

Tablica 29. Pregled pokazatelja o člancima objavljenim u časopisu *Scientometrics* 1978-2010 kao indikatora razvoja scientometrije

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
n članaka (% od ukupno radova i priloga)	325 (64.9)	791 (33.3)	1353 (22.3)	2469 (11.1)
citati koje članci primaju u razdoblju 1978-2012				
udio citiranih	95,1	94,1	92	93,1
medijan citata	8	7	6	6
n citirajućih časopisa	544	840	1306	1792
udio samocitata autora / časopisa	11,7/38,6	14,9/42,1	16/32,6	14,9/36,6
medijan/maksimum starosti primljenih citata	13/34	10/23	4/12	5/34
citati koje članci pružaju u razdoblju 1978-2010				
prosječan broj citiranih publikacija po članku	20,1	18,7	26,4	23,1
medijan starosti citiranih publikacija po članku	7	6	6	6,5
prosječan broj različitih citiranih časopisa po članku	6,7	6,1	9,9	8,3
ukupan broj citiranih časopisa	743	1340	2921	4004
autorstvo članaka				
udio višeautorskih	43,7	51,7	68,4	59,8
medijan autora po članku	1	2	2	2
ostali pokazatelji vezani uz članke				
udjeli primjenjenih, metodoloških i teorijskih članaka	50/25/25	55/27/18	58/30/12	56/28/15
***h*-indeks / *g*-index**	35/55	37/58	45/72	61/90

Budući da se radi o jednom, ali ključnom časopisu za disciplinu, broj radova odosno članaka praćen kroz vrijeme relativno je pouzdan indikator i upućuje na sve veću zainteresiranosti znanstvenika za scientometriju. Broj članaka u časopisu Scientometrics kroz promatrana razdoblja pokazuje izraziti trend rasta. Postotak rasta je u razdoblju 1989-1999 u odnosu na 1978-1988 bio 143,3%, a u razdoblju 2000-2010, u odnosu na 1989-1999 je iznosio 71%. Postotak rasta članaka je izraženiji od postotka rasta svih radova i priloga (iz drugog u odnosu na prvo 94.6%, to jest iz trećeg u odnosu na drugo 71%) jer udio broja članaka raste kroz vrijeme.

Udio članaka u odnosu na ukupan broj svih radova i priloga raste s 64,9% u prvom razdoblju do 92,8% u zadnjem razdoblju, a u cijelom skupu radova iznosi 84,2%. Analizom vrste radova i priloga utvrđene su promjene u njihovoj pojavnosti. Neke vrste priloga poput bibliografija, podatkovnih izvještaja i obavijesti, u ovom znanstvenom časopisu, s vremenom izlaze sve rijeđe. Može se zaključiti da u doba interneta koje karakterizira sve šira dostupnost informacija uz sve veću efikasnost računala, odnosno promjena paradigme komuniciranja i obrade podatake, ovakve publikacije gube na značaju. Budući da je porast broja članaka u trećem razdoblju u odnosu na drugo jednak porastu broja svih radova, možemo zaključiti da su promjene u objavama različitih vrsta radova nastuple u 90-ima što se podudara s pojavom i popularizacijom weba.

Podaci o citiranosti radova ukazuju na prihvaćenost i relevantnost članaka objavljenih u ovom časopisu u svim vremenskim razdobljima. Također, velik udio radova dobiva više citata što ukazuje na relevantnost cijelog tijela literature, više nego uključivanja manjeg broja "super-citiranih" članaka. Preko 90% članaka objavljenih u časopisu Scientometrics je citirano u svim razdobljima, s blagom prednošću ranijih razdoblja koja su imala više vremena biti citirana. Medijan citata članaka u cijelom razdoblju iznosi 6, što ukazuje na relativno visoku citiranost individualnih članaka.

Prosječna citiranost po članku kroz godine kao i za cjelokupno razdoblje vrlo je slična, što ukazuje na sličan status scientometrijskih članaka od početka objavljivanja časopisa. Prednost u citatima imaju stariji radovi što nije neobično pogotovo u odnosu na medijan starosti citata. Pored toga što su ovi radovi imali najviše vremena prikupiti citate, za očekivati je da se u prvim godinama časopisa objavljivalo više teorijskih radova s ciljem definiranja područja koji ostaju kontinuirano relevantni.

Više od 50% članaka objavljenih u časopisu *Scientometrics* u prva dva razdoblja dobivaju citate još barem 10 godina što je vidljivo iz medijana starosti primljenih citata u tablici 29. što ukazuje na relativno sporo zastarijevanje literature. Za usporebu, drugi autori prijavljuju prosječnu starost citata od 16 u antropologiji, 9 u ekonomiji i 7 u sociologiji (Sangam, 1999), odnosno 3 u računalnim znanostima (Cunningham i Bocock, 1995) te 10 godina u matematici, 8 u kemiji i 4 u fizici (Gupta, 1997). Maksimum citata u svakom od razdoblja odgovara duljini perioda od početka razdoblja do 2012. godine odnosno do kraja razdoblja u kojem se promatra citiranje. Navedeno ukazuje da radovi iz svih prikazanih razdoblja još uvjek dobivaju citate.

Analiza kocitiranosti najcitiranijih članaka prikazala je kocitatne grupe tih radova. Dobivene grupe odgovaraju interpretaciji da je evaluativna scientometrija najčešće primijećen dio scientometrije budući da najveća komponenta mreže kocitata (koja sadrži 90 odnosno 46% svih visoko citiranih članaka) prikazuje temu citatnih analiza za potrebu vrednovanja kao i vezanih pokazatelja poput *h*-indeksa. Druga najveća komponenta (29 čvorova odnosno 15% svih visoko citiranih članaka) vezana je uz mapiranje znanosti s posebnim naglaskom na odnos između znanosti i tehnologije.

Od svih citata koje su dobili članci u časopisu *Scientometrics*, 15% su samocitati autora. Obzirom na problematiku detekcije samocitata autora, ove podatke teško je usporediti s drugim područjima. Samocitati časopisa, pak, odnose 36,6% ukupno primljenih citata što je visoka samocitiranost, ali je i očekivana obzirom na jedinstvenu specijalizaciju časopisa u promatranom razdoblju. Za usporedbu s područjem s većim brojem časopisa može poslužiti istraživanje (Krauss, 2007) u kojem se proučilo 107 časopisa iz područja ekologije i pronašlo prosječno 12% samocitata časopisa.

Među časopisima koji najčešće citiraju časopis *Scientometrics*, JASIST i *Journal of Informetrics* zajedno pružaju 15% svih citata što zajedno s časopisom *Scienometrics* odnosi oko 50% dobivenih citata i govori o povezanosti ovih časopisa. Tu su i drugi časopisi iz područja informacijskih znanosti poput *Journal of Documentation* ili *Online Information Review*, ali nailazimo i na časopise iz tzv. STM područja poput PLoS One i *Czechoslovak Journal of Physics* što ide u prilog spominjanoj naklonosti scientometrije STM područjima i to pogotovo, radi dostupnih izvora podataka, u ranijim razdobljima.

Članci objavljeni u časopisu *Scientometrics* u prosjeku citiraju 23 publikacije s blagim povećanjem s 20 u prvom na 26 u zadnjem razdoblju. Prosječan broj različitih citiranih časopisa po članku raste s 6,7 u prvom na 9,9 u zadnjem razdoblju. Što se samih časopisa tiče, JASIST je i dalje na vrhu popisa: 40% svih članaka u časopisu *Scientometrics* citira ovaj časopis što uz informaciju o citiranosti *Scientometricsa* u JASIST-u ukazuje na snažnu povezanost između ovih časopisa. Uz ostale vezane časopise iz područja informacijskih znanosti, *Journal of Documentation*, *Journal of Information Science* i *Information Processing & Management*, članci u *Scientometricsu* se često pozivaju i na prestižne časopise iz područja prirodnih i primijenjenih znanosti *Science* ili *Nature* kao i časopisa koji odražavaju sociološku i organizacijsku prirodu scientometrije poput *Research Policy* i *Research Evaluation*.

Trendovi višeautorstva u časopisu Scientometrics pokazuju trendove slične onima u društvenim znanostima (Jokić i Zauder, 2013) što, uz informacije o citiranim i citirajućim časopisima, ide u prilog tretiranju scientometrije kao grane informacijskih znanosti. Udio višeautorskih članaka u prvom periodu iznosi 43,7% i raste do 68,4% u zadnjem periodu. Bez obzira na rast u udjelu višeautorskih radova, članci objavljeni u časopisu Scientometrics pokazuju "normalno" autorstvo u svim vremenskim periodima s relativno velikim udjelom jednoautorskih radova i s medijanom broja autora po članku koji iznosi dva za cijeli skup te za drugo i treće razdoblje. U prvom razdoblju 56% radova su jednoautorski.

Tablica 30. Pregled pokazatelja o autorima članaka objavljenih u časopisu *Scientometrics* 1978-2010 kao indikatora razvoja scientometrije

pokazatelj	1978-1988	1989-1999	2000-2010	1978-2010
n autora	350	807	1755	2595
medijan n članaka	1	1	1	1
medijan citata	9	7	7	7
n zemalja autora	26	39	63	67

Broj autora u uzorku raste shodno broju članaka, s porastom od 130, 6% u drugom razdoblju u odnosu na prvo, te 118% u trećem razdoblju u odnosu na prvo. Izraženiji porast u broju autora u trećem razdoblju u odnosu na porast u broju članaka (71%) vezan je i uz porast u udjelu višeautorskih radova.

U časopisu Scientometrics objavljuju podjednako autori iz više zemalja s različitih dijelova svijeta što ide u prilog tvrdnji da se radi o svjetski relevantnom časopisu, ali i o svjetski relevantnoj i prihvaćenoj tematici kada se interpretira uz podatke o porastu broja članaka i autora. Na popisu zemalja s više od 100 radova objavljenih u časopisu Scientoemtrics, zastupljena su tri kontinenta, na popisu zemalja s više od 40 radova, pet kontinenata, a na popisu s više od 20 objavljenih radova, šest kontinenata. Uz spomenutu Mađarsku i Nizozemsku, visoko produktivne autori u području scientometrije dolaze iz SAD, Belgije, Španjolske, Engleske, Indije, Francuske, Kine i Njemačke s čestom međunarodnom suradnjom.

Produktivnost autora u skladu je s interdisciplinarnom prirodom Scientometrije u svim razdobljima. Za pretpostaviti je da će velik broj, pogotovo primjenjenih članaka, objavljivati i znanstvenici iz raznih polja znanosti koji nemaju primaran interes za samu scientometriju već objavljuju istraživanja znanstvenih publikacija u vlastitom području. Vrlo mali udio autora je visoko produktivan u časopisu Scientometrics, a broj različitih autora u Scientometrics se kontinuirano povećava. Visoko produktivni autori koji kontinuirano objavljuju mogu se shvatiti ključnim za razvoj područja. Bez obzira na velik broj nisko produktivnih autora, u istraživanju su pronađeni autori koji nose značajan dio razvoja scientometrije.

Među najproduktivnijim autorima svakako treba istaknuti rad dvije europske grupe autora. Prva je vezana uz osnivače časopisa i situirana u Budimpešti. Riječ je o suradnji urednika i osnivača časopisa Tibora Brauna s najproduktivnijim autorom u promatranom skupu, Wolfgangom Glänzelom te Andrasom Schubertom. Druga grupa autora je leidenska grupa iz CWTS-a (The Centre for Science and Technology Studies, Leiden) odnosno Moed, van Raan i van Leeuwen čija suradnja je vidljiva u svim aspektima scientometrije.

Analiza suradnje putem mreža pokazala je stvaranje centralne komponente autora, što je važno za promatranje scientometrije kao kohezivnog područja. Gledano kroz vrijeme, ovaj skup radova počinje pokazivati kohezivnu strukturu koautorstva u 90-ima, a u 2000-ima poprima strukturu koja je, i radi povećanja broja radova i radi povećanja broja autora u ovom razdoblju, markantna za cijelo promatrano razdoblje. Navedeno, uz informacije o citirajućim i citiranim časopisima, ukazuje na konsolidaciju područja scientometrije kao zasebne discipline informacijskih znanosti. Bez obzira na kontinuirano velik broj nepovezanih autora područje pokazuje dovoljan broj kontinuiranih autora i takve značajke suradnje kako se može smatrati zasebnom disciplinom.

Kako bi prikazani rezultati bili relevantni i što bliži pravom stanju korištenih indikatora, bilo je potrebno osmisliti, kreirati i odraditi cjelovit pristup operacionalizaciji istraživanja koji je proveden programskim jezikom Python. Dio postupaka odrađen je standardiziranim, dio posebno razvijenim općenitim alatima (odnosno modulima za programski jezik Python), koji omogućuju dubinsku provjeru postupak i brz razvoj, a cjelokupni postupak implementiran je Python skriptama posebno pisanim za ove podatke. Naveden pristup omogućuje reaktivnost u kontekstu obrade nekog specifičnog skupa podataka te omogućuje ponovnu provedbu svih postupaka od pripreme podataka do izrade tablica i vizualizacija korištenih u tekstu.

S obzirom na znatnu podatkovnu problematiku u području, tradiciju adaptacije ili razvijanja novih pokazatelja u scientometriji, te brz razvoj novih postupkaka i mogućnosti uslijed informatizacije, predloženi pristup ocijenjen je kao najprimjerenije rješenje za istraživanja koja imaju potrebu pripreme i upravljanja kompleksnim skupovima podataka.

Istraživanje je razvilo i mogućnosti uključivanja cjelovitih tekstova radova u scientometrijske analize, odnosno proširivanje tih analiza pokazateljima iz teksta članaka. U ovom segmentnu posla u radu naglasak je u prvom redu stavljen na omogućavanje postupaka a manje na teoretskim mogućnostima.

Prikazana je problematika i metoda izrade korpusa ovog izvora kao krucijalan prvi korak za omogućavanje korištenja sadržaja tekstova radova u scientometrijskim istraživanjima. Na ovaj način metodološki obrađeni podaci u radu su navedeni s ciljem da se naglase njegove mogućnosti. Bez obzira na poteškoće prilikom računalnog baratanja tekstovima radova, alati i spoznaje o takvim mogućnostima se rapidno razvijaju, a mogućnost uključivanja samih predmeta promatranja u kvantitativna istraživanja prevažna je kako bi ju se propustilo. Zbog opsežnosti ovog istraživanja, detaljniji rad na sadržajnim analizama cjelovitih tekstova ostavljen je za neka buduća istraživanja.

5 LITERATURA

Abt, Helmut A. (2007). The future of single-authored papers. *Scientometrics*, 73(3):353-358.

Abt, Helmut. (2000). Do Important Papers Produce High Citation Counts?. *Scientometrics*, 48(1):65-70.

Acedo, Francisco José; Barroso, Carmen; Casanueva, Cristóbal; Galán, José Luis. (2006). Co-Authorship in Management and Organizational Studies: An Empirical and Network Analysis. *Journal of Management Studies*, 43(5):957-983.

Aksnes, Dag W. (2003). A macro study of self-citation. *Scientometrics*, 56(2):235-246.

Beaver, D.; Rosen, R. (1978). Studies in scientific collaboration Part I: The professional origins of scientific co-authorship. *Scientometrics*, 1(1):65-84.

Beaver, D.; Rosen, R. (1979). Studies in scientific collaboration Part II: Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-830. *Scientometrics*, 1(2):133-149.

Beaver, D.; Rosen, R. (1979). Studies in scientific collaboration Part III: Professionalization and the natural history of modern scientific co-authorship. *Scientometrics*, 1(3):231-245.

Beck, James W.; Beatty, Adam S.; Sackett, Paul R. (2013). On the Distribution of Job Performance: The Role of Measurement Characteristics in Observed Departures from Normality. *Personnel Psychology*.

van den Besselaar, Peter; Heimeriks, Gaston. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, 68(3):377-393.

Bird, Alexander. (1998). Philosophy of science. London: UCL Press.

Björneborn, L.; Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14):1216-1227.

Bonitz, Manfred. (1994). The multidimensional space of Scientometrics; The Derek John de Solla Price awards 1984-1993. *Scientometrics*, 29(1):3-14.

Borgatti, Stephen P.; Mehra, Ajay; Brass, Daniel J.; Labianca, Giuseppe. (2009). Network Analysis in the Social Sciences. *Science*, 323(5916):892-895.

Boyack, Kevin W.; Klavans, Richard; Börner, Katy. (2005). Mapping the backbone of science. *Scientometrics*, 64(3):351-374.

Bradford, S. (1934). On the scattering of papers on scientific subjects in scientific periodicals. *Engineering*, 137(193):86-86.

Braun, Tibor; Glänzel, Wolfgang; Schubert, András. (1985). Scientometric Indicators: A Thirty-Two Country Comparative Evaluation of Publishing Performance and Citation Impact. Singapore: World Scientific.

Broadus, Robert N. (1987). Early approaches to bibliometrics. *Journal of the American Society for Information Science*, 38(2):127-129.

Cainelli, Giulio; Maggioni, Mario A.; Uberti, T. Erika; De Felice, Annunziata. (2010). The strength of strong ties: co-authorship and productivity among Italian economists.

Callon, Michel; Courtial, Jean-Pierre; Laville, Francoise. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22(1):155-205.

Cherny, Arkady; Gilyarevsky, Ruggero. (2001). The Impact of V.V. Nalimov on Information Science. *Scientometrics*, 52(2):159-163.

Costas, Rodrigo; Bordons, María. (2008). Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, 77(2):267-288.

Creath, Richard. (2013). Logical Empiricism. U: Edward N. Zalta, ur., The Stanford Encyclopedia of Philosophy.Stanford.

Cronin, B. (1984). The Citation Process: The Role and Significance of Citations in Scientific Communication. Taylor Graham.

Cunningham, Sally Jo; Bocock, David. (1995). Obsolescence of computing literature. *Scientometrics*, 34(2):255-262.

De Haan, J. (1997). Authorship patterns in Dutch sociology. *Scientometrics*, 39(2):197-208.

De Stefano, Domenico; Fuccella, Vittorio; Vitale, Maria Prosperina; Zaccarin, Susanna. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, 35(3):370-381.

Diodato, Virgil P. (1994). Dictionary of Bibliometrics.Taylor & Francis.

Duque, Ricardo B.; Ynalvez, Marcus; Sooryamoorthy, R.; Mbatia, Paul; Dzorgbo, Dan-Bright S.; Shrum, Wesley. (2005). Collaboration Paradox: Scientific Productivity, the Internet, and Problems of Research in Developing Areas. *Social Studies of Science*, 35(5):755-785.

Egghe, Leo. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1):131-152.

Feinberg, Gary; Watnick, Beryl; Sacks, Arlene. (2011). Solo vs. Collaborative Research in the Social Sciences and Higher Education: Unraveling the Realities of Male-Female Research Publication Patterns in the Context of Gender Politics and Social Justice Issues. *Journal of Multidisciplinary Research*. 3(3):None-None.

Feist, Gregory. (2006). The psychology of science and the origins of the scientific mind. New Haven: Yale University Press.

Feyerabend, Paul. (1975). Against method. London New York: Verso.

Fox, Mary. (2004). R. K. Merton - Life time of influence. *Scientometrics*, 60(1):47-50.

Garfield, Eugene. (1978). Editorial statements. *Scientometrics*, 1(1):3-8.

Garfield, Eugene. (1979). Is citation analysis a legitimate evaluation tool?. *Scientometrics*, 1(4):359-375.

Garfield, Eugene. (1998). From citation indexes to informetrics: is the tail now wagging the dog?. *Libri*, 48(2):67-80.

Garfield, Eugene. (2004). The intended consequences of Robert K. Merton. *Scientometrics*, 60(1):51-61.

Garfield, Eugene. (2010). The evolution of the Science Citation Index. *International Microbiology*, 10(1):65-69.

Gauffriau, Marianne; Larsen, Peder; Maye, Isabelle; Roulin-Perriard, Anne; von Ins, Markus. (2007). Publication, cooperation and productivity measures in scientific research. *Scientometrics*, 73(2):175-214.

Gisvold, S.E. (1999). Citation analysis and journal impact factors: is the tail wagging the dog?. *Acta anaesthesiologica scandinavica*, 43(10):971-973.

Glänzel, Wolfgang. (2014). Greetings from the new Editor-in-Chief. *Scientometrics*, 98(1):3-4.

Glänzel, Wolfgang; Schoepflin, URS. (1994). Little scientometrics, big scientometrics ... and beyond?. *Scientometrics*, 30(2-3):375-384.

Glänzel, Wolfgang; Schubert, András. (1992). Some facts and figures on highly cited papers in the sciences, 1981-1985. *Scientometrics*, 25(3):373-380.

Glänzel, Wolfgang; Schubert, András. (2005). Analysing scientific networks through co-authorship. U: None, ur., Handbook of quantitative science and technology research. None: None.

Godfrey-Smith, Peter. (2003). Theory and reality: an introduction to the philosphy of science. Chicago: The University of Chicago press.

Godin, Benoît. (2006). On the origins of bibliometrics. *Scientometrics*, 68(1):109-133.

Granovsky, Yuri V. (2001). Is it possible to measure science? V.V. Nalimov's research in scientometrics. *Scientometrics*, 52(2):127-150.

Gupta, B. M. (1997). Analysis of distribution of the age of citations in theoretical population genetics. *Scientometrics*, 40(1):139-162.

Gupta, Usha. (1990). Obsolescence of physics literature: Exponential decrease of the density of citations to Physical Review articles with age. *Journal of the American Society for Information Science*, 41(4):282-287.

He, Zi-Lin; Geng, Xue-Song; Campbell-Hunt, Colin. (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university. *Research Policy*, 38(2):306-317.

Henk F. Moed. (2005). Citation Analysis in Research Evaluation. Springer.

Henry, J. (2008). The Scientific Revolution and the Origins of Modern Science. None: Palgrave Macmillan.

Hicks, Diana. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2):193-215.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569-16572.

Hoffman, Paul. (1987). The man who loves only numbers. *Atlantic Monthly*, 260(5):60-74.

Jacsó, Péter. (2005). As we may search: Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89(9):1537-1547.

Jacsó, Péter. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, 34(1):175 - 191.

Jokić, Maja. (2005). Bibliometrijski aspekti vrednovanja znanstvenog rada. Zagreb: Sveučilišna knjižara.

Jokić, Maja; Zauder, Krešimir. (2013). Bibliometrijska analiza časopisa Sociologija sela/Sociologija i prostor u razdoblju 1963.- 2012. *Sociologija i prostor*, 51(2 (196)):331-349.

Jokić, Maja; Zauder, Krešimir; Letina, Srebrenka. (2012). Karakteristike hrvatske nacionalne i međunarodne znanstvene produkcije u društveno-humanističkim zanostima i umjetničkom području za razdoblje 1991.-2005. Zagreb: Institut za društvena istraživanja.

Katz, J. S.; Martin, B. R. (1997). What is research collaboration?. *Research policy*, 26(1):1-18.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10-25.

Krauss, Jochen. (2007). Journal self-citation rates in ecological sciences. *Scientometrics*, 73(1):79-89.

Kronegger, Luka; Ferligoj, Anuška; Doreian, Patrick. (2011). On the dynamics of national scientific systems. *Quality and Quantity*, 45(5):989-1015.

Kronegger, Luka; Mali, Franc; Ferligoj, Anuška; Doreian, Patrick. (2012). Collaboration structures in Slovenian scientific communities. *Scientometrics*, 90(2):631-647.

Kuzhabekova, Aliya. (2011). Impact of co-authorship strategies on research productivity: A social-network analysis of publications in Russian cardiology.

Lawani, Stephen M. (1981). Bibliometrics: its theoretical foundations, methods and applications. *Libri*, 31(1):294-315.

van Leeuwen, T.N. (2005). Descriptive versus evaluative bibliometrics. U: Moed, H.F. and Glänzel, W. and Schmoch, Ulrich, ur., Handbook of Quantitative Science and Technology Research. New York: Kluwer Academic Publishers.

van Leeuwen, Thed N.; Visser, Martijn S.; Moed, Henk F.; Nederhof, Ton J.; Van Raan, Anthony FJ. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, 57(2):257-280.

Letina, S.; Zauder, K.; Jokić, M. (2012). Produktivnost hrvatskih psihologa: Scientometrijska analiza mreže suradnji na radovima indeksiranim u bazi WoS 1991-2010. *Suvremena psihologija*, 15(1):97-116.

Levitt, Jonathan; Thelwall, Mike. (2009). The most highly cited Library and Information Science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics*, 78(1):45-67.

Leydesdorff, Loet. (1987). Various methods for the mapping of science. *Scientometrics*, 11(5-6):295-324.

Leydesdorff, Loet; Vaughan, Liwen. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and Technology*, 57(12):1616-1628.

Lotka, Alfred James. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*, 16():317-323.

MacRoberts, Michael H; MacRoberts, Barbara R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5):342-349.

Mali, Franc; Kronegger, Luka; Doreian, Patrick; Ferligoj, Anuška. (2012). Dynamic scientific co-authorship networks. U: Scharnhorst, Andrea and Börner, Katy and Besselaar, Peter van den, ur., Models of Science Dynamics. None: Sage.

Marshakova, Irina V. (1973). System of document connections based on references. *Scientific and Technical Information Serial of VINITI*, 6(2):3-8.

Melin, Göran; Persson, Olle. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36(3):363-377.

Milojević, Staša. (2010). Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology*, 61(7):1410-1423.

Moed, Henk F.; De Bruin, Renger E.; Van Leeuwen, Th N. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, 33(3):381-422.

Moody, James. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American sociological review*, 69(2):213-238.

Nalimov, VV; Mulchenko, ZM. (1969). Naukometria. Moscow: Nauka.

Narin, Francis. (1976). Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity. Cherry Hill, N.J.: Computer Horizons.

Narin, Francis. (2012). Decades of progress, or the progress of decades?. *Scientometrics*, 92(2):391-393.

Narvaez-Berthelemot, Nora; Russell, Jane M. (2001). World distribution of social science journals: A view from the periphery. *Scientometrics*, 51(1):223-239.

Nederhof, Anton J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review. *Scientometrics*, 66(1):81-100.

Nederhof, Anton J.; Zwaan, Rolf A.; De Bruin, Renger E.; Dekker, P. J. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social and behavioural sciences: A comparative study. *Scientometrics*, 15(5-6):423-435.

Newman, Mark. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5200-5205.

Newman, Mark. (2010). Networks: An Introduction. None: Oxford University Press, USA.

Osareh, Farideh. (1996). Bibliometrics, Citation Analysis and Co-Citation Analysis: A Review of Literature I. *Libri*, 46(3):None-None.

Perc, Matjaž. (2010). Growth and structure of Slovenia's scientific collaboration network. *Journal of Informetrics*, 4(4):475-482.

Peritz, Bluma C. (1983). Are methodological papers more cited than theoretical or empirical ones? The case of sociology. *Scientometrics*, 5(4):211-218.

Petek, Marija. (2008). Personal name headings in COBIB: Testing Lotka's Law. *Scientometrics*, 75(1):175-188.

Phelan, TJ. (1999). A compendium of issues for citation analysis. *Scientometrics*, 45(1):117-136.

Prell, Christina. (2011). Social Network Analysis: History, Theory and Methodology. None: SAGE Publications Ltd.

de Solla Price, Derek. (1963). Little science, big science. None: Columbia University Press.

de Solla Price, Derek. (1978). Editorial statements. *Scientometrics*, 1(1):3-8.

Pritchard, Alan. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 4(25):348-349.

van Raan, Anthony F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3):491-502.

van Raan, A. F. J. (1998). In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics*, 43(1):129-139.

van Raan, A. F. J. (2005). Measuring science: Capita Selecta of Current Main Issues. U: None, ur., Handbook of Quantitative Science and Technology Research. None: Springer.

van Raan, A. F. J. (1997). Scientometrics: State-of-the-art. *Scientometrics*, 38(1):205-218.

van Rijnsoever, Frank J; Hessels, Laurens K; Vandeberg, Rens LJ. (2008). A resource-based view on the interactions of university researchers. *Research Policy*, 37(8):1255-1266.

Rip, A.; Courtial, J. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6(6):381-400.

Rosenberg, A. (2005). Philosophy of Science: A Contemporary Introduction. Routledge.

Sangam, SL. (1999). Obsolescence of literature in the field of psychology. *Scientometrics*, 44(1):33-46.

Sarkar, Sahotra; Pfeifer, Jessica. (2006). The philosophy of science: an encyclopedia. New York: Routledge.

Schreiber, Michael. (2008). The influence of self-citation corrections on Egghe's g-index. *Scientometrics*, 76(1):187-200.

Schummer, Joachim. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics*, 59(3):425-465.

Seol, Sung-Soo; Park, Jung-Min. (2008). Knowledge sources of innovation studies in Korea: A citation analysis. *Scientometrics*, 75(1):3-20.

Shaban, Sami; Aw, Tar-Ching. (2009). Trend towards multiple authorship in occupational medicine journals. *J Occup Med Toxicol*, 4(3):None-None.

Shapin, Steven. (1996). The scientific revolution. Chicago, IL: University of Chicago Press.

Shapiro, Fred R. (1992). Origins of bibliometrics, citation indexing, and citation analysis: The neglected legal literature. *Journal of the American Society for Information Science*, 43(5):337-339.

Shenton, Andrew K; Hay-Gibson, Naomi V. (2009). Bradford's law and its relevance to researchers. *Education for Information*, 27(4):217-230.

Simonton, Dean Keith. (2004). Creativity in science: Chance, logic, genius, and zeitgeist. None: Cambridge University Press.

Sivertsen, Gunnar; Larsen, Birger. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: an empirical analysis of the potential. *Scientometrics*, 91(2):567-575.

Small, H.; Sweeney, E. (1985). Clustering the Science Citation Index using co-citations Part I: A Comparison of methods. *Scientometrics*, 7(3-6):391-409.

Small, H.; Sweeney, E.; Greenlee, E. (1985). Clustering the science citation index using co-citations Part II: Mapping science. *Scientometrics*, 8(5-6):321-340.

Small, Henry. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265-269.

Sonnenwald, DH. (2007). Scientific Collaboration: a Synthesis of Challenges and Strategies. *Annual Review of Information Science and Technology*, 4(None):1-37.

Sugimoto, Cassidy R. (2014). Blaise Cronin wins the 2013 Derek John de Solla Price Medal. *Scientometrics*, 98(1):5-10.

Torres-Salinas, D.; Moed, H.F. (2009). Library Catalog Analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in Economics. *Journal of Informetrics*, 3(1):9-26.

Uebel, Thomas. (2008). Logical Empricism. U: Psillos, Stathis and Curd, Martin, ur., The Routledge Companion to Philosophy of Science. London and New York: Routledge.

Vinkler, Péter. (1994). Words and indicators. As scientometrics stands. *Scientometrics*, 30(2-3):495-504.

Wagner, Caroline S.; Leydesdorff, Loet. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10):1608-1618.

Waugh, Joanne; Ariew, Roger. (2008). The history of philosophy and the philosophy of science. U: Psillos, Stathis and Curd, Martin, ur., The Routledge Companion to Philosophy of Science. London and New York: Routledge.

White, Howard D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54(5):423-434.

White, Howard D; McCain, Katherine W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4):327-355.

Whitley, R. (2000). The Intellectual and Social Organization of the Sciences. Oxford: Oxford University Press.

Yanovsky, VI. (1981). Citation analysis significance of scientific journals. *Scientometrics*, 3(3):223-233.

Yuh-Shan, Ho. (2004). Citation review of Lagergren kinetic rate equation on adsorption reactions. *Scientometrics*, 59(1):171-177.

Zauder, Krešimir; Pečarić, Đilda; Tuđman, Miroslav. (2011). Sources for Scientific Frustrations: Productivity and Citation Data.

Zitt, Michel. (1991). A simple method for dynamic scientometrics using lexical analysis. *Scientometrics*, 22(1):229-252.

6 DODATAK 1. PYTHON KÔD

Svi postupci opisani u doktoratu provedeni su u programskom jeziku Python. Većina izračuna je direktno implementirano, a za neke se koriste dodatni moduli Numpy i IGraph. Python kôd je podijeljen u dva osnovna modula phd_skripte i phd_alati. Nazivi i komentari u samom kôdu su na engleskom jeziku obzirom da su riječi korištene u Pythonu i njegovim modulima riječi iz engleskog jezika tj. engleski je shvaćen kao *lingua franca* programskih jezika.

6.1 Alati

U nastavku je ispis korištenih funkcija općenite namjene.

```
import math

from collections import defaultdict, Counter
from itertools import combinations, chain
from operator import itemgetter

import igraph
import numpy

# DATA IO

import csv, json

csv.field_size_limit(500000)


def loadCsvDicts(pth, enc='utf-8', delim=';', fnames_trans=None,
                 check_lens=False):
    rows = []
    with open(pth, encoding=enc) as fle:
        r = csv.reader(fle, delimiter=delim)
        header = next(r)
        if fnames_trans:
            header = [fnames_trans[name] for name in header]
        for row in r:
            if check_lens:
                assert len(row) == len(header), "h={},i={} : {}".format(
                    len(header), len(row), pth)
            rows.append(dict(zip(header, row)))
    return header, rows


def saveCsv(header, data, pth, enc='utf-8', delim=';'):
    with open(pth, "w", encoding=enc, newline="\n") as fle:
        w = csv.writer(fle, delimiter=delim)
        w.writerow(header)
        for row in data:
            assert len(row) == len(header)
            w.writerow(row)


def loadJson(pth, enc="utf-8"):
    with open(pth, encoding=enc) as fle:
        data = json.load(fle)
    return data


def saveJson(data, pth, enc="utf-8"):
    with open(pth, "w", encoding=enc) as fle:
        json.dump(data, fle, indent=2)
```

```
# DATA

n_dec = 3

def mergeUntilDisjunct(st):
    st = set(map(frozenset, st))
    unsorted = set(chain.from_iterable(st))
    _s = set()
    for a in st:
        remove_from_b = set()
        new_group = set(a)
        for b in _s:
            if new_group & b:
                new_group.update(b)
                remove_from_b.add(b)
        _s -= remove_from_b
        _s.add(frozenset(new_group))
    assert set(chain.from_iterable(_s)) == unsorted  # no elements are lost
    return _s


def makeEVSummary(dset, percent=True, _ev=None, ndec=1):
    ex = next(iter(dset.values()))
    header = list(ex.keys())
    summary = dict.fromkeys(header, 0)
    for item in dset.values():
        for key in header:
            if item[key] is not _ev:
                summary[key] += 1
    if percent:
        n = len(dset)
        for key in summary:
            v = round(summary[key] / n * 100, ndec)
            v = int(v) if v == int(v) else v
            summary[key] = v
    return summary


def makeMultiEVSummary(*datas, _na="n/a"):
    data_table = []
    summaries = list(map(makeEVSummary, datas))
    all_atts = set()
    for summ in summaries:
        all_atts.update(summ)
    for key in sorted(all_atts):
        row = [key]
        for summ in summaries:
            row.append(summ.get(key, _na))
        data_table.append(row)
    return data_table


def joinDictReports(dicts, keys, names, first_name='pokazatelj'):
    header = [first_name]
    header.extend(names)
    report = []
    for h in keys:
        row = [h]
        for d in dicts:
            row.append(d[h])
        report.append(row)
    return report, header

# TEXT

import regex
from unicodedata import normalize


def compWSpace(s):
    """
    compresses ALL whitespace
    (striping any trailing or leading whitespace, tab, newlines)

    # >>> compWSpace(" \t \n word  \n\r \tword\r    \u00A0 word \n\xa0 word \r\n")
```

```
    # 'word word word word'
    """
    return ' '.join(s.split())


def fromString(v, mv):
    v = compWSpace(v)
    if v in mv:
        return None
    else:
        return v


def splitValue(v, s):
    if v:
        return [vv.strip() for vv in v.split(s)]
    else:
        return None


def toInt(v):
    if v is not None:
        return int(v)


def resolveDashes(v):
    if v is not None:
        return v.replace("-", "-").replace("-", "-")


def resolveAuthors(v, spl_on):
    if v is not None:
        return [vv.strip() for vv in v.split(spl_on)]


def normToAscii(s):
    """
    Normalizes unicode string to ASCII

    >>> normToAscii('đščćž')
    'dsccz'
    """
    # 'đ' needs special handling :P
    s = s.replace("\u0111", u'd')
    s = s.replace("\u0110", u'D')
    s = normalize('NFKD', s).encode('ascii', 'ignore')
    return s.decode()


def dropNonWordChars(s, r='', leave='-'):
    """
    drops all characters from the string which are not letters, digits, whitepace, underscore
    and any characters present in leave

    >>> dropNonWordChars('đšč.ćž lj')
    'đščćž lj'
    """
    return regex.sub(r'[^\w\s%s]' % leave, r, s, flags=regex.V1 | regex.UNICODE)


def minInfoString(s):
    ps = compWSpace(dropNonWordChars(normToAscii(s), ' ', '').lower())
    return ps


# author name help

def cleanFirstAuthorName(first, init_tres=2):
    first = first.strip().replace("-", " ").replace('.', " ")
    first = " ".join(first.split())
    # handle A F J
    if len(first) > 1 and len(first.split(" ")) == len(first.replace(" ", "")):
        first = first.replace(" ", "").upper()
    elif len(first) <= init_tres:
        first = first.upper().strip(' .') + '.'
    else:
        first = first.title()
```

```
    return first


def cleanAut(s, first_in_tres=3):
    if not s:
        return s
    ncomma = s.count(',')
    if ncomma:
        last, first, *rest = s.split(',')
        first = cleanFirstAuthorName(first, first_in_tres)
        last = last.strip().title()
        if rest:
            rest = " [{}]".format(' - '.join(rest).lower().strip())
        else:
            rest = ""
        return last + ', ' + first + rest
    else:
        return s.strip().title()


def pickBestName(*cands, **overrides):
    min_info = list(map(minInfoString, cands))
    if len(set(min_info)) == 1:
        return cands[0]
    lens = defaultdict(set)
    for i, l in enumerate(map(len, min_info)):
        lens[l].add(i)
    best_by_len = cands[min(lens[max(lens)])]
    if best_by_len in overrides:
        best_by_len = overrides[best_by_len]
    return best_by_len

# GENERAL HELPFUL FUNCTIONS

bad_chars = set(r"""<>:"/\|?*""")


def doiToFileName(doi):
    fn = doi.replace(':', '++').replace('/', '--')
    assert not set(fn) & bad_chars
    return fn


def fileNameToDoi(fname):
    return fname.replace('++', ':').replace('--', '/')


# GROUPING

def _makeGroupFunc(att, func):
    if isinstance(att, (str, int)):
        if func is None:
            _func = lambda x: x[att]
        else:
            _func = lambda x: func(x[att])
    elif isinstance(att, (tuple, list)):
        if func is None:
            _func = itemgetter(*att)
        else:
            ga = itemgetter(*att)
            _func = lambda x: func(ga(x))
    else:
        raise Exception("Unsupported att type")
    return _func


def groupToDisjunct(data, att, func = None, items=None):
    _func = _makeGroupFunc(att, func)
    g = defaultdict(set)
    if not items:
        items=data
    for id in items:
        key = _func(data[id])
        g[key].add(id)
    return dict(g)
```

```
def groupToOverlaping(data, att, func = None, items=None):
    _func = _makeGroupFunc(att, func)
    g = defaultdict(set)
    if not items:
        items=data
    for id in items:
        keys = _func(data[id])
        if keys:
            for key in keys:
                g[key].add(id)
        else:
            g[None].add(id)
    return dict(g)


def groupFirstToDisjunct(data, att, func = None, items=None):
    _func = _makeGroupFunc(att, func)
    g = defaultdict(set)
    if not items:
        items=data
    for id in items:
        keys = _func(data[id])
        if keys:
            g[keys[0]].add(id)
        else:
            g[None].add(id)
    return dict(g)


def groupIntervals(groups, vfunc=None):
    if vfunc:
        def keyFunc(v, groups=groups):
            if v is None:
                return None
            for mn, mx in groups:
                if mn <= vfunc(v) <= mx:
                    return "{}-{}".format(mn, mx)
            return "out of group bounds"
    else:
        def keyFunc(v, groups=groups):
            if v is None:
                return None
            for mn, mx in groups:
                if mn <= v <= mx:
                    return "{}-{}".format(mn, mx)
            return "out of group bounds"

    return keyFunc

# AGGREGATION FUNCTIONS
from math import modf, floor


def quantile(x, q=0.5, qtype=7, issorted=False):
    """
    Args:
       x - input data
       q - quantile
       qtype - algorithm
       issorted- True if x already sorted.

    Compute quantiles from input array x given q.For median,
    specify q=0.5.

    References:
       http://reference.wolfram.com/mathematica/ref/Quantile.html
       http://wiki.r-project.org/rwiki/doku.php?id=rdoc:stats:quantile

    Author:
    Ernesto P.Adorio Ph.D.
    UP Extension Program in Pampanga, Clark Field.
    from:http://adorio-research.org/wordpress/?p=125
    """
    if not issorted:
        y = sorted(x)
    else:
        y = x
```

```
    if not (1 <= qtype <= 9):
        return None  # error!

    # Parameters for the Hyndman and Fan algorithm
    abcd = [(0, 0, 1, 0),  # inverse empirical distrib.function., R type 1
            (0.5, 0, 1, 0),  # similar to type 1, averaged, R type 2
            (0.5, 0, 0, 0),  # nearest order statistic,(SAS) R type 3

            (0, 0, 0, 1),  # California linear interpolation, R type 4
            (0.5, 0, 0, 1),  # hydrologists method, R type 5
            (0, 1, 0, 1),
            # mean-based estimate(Weibull method), (SPSS,Minitab), type 6
            (1, -1, 0, 1),  # mode-based method,(S, S-Plus), R type 7
            (1.0 / 3, 1.0 / 3, 0, 1),  # median-unbiased ,  R type 8
            (3 / 8.0, 0.25, 0, 1)  # normal-unbiased, R type 9.
    ]

    a, b, c, d = abcd[qtype - 1]
    n = len(y)
    g, j = modf(a + (n + b) * q - 1)
    if j < 0:
        return y[0]
    elif j >= n:
        return y[n - 1]

    j = int(floor(j))
    if g == 0:
        return y[j]
    else:
        return y[j] + (y[j + 1] - y[j]) * (c + d * g)


def iqRange(r):
    r = sorted(r)
    q1 = quantile(r, 0.25, issorted=True)
    q3 = quantile(r, 0.75, issorted=True)
    return q3 - q1


def quantileReport(r):
    r = sorted(r)
    q1 = quantile(r, 0.25, issorted=True)
    q2 = quantile(r, 0.5, issorted=True)
    q3 = quantile(r, 0.75, issorted=True)
    iqr = q3 - q1
    bottom_m_out = q1 - iqr * 1.5
    top_m_out = q3 + iqr * 1.5
    bottom_ex_out = q1 - iqr * 3
    top_ex_out = q3 + iqr * 3
    rep = {
        'min': min(r),
        'Q1': q1,
        'Q2': q2,
        'Q3': q3,
        'max': max(r),
        'IQR': iqr,
        'bottom mild whisker': bottom_m_out,
        'bottom extreme whisker': bottom_ex_out,
        'top mild whisker': top_m_out,
        'top extreme whisker': top_ex_out,
    }
    return rep


def gIndex(clist):
    """
    g-index is the (unique) largest number such that the top *g* articles received (together)
at least *g*^2 citations
    (Egghe, 2006)
    """
    g = 0
    run_cit = 0
    for cit in sorted(clist, reverse=True):
        run_cit += cit
        if g * g < run_cit:
            g += 1
        else:
```

```
            break
    return g


def hIndex(clist):
    """
    clist: sequence of citation count per article len(clist) = n articles; sum(clist) = n
citations
    "A scientist has an *h*-index of *h*, if *h* of her *Np* papers have at least *h*
citations each,
    and the other (*Np* - *h*) papers have at most *h* citations each." (Hirsch, 2005)
    """
    h = 0
    for cit in sorted(clist, reverse=True):
        if cit > h:
            h += 1
        else:
            break
    return h


# GRAPH RELATED



def makeCoocurrenceGraph(node_data):
    """
    graph_data is an iterable over (string id, set item_ids)

    item ids are ids of all items on which an item occures
    """
    g = igraph.Graph()
    edges = dict()
    for id, items in node_data:
        g.add_vertex(id, items=set(items))
    for a, b in combinations(g.vs, 2):
        edge_items = a["items"] & b["items"]
        if edge_items:
            edges[a["name"], b["name"]] = edge_items
    g.add_edges(list(edges.keys()))
    for edge in g.es:
        edge['items'] = edges[g.vs[edge.source]["name"], g.vs[edge.target]["name"]]
    return g


def getAutGraph(sci_data, dois=None, att_name="authors"):
    if not dois:
        dois = sci_data.keys()
    aut_data = defaultdict(set)
    for doi in dois:
        i = sci_data[doi]
        if i[att_name]:
            for a in i[att_name]:
                aut_data[a].add(doi)
    return makeCoocurrenceGraph(aut_data.items())

def getAllGraphItems(g, att_name="items"):
    return set(chain.from_iterable(n[att_name] for n in g.vs))


def getLargestComponent(g):
    c = g.components()
    comps = Counter(map(len, c))
    max_n_vertices = max(comps)
    assert comps[max_n_vertices] == 1, "There are many largest components"
    sg = None
    for i, l in enumerate(c):
        if len(l) == max_n_vertices:
            sg = c.subgraph(i)
            break
    return sg


def avgMinPathLen(g):
    avg_min_paths = []
    for path_lens in g.shortest_paths():
        path_lens = list(filter(lambda x: not math.isinf(x) and x, path_lens))
```

```
        if path_lens:
            avg_min_paths.append(numpy.mean(path_lens))
    return numpy.mean(avg_min_paths)

def reportAuthorGraph(g, att_name='items', n_random=0):
    non_isolates = []
    papers = set()
    for i, n in enumerate(g.vs):
        d = g.degree(i)
        papers.update(n[att_name])
        if d > 0:
            non_isolates.append(i)

    degrees = g.degree()
    comps = Counter(map(len, g.components()))

    lg = getLargestComponent(g)

    report = {
        "n radova": len(papers),
        "n čvorova": g.vcount(),
        "n veza": g.ecount(),
        "gustoća": round(g.density(), n_dec),
        "dijametar": g.diameter(),
        "asortativnost": round(g.assortativity_degree(False), n_dec),
        "n artikulacijskih čvorova": len(g.cut_vertices()),

        "globalni koeficijent grupiranja": round(
            g.transitivity_avglocal_undirected("zero"), 2),

        "prosječna najkraća duljina puta": round(avgMinPathLen(g), n_dec),

        "prosječan stupanj centralnosti": round(numpy.mean(degrees), n_dec),
        "StD stupnja centralnosti": round(numpy.std(degrees), n_dec),
        "median stupnja centralnosti": quantile(degrees),
        "max stupanj centralnosti": max(degrees),

        "n izoliranih čvorova": comps[1],
        "n izoliranih dijada": comps[2],
        "n izoliranih trijada": comps[3],
        "n ostalih komponenti (n>3)": sum(comps[k] for k in comps if k > 3),

        "n čvorova u najvećoj komponenti": lg.vcount(),
        "n radova u najvećoj komponenti": len(getAllGraphItems(lg)),
        "n veza u najvećoj komponenti": lg.ecount(),
    }

    if n_random:
        d = g.density()
        n = g.vcount()
        spaths = []
        skoef = []
        for i in range(n_random):
            rg = igraph.Graph.Erdos_Renyi(n, d)
            spaths.append(avgMinPathLen(rg))
            skoef.append(rg.transitivity_avglocal_undirected("zero"))
        report.update({
            "slučajna prosječna najkraća duljina puta (n=%s)" % n_random: numpy.mean(spaths),
            "slučajni globalni koeficijent grupiranja (n=%s)" % n_random: numpy.mean(skoef),
        })
    else:
        report.update({
            "slučajna prosječna najkraća duljina puta (n=%s)" % n_random: "n/a",
            "slučajni globalni koeficijent grupiranja (n=%s)" % n_random: "n/a",
        })
    return report
```

6.1.1 Grafikoni

```
from collections import Counter
from itertools import repeat
from math import sqrt
from operator import itemgetter
```

```
import numpy as np
import matplotlib.pyplot as plt

cs_van_gogh = ('#CBD893', '#E9C872', '#73AFB7', '#FBEF99', '#DB959A', '#9693D8')
cs_bw = ('0.6', '0.9', '0.3', '0.1') # '0.2', '0.1'('0.6', '0.3', '0.9')

golden_mean = (sqrt(5) - 1.0) / 2.0

params = {'backend': 'agg',

          'xtick.labelsize': 10,
          'ytick.labelsize': 10,
          'axes.labelsize': 10,
          'axes.titlesize': 12,
          'axes.textsize': 12,
          'text.fontsize': 12,
          'legend.fontsize': 8,
          'legend.labelspacing':0.3,
          'legend.handletextpad':0.4,
          'legend.fancybox':True,
          'legend.markerscale':0.8,
          'axes.grid': False,
          'font.family':'serif',
          'savefig.dpi': 300,
          'ps.usedistiller': 'xpdf',
          'ps.fonttype': 42,
          'pdf.fonttype': 42,

          }

to_inch = dict(
    cm = lambda x: float(x) / 2.54,
    mm = lambda x: float(x) / 25.4,
)

def getXSize(x):
    try:
        return int(x)
    except ValueError:
        x = x.strip()
        return to_inch[x[-2:]](x[:-2].strip())


def getFigSize(width, square = False):
    xs = getXSize(width)
    if square:
        fsize = (xs, xs)
    else:
        fsize = (xs, xs * golden_mean)
    return fsize


def makeFig(x_size, square = False):
    xs = getXSize(x_size)
    if square:
        fsize = (xs, xs)
    else:
        fsize = (xs, xs * golden_mean)
    fig, ax = plt.subplots(figsize = fsize)
    return fig, ax


def barsV(ax, data, labels = False, colors=cs_van_gogh,
          gwidth = 0.8, hatches = False):

    nbv = len(data)
    ng = len(data[0])
    xpos = np.arange(ng)
    w = gwidth / nbv
    cw = 0.1
    tstep = (w * nbv) / 2
    xticks = [cw + n + tstep for n in xpos]

    assert len(colors) >= len(data)
    if not labels:
        labels = repeat(None)
```

```
    for v, c, l in zip(data, colors, labels):
        cur_xpos = xpos + cw
        boxes = ax.bar(cur_xpos, v, w, label=l, color=c)
        if hatches:
            for box in boxes:
                box.set_hatch(hatches)
        cw += w
    ax.set_xticks(xticks)
    return ax


def lineDisc(ax, data, labels=None, colors=None,
             line_style =  ('-', '--', '-.', ':'), lw = 1):

    ms = 8.
    x = np.arange(len(data[0]))
    if not labels:
        labels = repeat(None)
    if not colors:
        colors = repeat('k')
    for y, l, c, ls in zip(data, labels, colors, line_style):
        ax.plot(x, y, label = l, c=c, ls=ls, ms=ms, lw=lw)
    return ax


def histDisc(ax, v, color = cs_van_gogh[0]):
    fullrange = False#kwargs['fullrange']??
    cntr = Counter(v)
    if fullrange:
        mnx, mxx = fullrange
        lft, top = zip(*sorted(((k, cntr[k]) for k in range(mnx, mxx)),
                               key = itemgetter(0)))
    else:
        lft, top = zip(*sorted(cntr.items(), key = itemgetter(0)))
    lft_ = [i - 0.5 for i in lft]

    boxes = ax.bar(lft_, top, 1.0, color=color)#w,
    return ax

def boxplot(ax, v):
    ax.boxplot(v, 0,'k+', 0)#w,
    return ax


def pie(ax, ratios, labels, l_size=8, l_dist=1.1, colors = cs_van_gogh):
    """
    draw a piechart
    """
    ax.set_aspect('equal')
    wedg, out_lbls = ax.pie(ratios, labels=labels, labeldistance=l_dist,
                            colors=colors, pctdistance=0.8)
    for t in out_lbls:
        t.set_size(l_size)
        t.set_weight('bold')
    return ax
```

6.1.2 Mreže

```
from collections import Counter
from itertools import repeat
from math import sqrt
from operator import itemgetter
import numpy as np
import matplotlib.pyplot as plt

cs_van_gogh = ('#CBD893', '#E9C872', '#73AFB7', '#FBEF99', '#DB959A', '#9693D8')
cs_bw = ('0.6', '0.9', '0.3', '0.1') # '0.2', '0.1'('0.6', '0.3', '0.9')

golden_mean = (sqrt(5) - 1.0) / 2.0

params = {'backend': 'agg',

          'xtick.labelsize': 10,
```

```
          'ytick.labelsize': 10,
          'axes.labelsize': 10,
          'axes.titlesize': 12,
          'axes.textsize': 12,
          'text.fontsize': 12,
          'legend.fontsize': 8,
          'legend.labelspacing':0.3,
          'legend.handletextpad':0.4,
          'legend.fancybox':True,
          'legend.markerscale':0.8,
          'axes.grid': False,
          'font.family':'serif',
          'savefig.dpi': 300,
          'ps.usedistiller': 'xpdf',
          'ps.fonttype': 42,
          'pdf.fonttype': 42,

          }

to_inch = dict(
    cm = lambda x: float(x) / 2.54,
    mm = lambda x: float(x) / 25.4,
)

def getXSize(x):
    try:
        return int(x)
    except ValueError:
        x = x.strip()
        return to_inch[x[-2:]](x[:-2].strip())


def getFigSize(width, square = False):
    xs = getXSize(width)
    if square:
        fsize = (xs, xs)
    else:
        fsize = (xs, xs * golden_mean)
    return fsize


def makeFig(x_size, square = False):
    xs = getXSize(x_size)
    if square:
        fsize = (xs, xs)
    else:
        fsize = (xs, xs * golden_mean)
    fig, ax = plt.subplots(figsize = fsize)
    return fig, ax


def barsV(ax, data, labels = False, colors=cs_van_gogh,
          gwidth = 0.8, hatches = False):

    nbv = len(data)
    ng = len(data[0])
    xpos = np.arange(ng)
    w = gwidth / nbv
    cw = 0.1
    tstep = (w * nbv) / 2
    xticks = [cw + n + tstep for n in xpos]

    assert len(colors) >= len(data)
    if not labels:
        labels = repeat(None)

    for v, c, l in zip(data, colors, labels):
        cur_xpos = xpos + cw
        boxes = ax.bar(cur_xpos, v, w, label=l, color=c)
        if hatches:
            for box in boxes:
                box.set_hatch(hatches)
        cw += w
    ax.set_xticks(xticks)
    return ax
```

```
def lineDisc(ax, data, labels=None, colors=None,
             line_style =  ('-', '--', '-.', ':'), lw = 1):

    ms = 8.
    x = np.arange(len(data[0]))
    if not labels:
        labels = repeat(None)
    if not colors:
        colors = repeat('k')
    for y, l, c, ls in zip(data, labels, colors, line_style):
        ax.plot(x, y, label = l, c=c, ls=ls, ms=ms, lw=lw)
    return ax


def histDisc(ax, v, color = cs_van_gogh[0]):
    fullrange = False#kwargs['fullrange']??
    cntr = Counter(v)
    if fullrange:
        mnx, mxx = fullrange
        lft, top = zip(*sorted(((k, cntr[k]) for k in range(mnx, mxx)),
                               key = itemgetter(0)))
    else:
        lft, top = zip(*sorted(cntr.items(), key = itemgetter(0)))
    lft_ = [i - 0.5 for i in lft]

    boxes = ax.bar(lft_, top, 1.0, color=color)#w,
    return ax

def boxplot(ax, v):
    ax.boxplot(v, 0,'k+', 0)#w,
    return ax


def pie(ax, ratios, labels, l_size=8, l_dist=1.1, colors = cs_van_gogh):
    """
    draw a piechart
    """
    ax.set_aspect('equal')
    wedg, out_lbls = ax.pie(ratios, labels=labels, labeldistance=l_dist,
                            colors=colors, pctdistance=0.8)
    for t in out_lbls:
        t.set_size(l_size)
        t.set_weight('bold')
    return ax
```

6.2 Skripte

U ovom modulu nalazi se sav kôd koji je specifično radio sa opisanim skupovima podataka. Podijeljen je u dva osnovna dijela, pripremu i analizu. Priprema odrađuje sve korake od ulaznih skupova do pripremljenih skupova u čemu prolazi kroz više faza. Ovaj kôd ovisi o modulu phd_alati.

6.2.1 Glavna skripta

Sljedeća skripta je kontrolna za sve ostale procese. Ulazni podaci za ovu skriptu su svi ulazni bibliografski metapodaci, a izlazi svi pripremljeni skupovi podataka i sve tablice i sve podatkovne slike korištene u tekstu rada.

```
"""
Main script doing all preparation and analysis as described in Phd Thesis:
```

```
*The development of scientometrics as represented by the journal Scientometrics since the
begining of its publication in 1978 to 2010*
"""

from phd_tools import tools

from phd_scripts.static import data_paths
from phd_scripts.prep import base_prepare_input, sci_kp_rep, sci_correct
from phd_scripts.prep import sci_merge, cit_connect_wos, sci_make_final
from phd_scripts.an.an_main import addAnalysisDeliverables, printInventory, addStaticImages


# create structures to hold "deliverables"
tables = {}
figures = {}
reports = {}
deliverables = {
    "tables":tables,
    "figures":figures,
    "reports":reports
}

# 1. BASE PREPARE SETS
# 1.1 prepare inputs
sci_sets, cit_sets = base_prepare_input.prepareMainSets(deliverables)

print("\nPreparing Scientometrics dataset.\n")
# 2. PREPARE SCI
# 2.1 CORRECT SCI SETS
# 2.1.1 report on pre-corrected
# Pregled internih provjera za svaki skup i u odnosu na metapodatke od izdavača

t, h = sci_kp_rep.sciInconstTable(sci_sets)
tables['inconsistency_report'] = {
    "header":h,
    "data":t,
    "title":"Nepravilnosti u ulaznim skupovima podataka",
}
# 2.1.2 correct
sci_correct.correctSci(sci_sets)
# # 2.1.3 report on post-corrected
# corrected_report = sci_kp_rep.sciInconstTable(sci_sets)

# 3. MERGE AND DISAMBIGUATE SCIENTOMETRIC DATA
sci_merged = sci_merge.mergeSci(sci_sets)

print("- successful\n\n----\n")

# 4. CONNECT WoS CITATIONS
cit_wos = cit_connect_wos.prepareWosCit(sci_merged, cit_sets)

# 5. MAKE FINAL SCI DATASET
sci_final = sci_make_final.makeSciFinal(sci_merged, cit_wos)

# 6. MAKE DELIVERABLES

print("\nMaking deliverables.\n")
addAnalysisDeliverables(deliverables)
addStaticImages(deliverables)
tools.saveJson(deliverables, data_paths.p_deliv)
print("- successful, inventory:\n")
printInventory()
print("\n----\n")
```

6.2.2 Osnovna priprema ulaznih podataka

```
import re

from phd_tools import tools, parse_ris

from phd_scripts.static import data_paths
from phd_scripts.static.raw_data_info import *
```

```
def _preparePubSci(data, header):
    pub_dois = set()
    preproc_keys = set(header) - set(["authors"])

    for item in data:
        for k in preproc_keys:
            item[k] = tools.fromString(item[k], missing_values)
        # HANDLE AUTHORS
        v = item["authors"]
        if v:
            item["authors"] = [tools.fromString(a, missing_values) for a in v]
            assert all(item["authors"])
        else:
            item["authors"] = None
        item["year_pub"] = tools.toInt(item["year_pub"])
        item["volume"] = tools.toInt(item["volume"])
        item["issue"] = tools.resolveDashes(item["issue"])
        # CHECK DOIs PRESENT AND UNIQUE
        doi = item["doi"]
        assert doi, "missing doi in publisher data"
        assert doi not in pub_dois, "duplicate doi in publisher data"
        pub_dois.add(doi)
        assert item["publication_title"].lower() == sci_name


def _prepItemWosScop(item, keys, aut_split=';', rest_split=';'):
    #print(type(item["authors"]), keys)
    for k in keys:
        item[k] = tools.fromString(item[k], missing_values)
    # split needed attributes containing many values
    item["authors"] = tools.resolveAuthors(item["authors"], aut_split)
    item["references"] = tools.splitValue(item["references"], rest_split)
    # convert strings to integers
    item["citation_count"] = tools.toInt(item["citation_count"])
    item["year_pub"] = tools.toInt(item["year_pub"])
    # convert other values
    item["issue"] = tools.resolveDashes(item["issue"])

drop_ad_auts = re.compile(r"\[.*?\]")

def prepareWos(data, header):
    wos_ids = set()
    for item in data:
        _prepItemWosScop(item, header)
        wos_id = item["wos_id"]
        if item["inst_adresses"]:
            item["inst_adresses"]            =            tools.splitValue(drop_ad_auts.sub("",
item["inst_adresses"]), ';')
        assert wos_id, "missing wos id in wos citing data"
        assert wos_id not in wos_ids, "duplicate wos_id in wos citing data"
        wos_ids.add(wos_id)


def _prepareWosSci(data, header):
    prepareWos(data, header)
    for item in data:
        item["volume"] = tools.toInt(item["volume"])
        it = item["item_type"]
        item["item_type"] = wos_type_fix.get(it, it)
        assert item["publication_title"].lower() == sci_name


# extracts ids from scopus links
_re_sco_link_id = re.compile(r'eid=(.+?)&')
def _extractScopId(scop_link, re_sco_link_id = _re_sco_link_id):
    scop_id = re_sco_link_id.findall(scop_link)
    assert len(scop_id) == 1 and scop_id[0].startswith("2-s2.0-"), "bad scop id"
    return scop_id[0][7:]

_last_space = re.compile(" (?!.* )")
# >>> last_space.sub(', ', "Van Raan AFJ")
# 'Van Raan, AFJ'

def prepareScopus(data, header):
    sco_cit_ids = set()
    for item in data:
```

```
        # PREPARE
        if not item["citation_count"]:
            item["citation_count"] = "0"
        _prepItemWosScop(item, header, aut_split = ',')
        # scopus splits author names by space
        if item["authors"]:
            item["authors"] = [_last_space.sub(", ", aut) for aut in item["authors"]]
        if item["inst_addresses_with_aut"]:
            _insts = tools.splitValue(item["inst_addresses_with_aut"], ';')
            insts = []
            for i in _insts:
                i = tools.splitValue(i, ',')
                aut = ", ".join(i[:2])
                i = i[2:]
                insts.append((aut, i))

            item["inst_addresses_with_aut"] = insts
        # extract scopus id from scopus url
        sco_id = item["sco_id"] = _extractScopId(item["scopus_url"])
        # VALIDATE
        assert sco_id not in sco_cit_ids, "duplicate scop_id in scopus data"
        sco_cit_ids.add(sco_id)


# prepare SCOPUS SCIENTOMETRICS metadata
def _prepareScopusSci(data, header):
    prepareScopus(data, header)
    for item in data:
        item["volume"] = tools.toInt(item["volume"])
        assert item["publication_title"].lower() == sci_name


def getInputData():
    #scientometrics raw data
    sci_pub_head, sci_pub = parse_ris.parseRis(*data_paths.u_sci_izd)
    sci_wos_head, sci_wos = tools.loadCsvDicts(*data_paths.u_sci_wos, fnames_trans=fnames_wos)
    sci_sco_head,          sci_sco          =          tools.loadCsvDicts(*data_paths.u_sci_sco,
fnames_trans=fnames_scopus)
    #scientometrics citing data
    cit_wos_head, cit_wos = tools.loadCsvDicts(*data_paths.u_cit_wos, fnames_trans=fnames_wos)
    cit_sco_head,          cit_sco          =          tools.loadCsvDicts(*data_paths.u_cit_sco,
fnames_trans=fnames_scopus)
    assert sci_wos_head == cit_wos_head
    assert sci_sco_head == cit_sco_head

    #prepare all sets
    _preparePubSci(sci_pub, sci_pub_head)
    _prepareWosSci(sci_wos, sci_wos_head)
    _prepareScopusSci(sci_sco, sci_sco_head)
    prepareWos(cit_wos, cit_wos_head)
    prepareScopus(cit_sco, cit_sco_head)

    #make output
    #all used ids are checked unique by the above
    sci_pub = {i["doi"]:i for i in sci_pub}
    sci_wos = {i["wos_id"]:i for i in sci_wos}
    sci_sco = {i["sco_id"]:i for i in sci_sco}
    cit_wos = {i["wos_id"]:i for i in cit_wos}
    cit_sco = {i["sco_id"]:i for i in cit_sco}

    sci = {
        "pub":{"data":sci_pub, "header":sci_pub_head, "key_name":"doi"},
        "wos":{"data":sci_wos, "header":sci_wos_head, "key_name":"wos_id"},
        "sco":{"data":sci_sco, "header":sci_sco_head, "key_name":"sco_id"}
    }
    cit = {
        "wos":{"data":cit_wos, "header":cit_wos_head, "key_name":"wos_id"},
        "sco":{"data":cit_sco, "header":cit_sco_head, "key_name":"sco_id"},
    }
    return sci, cit


def dropAtts(dsets, keep_atts):
    ska = set(keep_atts)
    for dset_name in dsets:
        header = [n for n in dsets[dset_name]['header'] if n in keep_atts]
        for item in dsets[dset_name]['data'].values():
```

```
            for k in item.keys() - ska:
                del item[k]
        dsets[dset_name]['header'] = header

def prepedReport(sci_sets, cit_sets):
    sci_pub = sci_sets["pub"]["data"]
    sci_wos = sci_sets["wos"]["data"]
    sci_sco = sci_sets["sco"]["data"]

    cit_wos = cit_sets["wos"]["data"]
    cit_sco = cit_sets["sco"]["data"]

    ev_header = ["atribut", "sci_pub", "sci_wos", "sci_sco", "cit_wos", "cit_sco"]
    ev_summary = tools.makeMultiEVSummary(sci_pub, sci_wos, sci_sco, cit_wos, cit_sco)

    return {
        "n_sci_pub":len(sci_pub),
        "n_sci_wos":len(sci_wos),
        "n_sci_sco":len(sci_sco),
        "n_cit_wos":len(cit_wos),
        "n_cit_sco":len(cit_sco),

        "ev_table":(ev_header, ev_summary)
    }

def prepareMainSets(deliverables):
    print("\nPreparing all input datasets from text.\n")

    print("  Loading and preparing input data from text\n  (cleaning, missing values, types,
spliting and ids)")
    sci, cit = getInputData()
    print("  - successful\n")

    print("  Removing unneeded atts")
    dropAtts(sci, keep_atts_sci)
    dropAtts(cit, keep_atts_cit)
    print("  - successful\n")

    print("  Saving all datasets.")
    tools.saveJson(sci, data_paths.kp_sci_prepared)
    tools.saveJson(cit, data_paths.kp_cit_prepared)
    print("  - successful\n")

    print("- successful\n\n----\n")

    preped_report = prepedReport(sci, cit)
    ev_head, ev_d = preped_report["ev_table"]
    deliverables["reports"]["basic_prep_report"] = preped_report
    deliverables["tables"]["ev_summary"] = {
        "header":ev_head,
        "data":ev_d,
        "title":"Prazne vrijednosti preuzetih atributa u ulaznim skupovima podataka",
    }

    return sci, cit
```

6.2.3 Ispravljanje nekonzistentnosti u skupovima podataka o radovima u *Scientometrics*

```
from phd_tools import tools

from phd_scripts.static import data_paths
from phd_scripts.static import correct_p_sci_sets as cps


def correctSci(dsets):
    """
    Applies premade corrections.
    """
    print("  Applying premade corrections to sci_pub, sci_wos and sci_sco")
    cps.implementCorrections(cps.drop_p_sci_izd,
```

```
                            cps.correct_p_sci_izd, dsets["pub"]["data"])
    cps.implementCorrections(cps.drop_p_sci_wos,
                            cps.correct_p_sci_wos, dsets["wos"]["data"])
    cps.implementCorrections(cps.drop_p_sci_sco,
                            cps.correct_p_sci_sco, dsets["sco"]["data"])
    # APPLY CORRECTIONS TO PUBLISHER ISSUES PER VOLUME
    for item in dsets["pub"]["data"].values():
         vol, iss = item["volume"], item["issue"]
         item["issue"] = cps.correct_p_sci_izd_vol_issue.get( (vol, iss), iss)
    tools.saveJson(dsets, data_paths.kp_sci_corrected)
    print("  - successful\n")
```

6.2.4 Spajanje i razriješavanje vrijednosti podataka o radovima u časopisu *Scientometrics*

```
from os import path

from phd_tools import tools

from phd_scripts.prep import sci_authors, sci_insts
from phd_scripts.static import data_paths

_match_on_key = "{volume}:{issue}:{page_begin}"

pub_first = [
    'doi', 'year_pub', 'volume', 'issue',
    'page_begin', 'page_end', 'item_title',
    'abstract', 'publication_title'
]
take_from_wos = ['wos_id', 'inst_adresses']
take_from_sco = ['sco_id', 'inst_addresses_with_aut']
dis_atts = ["authors", "citation_count", "references", 'item_type']

wos_d = ["{}_{}".format(att, "wos") for att in dis_atts]
sco_d = ["{}_{}".format(att, "sco") for att in dis_atts]

joint_header = pub_first + take_from_wos + take_from_sco + wos_d + sco_d + ["authors_pub",
"pdf_path"]
assert len(joint_header) == len(set(joint_header)), "Joint header names are not unique"

def _connectOnDoiOrIssue(data, name, doi_map, issue_to_doi):
    undetected = set()
    for id, item in data.items():
        doi = item["doi"]
        if not doi:
            key = _match_on_key.format(**item)
            if key in issue_to_doi:
                doi = issue_to_doi[key]
        if doi:
            doi_map[doi][name].add(id)
        else:
            undetected.add(id)
    return undetected


def mapOnDoi(dsets):
    pub_data = dsets["pub"]["data"]
    sco_data = dsets["sco"]["data"]
    wos_data = dsets["wos"]["data"]

    pub_dois = set()
    doi_map = {}
    pub_iss_to_doi = {}

    #no duplicates on this must be present, done in analysing prepared & corrected data
    for doi, main_item in pub_data.items():
        pub_dois.add(doi)
        doi_map[doi] = dict(wos=set(), sco=set())
        key = _match_on_key.format(**main_item)
        pub_iss_to_doi[key] = doi
```

```
    undetect_wos = _connectOnDoiOrIssue(wos_data, "wos", doi_map, pub_iss_to_doi)
    undetect_scopus = _connectOnDoiOrIssue(sco_data, "sco", doi_map, pub_iss_to_doi)
    return doi_map, undetect_wos, undetect_scopus


def resolveDoiMap(doi_map):
    # make a single doi refer to a single wos and single sco item
    # keep track of many wos or sco arts mapped to a single doi
    wos_dup = set()
    sco_dup = set()
    for doi, ids in doi_map.items():
        wos, sco = ids["wos"], ids["sco"]
        if len(wos) > 1:
            wos_dup.add(frozenset(wos))
        elif len(wos) == 1:
            ids["wos"] = wos.pop()
        else:
            ids["wos"] = None

        if len(sco) > 1:
            sco_dup.add(frozenset(sco))
        elif len(sco) == 1:
            ids["sco"] = sco.pop()
        else:
            ids["sco"] = None
    return wos_dup, sco_dup


def makeDoiMap(sci_sets):
    print("  Mapping items on DOI or vol-issue-bp.")
    doi_map, un_wos, un_sco = mapOnDoi(sci_sets)
    dup_wos, dup_sco = resolveDoiMap(doi_map)
    print("  - successful\n")
    return doi_map, un_wos, un_sco, dup_wos, dup_sco


def _getMainItemUpdate(main_item, sub_item, atts):
    new_vals = {}
    for att in atts:
        sub_value = sub_item.get(att, None)
        main_value = main_item[att]
        if sub_value and not main_value:
            new_vals[att] = sub_value
    return new_vals


def _getOtherItemData(suff, item, take_atts, dis_atts):
    new_vals = {}
    for att in take_atts:
        new_vals[att] = item[att]
    for att in dis_atts:
        new_vals["{}_{}".format(att, suff)] = item[att]
    return new_vals


def _updateMainItem(main_item, name, id, data, pub_first, take_over, dis_atts):
    if id:
        sub_item = data[id]
        fill_in = _getMainItemUpdate(main_item, sub_item, pub_first)
        other_data = _getOtherItemData(name, sub_item, take_over, dis_atts)
        main_item.update(fill_in)
        main_item.update(other_data)


def makeMergedData(doi_map, dsets):
    sci_data = {}
    pub_data = dsets["pub"]["data"]
    sco_data = dsets["sco"]["data"]
    wos_data = dsets["wos"]["data"]

    #  pdf_doi  =  set(doitools.fileNameToDoi('.'.join(f.split('.')[:-1]))  for  d,  s,  f  in
os.walk(data_paths.full_texts))
    for doi in doi_map:
        # create item base
        pub_item = pub_data[doi]
        main_item = { k:pub_item.get(k, None) for k in joint_header }
        assert main_item["doi"] == doi
```

```
        main_item["authors_pub"] = pub_item['authors']

        pdf_path = path.join(data_paths.full_texts, str(main_item["year_pub"]),
                             tools.doiToFileName(doi) + '.pdf')
        if path.isfile(pdf_path):
            main_item["pdf_path"] = pdf_path
        else:
            main_item["pdf_path"] = None
        # update item with data from wos and scopus
        wos_id = doi_map[doi]["wos"]
        sco_id = doi_map[doi]["sco"]
        _updateMainItem(main_item,   "wos",   wos_id,   wos_data,   pub_first,   take_from_wos,
dis_atts)
        _updateMainItem(main_item,   "sco",   sco_id,   sco_data,   pub_first,   take_from_sco,
dis_atts)
        assert  set(main_item.keys())  ==  set(joint_header),  "Joint  items  to  joint  header
mismatch."
        sci_data[doi] = main_item
    return sci_data


def prepareMerged(sci_merged):
    hand_set = tools.loadJson(data_paths.h_hand)
    # disambiguate authors
    aut_reg, aut_name_counts = sci_authors.disambiguateAuthors(sci_merged)
    to_proper_aut = {}
    for d in aut_reg.values():
        for v in d["variants"]:
            assert v not in to_proper_aut, "Same variant for different author"
            to_proper_aut[v] = d["name"]
    # extract countries
    count = sci_insts.getDoiCountries(sci_merged)
    # update merged items
    for doi, item in sci_merged.items():
        item["type"] = hand_set[doi]["type"]
        pf = hand_set[doi]["prim_focus"]
        # take all different prim_focuses keeping original order,
        # just in case multiple applied during hand classification
        if pf:
            _pf = []
            for p in pf:
                if p not in _pf:
                    _pf.append(p)
            item["prim_focus"] = pf
        else:
            item["prim_focus"] = None
        item["countries"] = count[doi]
        auts = filter(None, (item["authors_pub"], item["authors_wos"], item["authors_sco"]))
        auts = set(tuple(to_proper_aut[a] for a in at) for at in auts)
        if auts:
            # aftedr disambiguation, all three author sets must produce same value for author
field
            assert len(auts) == 1, doi + ' - ' + str(auts)
            item["authors"] = auts.pop()
        else:
            item["authors"] = None


def mergeSci(sci_sets):
    print("  Merging scientometric data on mapped ids.")
    doi_map, un_wos, un_sco, dup_wos, dup_sco = makeDoiMap(sci_sets)
    # raise error if any WoS or Sco unconnected or connected to more than one
    assert not un_wos and not un_sco and not dup_wos and not dup_sco
    # raise error if not all items have connection data
    assert len(doi_map) == len(sci_sets['pub']['data'])
    sci_merged = makeMergedData(doi_map, sci_sets)
    print("  - successful\n")
    print("  Preparing merged data.\n")
    prepareMerged(sci_merged)
    print("  - successful\n")
    print("  Saving merged data.")
    tools.saveJson(sci_merged, data_paths.kp_sci_merged)
    print("  - successful\n")
    return sci_merged
```

6.2.4.1 Razriješavanje podataka o autorima

```
import difflib
from collections import defaultdict, Counter
from itertools import combinations, chain
from operator import itemgetter

from phd_scripts.static import data_paths
from phd_scripts.static.res_sci_auts import *

from phd_tools import tools

data_names = ["pub", "wos", "sco"]
aut_att_names = ["authors_pub", "authors_wos", "authors_sco"]
geta = itemgetter(*aut_att_names)


def disambiguateAuthors(sci_merged):
    print("    Disambiguating author names.")
    best_names, clean_aut_sets, merged_namesets, clean_to_dirty = disambiguateSciAuthors(
        sci_merged)
    aut_reg = getAutData(sci_merged, best_names, clean_to_dirty)
    aut_name_counts = autReport(sci_merged)
    all_vars = set()
    for a in aut_reg.values(): all_vars.update(a['variants'])
    assert len(all_vars) == aut_name_counts[
        'tot_raw'], "more variants in data than reckognized for disambiguated authors"
    assert len(aut_reg) == len(set(map(tools.normToAscii,
                                       aut_reg))), "normalized names of authors not unique"
    aut_name_counts["n_merged_namesets"] = len(merged_namesets)
    aut_name_counts["n_aut_names"] = len(aut_reg)
    tools.saveJson(aut_reg, data_paths.h_sci_authors)
    print("    - successful\n")
    return aut_reg, aut_name_counts


def disambiguateSciAuthors(sci_merged):
    authorships = list(map(geta, sci_merged.values()))

    aut_sets = set()
    for auts in authorships:
        merged = set(frozenset(a) for a in zip(*filter(bool, auts)))
        aut_sets.update(merged)

    clean_aut_sets = set()
    clean_to_dirty = defaultdict(set)
    for aset in aut_sets:
        clean_set = set()
        for a in aset:
            clean_a = tools.cleanAut(a)
            clean_set.add(clean_a)
            clean_to_dirty[clean_a].add(a)
        clean_aut_sets.add(frozenset(clean_set))

    merged_namesets = tools.mergeUntilDisjunct(clean_aut_sets)

    best_names = {}
    for ns in merged_namesets:
        best_names[
            tools.pickBestName(*sorted(ns, reverse=True), **rename_best)] = ns

    for pref, non_pref in merge_best:
        best_names[pref] = best_names[pref] | best_names[non_pref]
        del best_names[non_pref]

    return best_names, clean_aut_sets, merged_namesets, clean_to_dirty


def getSuspiciousSets(merged_namesets):
    suspicious_sets = set()
    for ns in merged_namesets:
        for a, b in combinations(ns, 2):
            s = difflib.SequenceMatcher(None, a, b)
            if s.ratio() < 0.7:
                suspicious_sets.add(ns)
```

```
    return suspicious_sets


def getAutData(sci_merged, best_names, clean_to_dirty):
    changes = dict()
    for n, vars in best_names.items():
        for nn in vars:
            assert nn not in changes
            changes[nn] = n
    aut_reg = defaultdict(dict)
    for doi, item in sci_merged.items():
        if any(geta(item)):
            for variants in zip(*filter(bool, geta(item))):
                proper = set(changes[tools.cleanAut(v)] for v in variants)
                assert len(proper) == 1
                a = proper.pop()
                aut_reg[a].setdefault("items", set()).add(doi)
    aut_reg = dict(aut_reg)
    for aut in aut_reg:
        vars = set()
        # revert to names as they are in the data
        for var in best_names[aut]:
            vars.update(clean_to_dirty[var])
        vars.update(clean_to_dirty[aut])
        adict = aut_reg[aut]
        adict["name"] = aut
        adict["items"] = list(adict["items"])
        adict["variants"] = list(vars)
    return aut_reg


def autReport(sci_merged):
    wrap_none = lambda x: [x] if x is None else x
    wrap_none_clean = lambda x: [x] if x is None else map(tools.cleanAut, x)

    counts = dict()
    counts_clean = dict()
    count_none = dict()
    all_auts = set()
    all_auts_clean = set()
    for data_name in ["pub", "wos", "sco"]:
        aut_values = list(
            map(itemgetter("authors_" + data_name), sci_merged.values()))

        counts[data_name] = Counter(
            chain.from_iterable(map(wrap_none, aut_values)))
        counts_clean[data_name] = Counter(
            chain.from_iterable(map(wrap_none_clean, aut_values)))

        count_none[data_name] = counts[data_name].pop(None)
        count_none_clean = counts_clean[data_name].pop(None)
        assert count_none_clean == count_none[data_name]

        all_auts.update(counts[data_name])
        all_auts_clean.update(counts_clean[data_name])

    return {
        'tot_raw': len(all_auts),
        'tot_clean': len(all_auts_clean),
        'pub_raw': len(counts["pub"]),
        'pub_clean': len(counts_clean["pub"]),
        'wos_raw': len(counts["wos"]),
        'wos_clean': len(counts_clean["wos"]),
        'sco_raw': len(counts["sco"]),
        'sco_clean': len(counts_clean["sco"]),
    }
```

6.2.4.2 Ekstrakcija zemalja iz adresa

```
from phd_tools import tools

from phd_scripts.static.res_sci_insts import *
from phd_scripts.static import data_paths
```

```
def repInstsWithAuthors(sci_merged):
    aut_no_inst = set()
    wrong_n_inst = set()
    empty_inst = set()

    for doi, item in sci_merged.items():
        aut_sco = item["authors_sco"]
        if not aut_sco:
            continue
        insts = item['inst_addresses_with_aut']
        if not insts:
            aut_no_inst.add(doi)
        if len(insts) != len(aut_sco):
            wrong_n_inst.add(doi)
        for i in insts:
            if not i[1]:
                empty_inst.add(doi)
                break
    print("n aut no insts: %s" % len(aut_no_inst))
    print("n wrong n insts: %s" % len(wrong_n_inst))
    print("n with empty insts: %s" % len(empty_inst))
    return aut_no_inst | wrong_n_inst | empty_inst


def getAddressCountries(sci_merged, doiset):
    instset = set()
    for doi in doiset:
        insts = sci_merged[doi]['inst_addresses_with_aut']
        if insts:
            for i in insts:
                instset.add(tuple(i[1]))

    header, _cnt = tools.loadCsvDicts(data_paths.a_countries)
    countries = {}
    for c in _cnt:
        for var in [s.strip() for s in c["variants"].split(';')]:
            assert var not in countries
            countries[var.lower()] = c["full"]

    inst_country = {}
    no_countries = set()
    many_countries = set()
    for inst in instset:
        if inst in hand_extracted:
            inst_country[inst] = hand_extracted[inst]
            continue

        this_countries = set()
        for part in inst:
            if part.lower() in countries:
                this_countries.add(countries[part.lower()])

        if len(this_countries) == 0:
            no_countries.add(inst)
        elif len(this_countries) > 1:
            many_countries.add(inst)
        else:
            inst_country[inst] = this_countries.pop()

    return inst_country, no_countries, many_countries


def getAddressWosCountries(sci_merged, doiset):
    instset = set()
    for doi in doiset:
        insts = sci_merged[doi]['inst_adresses']
        if insts:
            instset.update(insts)

    header, _cnt = tools.loadCsvDicts(data_paths.a_countries)
    countries = {}
    for c in _cnt:
        for var in [s.strip() for s in c["variants"].split(';')]:
            assert var not in countries
            countries[var.lower()] = c["full"]
```

```
    inst_country = {}
    no_countries = set()
    many_countries = set()
    for inst in instset:
        if inst in hand_wos_extracted:
            inst_country[inst] = hand_wos_extracted[inst]
            continue

        this_countries = set()
        _inst = tools.dropNonWordChars(inst.lower())

        for c in countries:
            if c in _inst:
                this_countries.add(countries[c])
        if len(this_countries) == 0:
            no_countries.add(inst)
        elif len(this_countries) > 1:
            many_countries.add(inst)
        else:
            inst_country[inst] = this_countries.pop()

    return inst_country, no_countries, many_countries


def getDoiCountries(sci_merged):
    # bad_dois = repInstsWithAuthors(sci_merged)
    ad_to_co, no_co, many_co = getAddressWosCountries(sci_merged,
                                                      sci_merged.keys())
    assert not no_co and not many_co
    data = {}
    for doi, item in sci_merged.items():
        if item["inst_adresses"]:
            data[doi] = list(
                set(ad_to_co[inst] for inst in item["inst_adresses"]))
        else:
            data[doi] = None
    return data
```

6.2.5 Povezivanje WoS citirajućih radova s citiranim radovima u *Scientometrics*

```
from collections import defaultdict
from operator import itemgetter

from phd_tools import tools

from phd_scripts.static import data_paths
from phd_scripts.static.res_sci_refs import *

jou_correct = {
    'JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE':'JOURNAL OF THE AMERICAN SOCIETY
FOR INFORMATION SCIENCE AND TECHNOLOGY',
    'ZHURNAL NEVROPATOLOGII I PSIKHIATRII IMENI S S KORSAKOVA':'ZHURNAL NEVROLOGII I
PSIKHIATRII IMENI S S KORSAKOVA',
    'INVESTIGACION BIBLIOTECOLOGIA':'INVESTIGACION BIBLIOTECOLOGICA',
}

known_bad_citing = {
    # no references
    'WOS:000285219000008',
    # no references to sci
    'WOS:000243213200005', 'WOS:000285128100001', 'WOS:000281810500001'
}

def dropBadWoSItems(wos_data):
    for ref in known_bad_citing:
        del wos_data[ref]

def correctCitJournals(wos_data):
    for item in wos_data.values():
        pt = item['publication_title']
```

```
        pt = jou_correct.get(pt, pt)
        if pt.startswith('ASIST'):
            pt = 'ASIST ANNUAL MEETING - PROCEEDINGS'
        if pt.startswith('ARTIFICIAL NEURAL NETWORKS:'):
            pt = 'ARTIFICIAL NEURAL NETWORKS - PROCEEDINGS'
        item['publication_title'] = pt.title()

def disambiguateCitSciAuthors(wos_data):
    auts = tools.loadJson(data_paths.h_sci_authors)
    to_proper_aut = {}
    for d in auts.values():
        for v in d["variants"]:
            assert v not in to_proper_aut, "Same variant for different author"
            to_proper_aut[v] = d["name"]
    for item in wos_data.values():
        if item["authors"]:
            item["authors"] = [to_proper_aut.get(a, a) for a in item["authors"]]



def getJouSelfCit(sci_data, wos_header):
    data = {}
    for item in sci_data.values():
        if item["references_wos"]:
            if sum(1 for _ in filter(lambda ref: 'SCIENTOMETRICS' in ref and ref not in
known_bad_refs, item["references_wos"])):
                wos_item = {}
                for att in wos_header:
                    wos_item[att] = item.get(att)
                wos_item["references"] = item["references_wos"]
                data[item["wos_id"]] = wos_item
    return data

def getSciRefs(wos_data, att_name = 'references'):
    sci_refs = set()
    no_refs = set()
    no_sciento_refs = set()
    for wos_id, item in wos_data.items():
        refs = item[att_name]
        if not refs:
            no_refs.add(wos_id)
            continue
        n_refs_to_sciento = 0
        for ref in refs:
            if 'SCIENTOMETRICS' in ref:
                sci_refs.add(ref)
                n_refs_to_sciento += 1
        if not n_refs_to_sciento:
            no_sciento_refs.add(wos_id)
    return sci_refs, no_refs, no_sciento_refs

def dropBadRefs(refset, known_bad_refs = known_bad_refs, hand_connected = hand_connected):
    bad_refs = set()
    after2010 = set()
    for ref in refset:
        els = [s.strip() for s in ref.split(',')]
        if len(els) < 3:
            bad_refs.add(ref)
            continue
        try:
            year = int(els[1])
            if year < 1978:
                bad_refs.add(ref)
                continue
            elif year > 2010:
                after2010.add(ref)
                continue
        except ValueError:
            bad_refs.add(ref)
    # disregard hand_connected and known bad refs
    bad_refs -= known_bad_refs | hand_connected.keys()
    assert not bad_refs, str(bad_refs)
    return (refset - known_bad_refs), after2010 # - after2010

_key = '{volume}:{page_begin}'
disregard_dois = {'10.1007/BF02457983'}
```

```
def refsToSci(refset, sci_data, _key = _key):

    # volume:page_begin to doi dictionary
    vp_to_doi = {_key.format(**sci_data[doi]):doi for doi in (sci_data.keys() -
disregard_dois)}
    assert len(vp_to_doi) == len(sci_data) - len(disregard_dois)
    # author first year + year to doi dictionary, remove all items witch map to more than one
doi
    ay_to_doi = defaultdict(set)
    for doi, item in sci_data.items():
        if item['authors']:
            fasur = item['authors'][0].split(',')[0]
            ay_to_doi[fasur.lower() + str(item['year_pub'])].add(doi)
    for k, v in list(ay_to_doi.items()):
        if len(v) > 1:
            del(ay_to_doi[k])
        else:
            ay_to_doi[k] = v.pop()

    matched = defaultdict(set)
    unconnected = set()
    for ref in refset:
        # check if preconnected
        if ref in hand_connected:
            matched[ref].add(hand_connected[ref])
            continue

        els = [s.strip() for s in ref.split(',')]
        # try to match on doi
        if els[-1].startswith('DOI'):
            ref_doi = els[-1][3:].strip()
            if ref_doi in sci_data:
                matched[ref].add(ref_doi)
                continue
            els.pop()
        # try to match on volume-begin_page
        if els[-1].startswith('P') and els[-2].startswith('V'):
            key = '{}:{}'.format(els[-2][1:], els[-1][1:])
            if key in vp_to_doi:
                matched[ref].add(vp_to_doi[key])
                continue
        # try to match on aut_surname-year for all such keys that in
        # scientometrics data have only one matching doi
        key = els[0].split()[0].lower() + els[1]
        if key in ay_to_doi:
            matched[ref].add(ay_to_doi[key])
            continue
        unconnected.add(ref)
    # assert not unconnected
    return dict(matched), unconnected


def connectSciCites(wos_data, sci_data, wos_header):

    sci_self_cit = getJouSelfCit(sci_data, wos_header)
    assert not wos_data.keys() & sci_self_cit.keys()
    wos_data.update(sci_self_cit)

    raw_refs, no_refs, no_sciento_refs = getSciRefs(wos_data)
    assert not no_refs, str(no_refs)
    assert not no_sciento_refs, str(no_sciento_refs)

    clean_refs, after2010 = dropBadRefs(raw_refs)

    matched, unconnected = refsToSci(clean_refs, sci_data)
    assert not unconnected, str(sorted(unconnected))

    for ref in matched:
        assert(len(matched[ref]) == 1)
        matched[ref] = matched[ref].pop()

    tools.saveJson(matched, data_paths.kp_cit_matches)

    # update items with citation info
    no_cites = set()
    con_doi = set()
    dois11 = set(d for d in sci_data if sci_data[d]['year_pub'] > 2010)
```

```
    for wos_id in list(wos_data.keys()):
        item = wos_data[wos_id]
        refs = item['references']
        cites = set(filter(None, (matched.get(r) for r in refs)))
        if not cites:
            no_cites.add(wos_id)
        item['cites'] = list(cites)
        # drop items containing only references to papers after 2010
        if not cites - dois11:
            del wos_data[wos_id]
            continue
        con_doi.update(cites)

    assert con_doi.issubset(sci_data), con_doi - sci_data.keys()
    # item without any cites to scientometrics indicates a mistake in the selection process
    assert not no_cites, no_cites

    # reportOnRefMatch(wos_data, raw_refs, clean_refs, matched, con_doi)


def reportOnRefMatch(wos_data, raw_refs, clean_refs, matched, con_doi):
    all_refs = []
    for refs in map(itemgetter("references"), wos_data.values()):
        if refs:
            all_refs.extend(refs)

    print("n citing items: ", len(wos_data))
    print("n all refs: ", len(all_refs))
    print("n diff refs: ", len(set(all_refs)))
    print("n containing sciento: ", len(raw_refs))
    print("n matched refs: ", len(matched))
    print("n known unconnected: ", len(known_bad_refs))
    print("n hand connected: ", len(hand_connected))
    print("n cited dois: ", len(con_doi))

    #print()
    #ref_sample = sample(all_refs, 10)
    #for ref in ref_sample:
    #    print(ref)


def prepareWosCit(sci_data, cit_sets):
    # cit_sets = dataio.loadJson(data_paths.kp_cit_prepared)
    # sci_data = dataio.loadJson(data_paths.p_sci)
    print("Preparing WoS citations.")
    wos_data = cit_sets["wos"]["data"]
    wos_header = cit_sets["wos"]["header"]

    dropBadWoSItems(wos_data)
    correctCitJournals(wos_data)
    disambiguateCitSciAuthors(wos_data)

    connectSciCites(wos_data, sci_data, wos_header)

    tools.saveJson(wos_data, data_paths.p_wos_cit)
    print("- successful\n\n----\n")
    return wos_data

# if __name__ == '__main__':
#
#     prepareWosCit()
```

6.2.6 Izrada pripremljenog skupa svih podataka

```
from collections import defaultdict
from operator import itemgetter

from phd_tools import tools

from phd_scripts.static import data_paths


take_over = [
```

```
    "doi", "wos_id", "sco_id", "authors", "item_title", "publication_title",
    "year_pub", "volume", "issue", "page_begin", "page_end", "type",
    "prim_focus", "citation_count_wos", "citation_count_sco", "countries",
    "pdf_path", "abstract", "references_wos"
]

sci_header = [
    "doi", "wos_id", "sco_id", "authors", "item_title", "publication_title",
    "year_pub", "volume", "issue", "page_begin", "page_end", "type",
    "prim_focus", "citation_count_wos", "citation_count_sco", "citation_count",
    "cited_by", "countries", "pdf_path", "abstract", "references_wos"
]


def makeSciFinal(sci_merged, cit_wos):
    # print(sci_merged[next(iter(sci_merged))])
    # print(cit_wos[next(iter(cit_wos))])
    print("Making final data output.")
    cited_by = defaultdict(set)
    for wos_id, citing in cit_wos.items():
        for cited in citing["cites"]:
            cited_by[cited].add(wos_id)
        # citing["cites"] = list(citing["cites"])
    new_sci_data = dict()
    for doi, item in sci_merged.items():
        new_item = { k:item[k] for k in take_over}
        new_item["citation_count"] = len(cited_by[doi])
        new_item["cited_by"] = list(cited_by[doi])
        new_sci_data[doi] = new_item
    data = {
        "sci":new_sci_data,
        "sci_header":take_over,
        "cit":cit_wos,
    }
    tools.saveJson(data, data_paths.p_sci)
    tools.saveCsv(sci_header, map(itemgetter(*sci_header),
                                  data["sci"].values()), data_paths.p_sci_csv)
    print("- successful\n\n----\n")
    return data
```

6.2.7 Analiza pripremljenog skupa

6.2.7.1 Kontrolna skripta

```
import os, shutil
from phd_tools import tools

from phd_scripts.an.parts import table_spec, chart_spec, graph_spec
from phd_scripts.static import data_paths
from phd_scripts.an.dload import loadFinal


custom_titles = {
    "preuzimanje_i_priprema": "Preuzimanje i priprema ulaznih podataka za analizu",
    "osnovna_priprema": "Preliminarna priprema svih skupova podataka",
    "validacija_i_promjene": "Provjera vrijednosti za različite skupove podataka o radovima iz
časopisa *Scientometrics*",
    "graph_example_non_direct": "Primjer slučajno generiranog neusmjerenog grafa",
    "graph_example_direct": "Primjer slučajno generiranog usmjerenog grafa",
    "spajanje_sci": "Spajanje zapisa o radovima iz časopisa *Scientometrics*",
    "metrics": "Scientometrija, bibliometrija, informetrija ...",
    "graph_countries_art10":   "Mreža   međunarodne   suradnje   na   radovima   u   časopisu
*Scientometrics*",
    "graph_aut_full_1978-1988": "Mreža suradnje autora na člancima u časopisu "
                                "*Scientometrics* 1978-1988",
    "graph_aut_full_1989-1999": "Mreža suradnje autora na člancima u časopisu *Scientometrics*
1989-1999",
    "graph_aut_full_2000-2010": "Mreža suradnje autora na člancima u časopisu *Scientometrics*
2000-2010",
    "graph_aut_full_1978-2010": "Mreža suradnje autora na člancima u časopisu *Scientometrics*
1978-2010",
```

```
    "graph_aut_full_main_comp_1978-2010": "Glavna komponenta u mreži suradnje autora na
člancima u časopisu *Scientometrics* 1978-2010",
    "priprema_autora": "Ujednačavanje imena autora",
    "citatni_indeks": "Pojednostavljen prikaz citatnog indeksa",
    "citirani_citirajuci": "Citirajući rad, citat i citirani rad",
    "sci_cit_styles": "Stilovi navođenja literature korišteni u člancima u časopisu
*Scientometrics*",
    "classifier_gui": "Prikaz sučelja razvijenog za efikasnu klasifikaciju radova i priloga u
časopisu *Scientometrics*",
    "graph_key_aut": "Mreža suradnje autora koji su objavili više od 20 članaka "
                     "u časopisu *Scientometrics*",

}

figures = dict()


def addAnalysisDeliverables(deliverables, do_tables=True, do_charts=True,
                            do_graphs=True):
    # TODO USE these, make copies
    _tables = deliverables["tables"]
    _figs = deliverables["figures"]

    print("Loading data.")
    data = loadFinal()
    if do_tables:
        print("Making tables (n={})".format(len(table_spec)))
        tables = makeTables(data, table_spec)
        _tables.update(tables)

    if do_charts:
        print("Making figures (n={}).".format(len(chart_spec)))
        figs = makeCharts(data, chart_spec)
        _figs.update(figs)

    if do_graphs:
        print("Making graphs (n={}).".format(len(graph_spec)))
        figs = makeGraphs(data, graph_spec)
        _figs.update(figs)

    return deliverables


def addStaticImages(deliverables):
    figs = deliverables["figures"]
    p = data_paths.in_custom_img
    custom_images = os.listdir(p)
    for i in custom_images:
        tpath = data_paths.out_fig + i
        shutil.copy(p + i, tpath)
        n = '.'.join(i.split('.')[:-1])
        assert n not in figs
        rp = '.' + tpath[tpath.index('/parts'):]
        fig_item = {
            "resource_path": rp,
            "title": custom_titles.get(n, False),
        }
        figs[n] = fig_item


def reprFunc(item):
    return item['func'].__module__ + '.' + item['func'].__name__


def makeTables(data, table_spec):
    tables = {}
    for name, table_item in table_spec.items():
        d, h = table_item["func"](**data)

        table_item = table_item.copy()
        table_item["header"] = h
        table_item["data"] = d
        table_item['func'] = reprFunc(table_item)
        tables[name] = table_item
    return tables
```

```
def makeCharts(data, figures):
    import matplotlib
    from phd_tools import charttools
    matplotlib.rcParams.update(charttools.params)
    figs = {}
    for name, fig_item in figures.items():
        pth = fig_item["path"]
        fig = fig_item["func"](**data)
        fig.savefig(pth)
        fig_item = fig_item.copy()
        fig_item["resource_path"] = '.' + pth[pth.index('/parts'):]
        fig_item['func'] = reprFunc(fig_item)
        figs[name] = fig_item
    return figs


def makeGraphs(data, graphs):
    figs = {}
    for name, fig_item in graphs.items():
        fig_item = fig_item.copy()
        fig_item["func"](**data)
        pth = fig_item["path"]
        fig_item["resource_path"] = '.' + pth[pth.index('/parts'):]
        fig_item['func'] = reprFunc(fig_item)
        figs[name] = fig_item
    return figs


def printInventory():
    delivs = tools.loadJson(data_paths.p_deliv)
    fstr = "{} [ {} ]: {}"
    print("PHD Deliverables:\n")
    print("  Tables ({}):".format(len(delivs["tables"])))
    for t, d in delivs["tables"].items():
        print(fstr.format(" " * 4, t, d["title"]))
    print()
    print("  Figures ({}):".format(len(delivs["figures"])))
    for t, d in delivs["figures"].items():
        print(fstr.format(" " * 4, t, d["title"]))
```

6.2.7.2 Osnovni bibliometrijski pokazatelji

```
from collections import defaultdict, Counter
import numpy as np

from itertools import chain
from   phd_scripts.an.dload   import   aprod_group_names,   aut_prod_groups,   focus_sci_names,
ygroup_sci_names, \
    focus_legend
from phd_scripts.static import data_paths
from phd_tools import tools, charttools


def reportBase(sci, sci_groups, dois=None):
    dois = set(sci) if not dois else set(dois)

    met = sci_groups["prim_focus_art10"]["methods"]
    the = sci_groups["prim_focus_art10"]["theory"]
    apl = sci_groups["prim_focus_art10"]["applied"]

    all_auts = []
    n_auts_per_paper = []
    cits = []
    vols = set()
    issues = set()
    for doi in dois:
        item = sci[doi]
        vols.add(item['volume'])
        vols.add(item['issue'])
        a = item["authors"]
        c = item["citation_count"]
        if a:
            all_auts.extend(a)
            n_auts_per_paper.append(len(a))
```

```
        if c is None:
            cits.append(0)
        else:
            cits.append(c)

    aut_dist = Counter(all_auts)
    n_papers_per_aut = list(aut_dist.values())
    no_auts = aut_dist.pop(None, 0)

    rep = {
        "n radova": len(dois),
        "n članaka": len(dois & sci_groups["art"]),

        "n autora": len(aut_dist),
        "n radova bez autora": no_auts,

        "n volumena": len(vols),
        "n brojeva": len(issues),

        "n citata": sum(cits),
        "n citiranih radova": sum(1 for c in cits if c),

        "n jednoautorskih radova": sum(1 for n in n_auts_per_paper if n == 1),
        "n višeautorskh radova": sum(1 for n in n_auts_per_paper if n > 1),

        "n dvoautorskih radova": sum(1 for n in n_auts_per_paper if n == 2),
        "n troautorskih radova": sum(1 for n in n_auts_per_paper if n == 3),
        "n radova s 4 do 10 autora": sum(1 for n in n_auts_per_paper if 3 < n < 11),
        "n radova s preko 10 autora": sum(1 for n in n_auts_per_paper if n > 10),

        "prosječno autora po radu": round(np.mean(n_auts_per_paper), 2),
        "median autora po radu": tools.quantile(n_auts_per_paper),
        "max autora po radu": max(n_auts_per_paper),

        "prosječno radova po autoru": round(np.mean(n_papers_per_aut), 2),
        "median radova po autoru": tools.quantile(n_papers_per_aut),
        "max radova po autoru": max(n_papers_per_aut),

        "n teorijskih članaka": len(dois & the),
        "n metodoloških članaka": len(dois & met),
        "n primijenjenih članaka": len(dois & apl),
    }
    return rep


def tablePublishOverviewYgroup(sci, sci_groups, **kwargs):
    names = [
        "n radova",
        "n članaka",
        "n teorijskih članaka",
        "n metodoloških članaka",
        "n primijenjenih članaka",
        "n autora",
    ]

    year_keys = '1978-1988', '1989-1999', '2000-2010', '1978-2010', '2011-2012'
    reps = [reportBase(sci, sci_groups, sci_groups['ygroup'][k]) for k in year_keys]
    d, h = tools.joinDictReports(reps, names, year_keys)
    return d, h
tablePublishOverviewYgroup.title = "Pregled objavljivanja svih radova u časopisu
*Scientometrics* 1978-2012"


names_authorship = [
        "n radova",
        "n jednoautorskih radova",
        "n višeautorskh radova",

        "n dvoautorskih radova",
        "n troautorskih radova",
        "n radova s 4 do 10 autora",
        "n radova s preko 10 autora",

        "prosječno autora po radu",
        "median autora po radu",
        "max autora po radu",
    ]
```

```
def tableAuthorshipYgroupArt10(sci, sci_groups, **kwargs):

    reps = [reportBase(sci, sci_groups, sci_groups['ygroup_art10'][k]) for k in
ygroup_sci_names]
    d, h = tools.joinDictReports(reps, names_authorship, ygroup_sci_names)
    return d, h
tableAuthorshipYgroupArt10.title = "Autorstvo članaka objavljenih u časopisu *Scientometrics*
1978-2010 u odnosu na godine objave"


def tableAuthorshipFocusArt10(sci, sci_groups, **kwargs):
    reps = [reportBase(sci, sci_groups, sci_groups['focus_art10'][k]) for k in
focus_sci_names]
    d, h = tools.joinDictReports(reps, names_authorship, focus_sci_names)
    return d, h
tableAuthorshipFocusArt10.title = "Autorstvo članaka u *Scientometrics* " \
                                  "1978-2010 u odnosu na tematiku članaka"


def tableAuthorshipPrimFocusArt10(sci, sci_groups, **kwargs):
    reps = [reportBase(sci, sci_groups, sci_groups['prim_focus_art10'][k]) for k in
focus_sci_names]
    d, h = tools.joinDictReports(reps, names_authorship, focus_sci_names)
    return d, h
tableAuthorshipPrimFocusArt10.title = "Autorstvo članaka u *Scientometrics* " \
                                      "1978-2010 u odnosu na tematiku članaka"


def tableAuthorsPubYgroupArt10(sci, sci_groups, **kwargs):
    names = [
        "n radova",
        "n autora",

        "prosječno radova po autoru",
        "median radova po autoru",
        "max radova po autoru",
    ]

    reps = [reportBase(sci, sci_groups, sci_groups['ygroup_art10'][k]) for k in
ygroup_sci_names]
    d, h = tools.joinDictReports(reps, names, ygroup_sci_names)

    d[2][0] = "prosječno članaka po autoru"
    d[3][0] = "median članaka po autoru"
    d[4][0] = "max članaka po autoru"

    return d, h
tableAuthorsPubYgroupArt10.title = "Pregled produktivnosti autora u časopisu *Scientometrics*
1978-2010"


def tableAuthorsPubNpubArt10(sci, sci_groups, **kwargs):
    aut_groups_art10 = tools.groupToOverlaping(sci, "authors",
items=sci_groups['ygroup_art10']['1978-2010'])
    key_func = tools.groupIntervals(aut_prod_groups, len)
    aut_prod_data = defaultdict(lambda: {"n radova": set(), "n autora": 0})
    for dois in aut_groups_art10.values():
        aut_prod_data[key_func(dois)]["n radova"].update(dois)
        aut_prod_data[key_func(dois)]["n autora"] += 1

    d = []
    h = ["n članaka po autoru", "n autora", "n članaka"]
    for k in aprod_group_names:
        row = [
            k,
            aut_prod_data[k]["n autora"],
            len(aut_prod_data[k]["n radova"]),
        ]
        d.append(row)
    return d, h
tableAuthorsPubNpubArt10.title = "Produktivnost autora s obzirom na broj " \
                                 "članaka objavljenih u *Scientometricsu*"


def tableTypesYears2010(sci, sci_groups, **kwargs):
    h = ["vrsta rada"] + list(ygroup_sci_names)
```

```
    d = []
    for typ in sorted(sci_groups["type10"]):
        row = [typ]
        for ygroup in ygroup_sci_names:
            dois = sci_groups["type10"][typ] & sci_groups["ygroup"][ygroup]
            row.append(len(dois))
        d.append(row)
    return d, h
tableTypesYears2010.title = "Broj radova prema vrsti i godini objave"


base_table_spec = [
    (tablePublishOverviewYgroup, data_paths.t_ygroup_base_overview),
    (tableAuthorshipYgroupArt10, data_paths.t_authorship_ygroup_art10),
    (tableAuthorshipFocusArt10, data_paths.t_authorship_focus_art10),
    (tableAuthorshipPrimFocusArt10, data_paths.t_authorship_prim_focus_art10),
    (tableAuthorsPubYgroupArt10, data_paths.t_authors_pub_ygroup_art10),
    (tableAuthorsPubNpubArt10, data_paths.t_authors_pub_npap_art10),
    (tableTypesYears2010, data_paths.t_types_years_10),
]


def figureArtsAuts10(sci, sci_groups, **kwargs):
    ycounts = []
    autcounts = []
    years = sci_groups["year_art10"]
    for y in sorted(years):
        ycounts.append(len(years[y]))
        autcounts.append(len(set(chain.from_iterable(sci[doi]["authors"]     for     doi     in
years[y]))))
    # auts and papers per year
    fig, ax = charttools.makeFig("11.5cm")
    charttools.lineDisc(ax, [ycounts, autcounts], line_style = ['-', '--'],
                        labels=["n članaka", "n autora"], lw=2)
    ax.set_xticklabels(xtick_years)
    ax.legend(loc="upper left")
    return fig
figureArtsAuts10.title = "Broj članaka i autora članaka po godinama objave"

def figurePrimFocusYear(sci, sci_groups, **kwargs):
    vals = []
    years = range(1978, 2011)
    for focus in focus_sci_names:
        f = []
        for y in years:
            yset = sci_groups['year_art10'][y]
            nyear = len(sci_groups["prim_focus_art10"][focus] & yset)
            f.append(nyear)
        vals.append(f)
    # auts and papers per year
    fig, ax = charttools.makeFig("11.5cm")
    charttools.lineDisc(ax, vals, line_style = ['-', '--', '-.'], lw=2,
                        labels=focus_legend)
    ax.set_xticklabels(xtick_years)
    ax.legend(loc="upper left")
    return fig
figurePrimFocusYear.title = "Tematika članaka po godinama objave"

xtick_years = [1978, 1983, 1988, 1993, 1998, 2003, 2008, 2013]

def figurePapsArts10(sci, sci_groups, **kwargs):
    ycounts = []
    autcounts = []
    years = sci_groups["year"]
    for y in sorted(years):
        ycounts.append(len(years[y]))
        autcounts.append(sum(1 for doi in years[y] if sci[doi]["type"] == "article"))
    # auts and papers per year
    fig, ax = charttools.makeFig("11.5cm")
    charttools.lineDisc(ax, [ycounts, autcounts], line_style = ['-', '--'],
                        lw=2, labels= ['radovi i prilozi', 'članci'])
    ax.set_xticklabels(xtick_years)
    ax.legend(loc = 'upper left')
    return fig
figurePapsArts10.title = "Broj svih radova i članaka po godinama objave"
```

```
def multiAutPropArts10(sci, sci_groups, **kwargs):
    props = []
    years = sci_groups["year_art10"]
    for y in sorted(years):
        tot_arts = len(years[y])
        n_multi    =    sum(1    for    doi    in    years[y]    if    sci[doi]["authors"]    and
len(sci[doi]["authors"]) > 1)
        props.append(n_multi/tot_arts)
    fig, ax = charttools.makeFig("11.5cm")
    charttools.lineDisc(ax, [props], line_style = ['-'], lw=2)
    ax.set_xticklabels(xtick_years)
    return fig
multiAutPropArts10.title = "Proporcija višeautorskih članaka po godinama objave"

from matplotlib import pyplot as plt

def piesFocusYGroups(sci, sci_groups, **kwargs):
    fs = list(charttools.getFigSize("15cm"))
    fs[1] = fs[0]
    fig, axes = plt.subplots(2, 2, figsize = fs)

    gy = sci_groups["ygroup_art10"]
    fp = gy['1978-1988']
    sp = gy['1989-1999']
    tp = gy['2000-2010']
    ap = gy['1978-2010']

    met = sci_groups["prim_focus_art10"]["methods"]
    the = sci_groups["prim_focus_art10"]["theory"]
    apl = sci_groups["prim_focus_art10"]["applied"]


    full = [len(ap & apl), len(ap & met), len(ap & the)]
    f = [len(fp & apl), len(fp & met), len(fp & the)]
    s = [len(sp & apl), len(sp & met), len(sp & the)]
    t = [len(tp & apl), len(tp & met), len(tp & the)]


    def pie(ax, ratios, t):
        ax.set_aspect('equal')
        ax.set_title(t)
        wedg = ax.pie(ratios, colors=charttools.cs_van_gogh,pctdistance=0.6,
                      autopct='%1.1f%%')

        return wedg[0][:3]

    pie(axes[0][0], f, '1978-1988')
    pie(axes[0][1], s, '1989-1999')
    pie(axes[1][0], t, '2000-2010')
    wedg = pie(axes[1][1], full, '1978-2010')
    fig.legend(wedg, focus_legend, loc='lower center')
    fig.subplots_adjust(0, 0, 1, 0.92)
    return fig
piesFocusYGroups.title = "Udijeli radova po tematici za sva vremenska " \
                         "razdoblja"


def piesFocusYGroupsHigh(sci, sci_groups, **kwargs):
    fs = list(charttools.getFigSize("15cm"))
    fs[1] = fs[0]
    fig, axes = plt.subplots(2, 2, figsize = fs)

    hc = sci_groups["highly_cited_art10"]
    gy = sci_groups["ygroup_art10"]
    fp = gy['1978-1988'] & hc
    sp = gy['1989-1999'] & hc
    tp = gy['2000-2010'] & hc
    ap = gy['1978-2010'] & hc

    met = sci_groups["prim_focus_art10"]["methods"]
    the = sci_groups["prim_focus_art10"]["theory"]
    apl = sci_groups["prim_focus_art10"]["applied"]


    full = [len(ap & apl), len(ap & met), len(ap & the)]
    f = [len(fp & apl), len(fp & met), len(fp & the)]
    s = [len(sp & apl), len(sp & met), len(sp & the)]
```

```
    t = [len(tp & apl), len(tp & met), len(tp & the)]


    def pie(ax, ratios, t):
        ax.set_aspect('equal')
        ax.set_title(t)
        wedg = ax.pie(ratios, colors=charttools.cs_van_gogh,pctdistance=0.6,
                      autopct='%1.1f%%')

        return wedg[0][:3]

    pie(axes[0][0], f, '1978-1988')
    pie(axes[0][1], s, '1989-1999')
    pie(axes[1][0], t, '2000-2010')
    wedg = pie(axes[1][1], full, '1978-2010')
    fig.legend(wedg, focus_legend, loc='lower center')
    fig.subplots_adjust(0, 0, 1, 0.92)
    return fig
piesFocusYGroupsHigh.title = "Udjeli teorijskih, metodoloških i primijenjenih visokocitiranih
članaka"

base_chart_spec = charts = [
    (figureArtsAuts10, data_paths.f_arts_auts_10),
    (figurePapsArts10, data_paths.f_paps_arts_10),
    (multiAutPropArts10, data_paths.f_multi_aut_prop_art10),
    (figurePrimFocusYear, data_paths.f_prim_focus_year),
    (piesFocusYGroups, data_paths.f_prim_focus_pies),
    (piesFocusYGroupsHigh, data_paths.f_prim_focus_pies_high),
]
```

6.2.7.3 Analiza koautrostva

```
from collections import defaultdict
from itertools import chain, combinations
from operator import itemgetter

from phd_tools import tools, charttools, graphtools

from phd_scripts.an import dload
from phd_scripts.static import data_paths

n_random = 1

def tableAutGraphOverviewYGroupArt10(sci, sci_groups, **kwargs):
    names = [
        'n radova',
        "n radova u najvećoj komponenti",
        'n čvorova',
        "n čvorova u najvećoj komponenti",
        'n veza',
        "n veza u najvećoj komponenti",
    ]
    reps = []
    for k in dload.ygroup_sci_names:
        dois = sci_groups['ygroup_art10'][k]
        graph = graphtools.getAutGraph(sci, dois)
        reps.append(graphtools.reportAuthorGraph(graph))
    d, h = tools.joinDictReports(reps, names, dload.ygroup_sci_names)
    for i in range(1, 6, 2):
        d[i][0] = "  u najvećoj komponenti"
        for ii in range(1, 5):
            d[i][ii] = "{} ({}%)".format(d[i][ii], round(d[i][ii] / d[i-1][ii] * 100), 2)
    return d, h
tableAutGraphOverviewYGroupArt10.title = "Pregled mreže koautorstva na člancima objavljenim u
časopisu *Scientometrics* 1978-2010"

def tableAutsPeriodArt10(sci, sci_groups, **kwargs):
    auts = []
    for k in dload.ygroup_sci_names[:-1]:
        dois = sci_groups['ygroup_art10'][k]
        auts.append( set(chain.from_iterable(sci[doi]["authors"] for doi in dois)) )
    first, second, third = auts

    table = [
```

```
        ['n autora', first, second, third],
        ['n privremenih', first - (second | third), second - (first | third), third - (first |
second) ],
        ['n pridošlih', first, second - first, third - (first | second) ],
        ['n terminatora', first - (second | third), second - third, third ],
        ['n kontinuiranih', (first & second) | (first & third), second & (first | third),
(first & third) | (second & third)],
    ]

    for row in table:
        for i, v in enumerate(row[1:], 1):
            row[i] = len(v)
    return table, ['pokazatelj'] + list(dload.ygroup_sci_names[:-1])
tableAutsPeriodArt10.title = "Promjene u broju autora po razdoblju"

def tableGraphFeaturesYGroupArt10(sci, sci_groups, **kwargs):

    names = [
        'n radova',
        'n čvorova',
        'n veza',
        "gustoća",
        "dijametar",
        "asortativnost",
        "n artikulacijskih čvorova",
        "globalni koeficijent grupiranja",
        "slučajni globalni koeficijent grupiranja (n=%s)" % n_random,
        "prosječna najkraća duljina puta",
        "slučajna prosječna najkraća duljina puta (n=%s)" % n_random,
    ]
    reps = []
    for k in dload.ygroup_sci_names:
        dois = sci_groups['ygroup_art10'][k]
        graph = graphtools.getAutGraph(sci, dois)
        reps.append(graphtools.reportAuthorGraph(graph, n_random=n_random))
    d, h = tools.joinDictReports(reps, names, dload.ygroup_sci_names)
    return d, h

tableGraphFeaturesYGroupArt10.title = "Značajke mreže koautorstva na člancima objavljenim u
časopisu *Scientometrics* 1978-2010"

def tableGraphFeaturesMainCompYGroupArt10(sci, sci_groups, **kwargs):
    names = [
        'n radova',
        'n čvorova',
        'n veza',
        "gustoća",
        "dijametar",
        "asortativnost",
        "n artikulacijskih čvorova",
        "prosječan stupanj centralnosti",
        "globalni koeficijent grupiranja",
        "slučajni globalni koeficijent grupiranja (n=%s)" % n_random,
        "prosječna najkraća duljina puta",
        "slučajna prosječna najkraća duljina puta (n=%s)" % n_random,
    ]
    graph = graphtools.getAutGraph(sci, sci_groups['ygroup_art10'][dload.ygroup_sci_names[-
1]])
    mc = graphtools.getLargestComponent(graph)
    rep = graphtools.reportAuthorGraph(mc, n_random=n_random)
    d = [[k, rep[k]] for k in names]

    return d, ["pokazatelj", "n"]
tableGraphFeaturesMainCompYGroupArt10.title = "Značajke glavne komponente mreže koautorstva na
člancima objavljenim u časopisu *Scientometrics* 1978-2010"

def tableCoopFeaturesYGroupArt10(sci, sci_groups, **kwargs):
    names = [
        "prosječan stupanj centralnosti",
        "StD stupnja centralnosti",
        "median stupnja centralnosti",
        "max stupanj centralnosti",
        "n izoliranih čvorova",
        "n izoliranih dijada",
        "n izoliranih trijada",
        "n ostalih komponenti (n>3)",
    ]
```

```
    reps = []
    for k in dload.ygroup_sci_names:
        dois = sci_groups['ygroup_art10'][k]
        graph = graphtools.getAutGraph(sci, dois)
        reps.append(graphtools.reportAuthorGraph(graph))
    d, h = tools.joinDictReports(reps, names, dload.ygroup_sci_names)
    return d, h
tableCoopFeaturesYGroupArt10.title = "Suradnja na člancima objavljenim u časopisu
*Scientometrics* 1978-2010"

def getCoop(sci_groups):
    aut_data = sci_groups['auts_art10']
    coops = defaultdict(lambda: dict())
    for a, b in combinations(aut_data, 2):
        n_coop = len(aut_data[a] & aut_data[b])
        if n_coop:
            coops[a][b] = n_coop
            coops[b][a] = n_coop
    return coops

def getTopCoops(a, coops):
    c = coops[a]
    if not c:
        return 0, ["bez suradnje"]
    top = max(c.values())
    ca = []
    for b in c:
        if c[b] == top:
            ca.append(b)
    ca.sort()
    return top, ca


def tableTopAutsOverviewArt10(sci, sci_groups, more_than=19, **kwargs):
    d = []
    h = ["autor", "*n* radova", "*n* citata", "*h*-index",
         "*g*-index", "najčešća suradnja", "min-max godina objave"]

    aut_data = sci_groups['auts_art10']
    coops = getCoop(sci_groups)
    for a, dois in aut_data.items():
        if len(dois) > more_than:
            years_pub = [sci[doi]["year_pub"] for doi in dois]
            cits = [len(sci[doi]["cited_by"]) for doi in dois]
            cps = getTopCoops(a, coops)
            row = [
                a,
                len(dois),
                sum(cits),
                tools.hIndex(cits),
                tools.gIndex(cits),
                "{}:{}".format(cps[0], ';'.join(cps[1])),
                "{}-{}".format(min(years_pub), max(years_pub)),
            ]
            d.append(row)
    d.sort(key=itemgetter(4), reverse=True)
    return d, h
tableTopAutsOverviewArt10.title = "Pregled autora koji su u časopisu *Scientometrics* 1978-
2010 objavili više od 19 članaka"


def tableCountriesOverviewArt10(sci, sci_groups, more_than=19, **kwargs):
    d = []
    h = ["autor", "broj radova", "min-max godina objave"]
    cdata = sci_groups['countries_art10']
    for c, dois in cdata.items():
        if len(dois) > more_than:
            years_pub = [sci[doi]["year_pub"] for doi in dois]
            row = [
                c,
                len(dois),
                "{}-{}".format(min(years_pub), max(years_pub))
            ]
            d.append(row)
    d.sort(key=itemgetter(1), reverse=True)
    return d, h
```

```
tableCountriesOverviewArt10.title = "Pregled zemalja koje su zastupljene u časopisu
*Scientometrics* s više od 19 članaka"

aut_table_spec = [
    (tableAutGraphOverviewYGroupArt10, data_paths.t_aut_graph_overview_ygroup_art10),
    (tableGraphFeaturesYGroupArt10, data_paths.t_aut_graph_features_ygroup_art10),
    (tableCoopFeaturesYGroupArt10, data_paths.t_coop_features_ygroup_art10),
    (tableTopAutsOverviewArt10, data_paths.t_top_auts_overview_art10),
    (tableAutsPeriodArt10, data_paths.t_auts_period_art10),
    (tableGraphFeaturesMainCompYGroupArt10, data_paths.t_aut_graph_features_mc_ygroup_art10),
    (tableCountriesOverviewArt10, data_paths.t_countries_overview_art10),
]

from matplotlib import pyplot as plt


def distAutsPerPap(sci, sci_groups, **kwargs):
    aut_ns = []
    for doi in sci_groups['ygroup_art10']['1978-2010']:
        auts = sci[doi]["authors"]
        if auts:
            aut_ns.append(len(auts))
    fig, ax = charttools.makeFig("11.5cm")
    charttools.histDisc(ax, aut_ns)
    return fig
distAutsPerPap.title = "Distribucija radova po broju autora"

def distDegreePerAut(sci, sci_groups, **kwargs):
    figs, axes = plt.subplots(2, 2)
    row = 0
    col = 0
    for k in dload.ygroup_sci_names:
        if col > 1:
            row += 1
            col = 0
        graph = graphtools.getAutGraph(sci, sci_groups['ygroup_art10'][k])
        charttools.histDisc(axes[row][col], graph.degree())
        axes[row][col].set_title(k)
        col += 1
    return figs
distDegreePerAut.title = "Distribucija stupnja centralnosti po autoru"

aut_chart_spec = [
    (distAutsPerPap, data_paths.f_dist_auts_per_pap_art10),
    (distDegreePerAut, data_paths.f_dist_degree_per_aut),
]
```

6.2.7.4 Citatna analiza

```
from collections import Counter, defaultdict
from itertools import chain, combinations
from operator import itemgetter

import numpy as np

from phd_tools import tools, charttools

from phd_scripts.static import data_paths
from phd_scripts.an.dload import ygroup_sci_names, focus_sci_names, \
    ygroup_cit_names


def reportCitingBase(sci_data, cit_data, dois=None, cit_keys=None):
    dois = set(sci_data) if not dois else dois
    cit_keys = set(cit_data) if not cit_keys else cit_keys

    citing_ids = set()
    all_cited_dois = []
    aut_selfcited_dois = []
    jou_selfcited_dois = []
    all_citing_refs = []

    cite_sci_per_paper = []
    refs_per_paper = []
```

```
ratios_refs_to_sci = []

cites_age = []

for cit_key in cit_keys:
    item = cit_data[cit_key]
    cites = set(item["cites"]) & dois
    if not cites:
        continue

    if item["authors"]:
        citing_auts = set(item["authors"])
        for doi in cites:
            ca = sci_data[doi]["authors"]
            if ca and set(sci_data[doi]["authors"]) & citing_auts:
                aut_selfcited_dois.append(doi)

    if item["publication_title"] == "Scientometrics":
        jou_selfcited_dois.extend(cites)

    year_pub = item["year_pub"]
    cites_age.append(
        [year_pub - sci_data[doi]["year_pub"] for doi in cites])

    citing_ids.add(cit_key)
    all_cited_dois.extend(cites)
    # include all cites to scientometrics, not just in this subset,
    # "items that cited scientometirics in sets have this n total sciento cites"
    cite_sci_per_paper.append(len(item["cites"]))
    all_citing_refs.extend(item["references"])
    refs_per_paper.append(len(item["references"]))
    ratios_refs_to_sci.append(len(item["cites"]) / len(item["references"]))

cits = Counter(all_cited_dois)
selfcits_author = Counter(aut_selfcited_dois)
selfcits_jou = Counter(jou_selfcited_dois)
refs = Counter(all_citing_refs)

# TO DO, add proper range desc

rep = {

    "n citata": sum(cits.values()),
    "n Scientometrics radova": len(dois),
    "n citiranih radova": len(cits),
    "prosječno citata po citiranom radu": round(
        np.mean(list(cits.values())), 2),
    "medijan citata po citiranom radu": tools.quantile(cits.values()),
    "iqr citata po citiranom radu": tools.iqRange(cits.values()),
    "maks. citata po citiranom radu": max(cits.values()),

    "n citirajućih radova": len(citing_ids),
    "n citirajućih časopisa": len(set(
        cit_data[wos_id]["publication_title"] for wos_id in citing_ids)),

    "n referenci svih citirajućih radova": sum(refs_per_paper),
    "n različitih referenci svih citirajućih radova": len(refs),
    "mean referenci po radu": round(np.mean(refs_per_paper), 2),
    "medijan referenci po radu": tools.quantile(refs_per_paper),
    "IQR referenci po radu": tools.iqRange(refs_per_paper),
    "maks. referenci po radu": max(refs_per_paper),

    "prosječno citata po citirajućem radu": round(
        np.mean(cite_sci_per_paper), 2),
    "medijan citata po citirajućem radu": tools.quantile(
        cite_sci_per_paper),
    "IQR citata po citirajućem radu": tools.iqRange(refs_per_paper),

    "prosječan udio citata na Sci. u referencama": round(
        np.mean(ratios_refs_to_sci), 2),

    "n samocitata časopisa": sum(selfcits_jou.values()),
    "n samocitata autora": sum(selfcits_author.values()),

    "medijan starost citata": tools.quantile(
        map(tools.quantile, cites_age)),
    "maks. starost citata": max(chain.from_iterable(cites_age))
```

```
    }
    return rep


def tableTopCitingJournalsArt10(sci, cit, sci_groups, **kwargs):
    dois_art10 = sci_groups['ygroup_art10']['1978-2010']
    jou = defaultdict(set)

    for cit_id, cit_item in cit.items():
        jou[cit_item["publication_title"]].add(cit_id)

    top_20 = sorted(jou, key=lambda x: sum(len(cit[i]['cites']) for i in jou[
        x]), reverse=True)[:20]
    h = ["časopis", "n citata", "n citirajućih radova", "n citiranih radova",
         "prosječno udjela citata u referencama"]
    rep = []
    for j in top_20:
        _row = reportCitingBase(sci, cit, dois_art10, jou[j])
        row = [
            j,
            _row["n citata"],
            _row["n citirajućih radova"],
            _row["n citiranih radova"],
            _row["prosječan udio citata na Sci. u referencama"],
        ]
        rep.append(row)
    rep.sort(key=itemgetter(1), reverse=True)
    return rep, h


tableTopCitingJournalsArt10.title = \
    "Top 20 časopisa indeksiranih u WoS-u 1978-2012 koji " \
    "su objavljivali radove koji citiraju časopis *Scientometrics*"


def tableCitationsOverview(sci, cit, sci_groups, **kwargs):
    h_table_cit1 = [
        "n citirajućih radova",
        "n citirajućih časopisa",
        "n citata",
        "maks. citata po citiranom radu",
        "n citiranih radova",
        "n Scientometrics radova",
    ]
    dois_art10 = sci_groups['primary']

    base_rep = reportCitingBase(sci, cit)
    art10_rep = reportCitingBase(sci, cit, dois_art10)
    rest_rep = reportCitingBase(sci, cit, sci.keys() - dois_art10)
    d, h = tools.joinDictReports([base_rep, art10_rep, rest_rep], h_table_cit1,
                                 ["svi radovi", "članci 2010", "ostalo"])
    return d, h


tableCitationsOverview.title = \
    "Pregled citata svih radova u časopisu Scientometrics " \
    "u časopisima indeksiranim u WoS citatnim indeksima 1978-2012"


def tableCitationsOverviewNonArt(sci, cit, sci_groups, **kwargs):
    h = ["vrsta rada"] + list(ygroup_sci_names)
    t = []

    def sumCits(sci, dois):
        return sum(len(sci[doi]["cited_by"]) for doi in dois)

    for type_name, tset in sci_groups['type10'].items():
        row = [type_name]
        for ygroup in ygroup_sci_names:
            row.append(sumCits(sci, sci_groups["ygroup"][ygroup] & tset))
        t.append(row)
    t.sort(key=itemgetter(3))
    return t, h


tableCitationsOverviewNonArt.title = "Pregled citata prema vrstama radova i priloga"
```

```
def tableCitingPapersYgroupArt10(sci, cit, sci_groups, cit_groups, **kwargs):
    dois_art10 = sci_groups['ygroup_art10']['1978-2010']

    names = [
        "n citirajućih časopisa",
        "n citirajućih radova",
        "n citiranih radova",
        "n citata",
        "prosječno citata po citirajućem radu",
        "medijan citata po citirajućem radu",
        "IQR citata po citirajućem radu",
        "n referenci svih citirajućih radova",
        "mean referenci po radu",
        "medijan referenci po radu",
        "prosječan udio citata na Sci. u referencama",
    ]
    reps = [reportCitingBase(sci, cit, dois_art10, cit_groups['ygroup'][k]) for
            k in ygroup_cit_names]
    d, h = tools.joinDictReports(reps, names, ygroup_cit_names)
    return d, h


tableCitingPapersYgroupArt10.title = \
    "Pregled radova objavljenih u časopisima u WoS " \
    "citatnim indeksima 1978-2012 koji su citirali članke " \
    "objavljene u časopisu *Scientometrics* 1978-2010"

cited_names = [
    "n Scientometrics radova", "n citiranih radova", "n citata",
    "n citirajućih radova", "n citirajućih časopisa",
    "prosječno citata po citiranom radu", "medijan citata po citiranom radu",
    "iqr citata po citiranom radu", "maks. citata po citiranom radu"
]


def tableCitedPapersYgroupArt10(sci, cit, sci_groups, **kwargs):
    reps = [reportCitingBase(sci, cit, sci_groups['ygroup_art10'][k]) for k in
            ygroup_sci_names]
    d, h = tools.joinDictReports(reps, cited_names, ygroup_sci_names)
    return d, h


tableCitedPapersYgroupArt10.title = \
    "Pregled citata članaka u časopisu *Scientometrics* " \
    "1978-2010 u svim časopisima indeksiranim u WoS citatnim indeksima 1978-2012"


def tableCitedPapersFocusArt10(sci, cit, sci_groups, **kwargs):
    reps = [reportCitingBase(sci, cit, sci_groups['focus_art10'][k]) for k in
            focus_sci_names]
    d, h = tools.joinDictReports(reps, cited_names, focus_sci_names)
    return d, h


tableCitedPapersFocusArt10.title = \
    "Pregled citata članaka u časopisu *Scientometrics* u odnosu na njihov fokus"


def tableCitedPapersPrimFocusArt10(sci, cit, sci_groups, **kwargs):
    reps = [reportCitingBase(sci, cit, sci_groups['prim_focus_art10'][k]) for k
            in focus_sci_names]
    d, h = tools.joinDictReports(reps, cited_names, focus_sci_names)
    return d, h


tableCitedPapersPrimFocusArt10.title = \
    "Pregled citata članaka u časopisu *Scientometrics* u odnosu na njihov primaran fokus"


def tableCitedPapersHighlyFocusArt10(sci, cit, sci_groups, **kwargs):
    reps = [reportCitingBase(sci, cit,
                             sci_groups['prim_focus_art10'][k] & sci_groups[
                                 'highly_cited_art10']) for k in
            focus_sci_names]
    d, h = tools.joinDictReports(reps, cited_names, focus_sci_names)
    return d, h
```

```
tableCitedPapersHighlyFocusArt10.title = \
    "Pregled citata visoko citiranih članaka u časopisu *Scientometrics* u odnosu na njihov
primaran fokus"


def tableSelfCitedPapersYgroupArt10(sci, cit, sci_groups, cit_groups, **kwargs):
    # dois_art10 = sci_groups['ygroup_art10']['1978-2010']
    names = [
        "n citata",
        "n samocitata časopisa",
        "n samocitata autora",
    ]

    reps = [reportCitingBase(sci, cit, sci_groups['ygroup_art10'][k]) for k in
            ygroup_sci_names]
    d, h = tools.joinDictReports(reps, names, ygroup_sci_names)

    reps_sciento = [reportCitingBase(sci, cit, sci_groups['ygroup_art10'][k],
                                     cit_groups['jou']['sciento']) for k in
                    ygroup_sci_names]
    dj, hj = tools.joinDictReports(reps_sciento, names, ygroup_sci_names)
    for row in dj:
        row[0] = "  u Scientometrics"

    reps_nonsciento = [reportCitingBase(sci, cit, sci_groups['ygroup_art10'][k],
                                        cit_groups['jou']['ostalo']) for k in
                       ygroup_sci_names]
    dn, hn = tools.joinDictReports(reps_nonsciento, names, ygroup_sci_names)
    for row in dn:
        row[0] = "  u ostalim časopisima"

    #d.insert(1, dj[0])
    #d.insert(2, dn[0])
    #d.insert(4, dj[1])
    #d.insert(5, dn[1])
    d.append(dj[2])
    d.append(dn[2])

    return d, h


tableSelfCitedPapersYgroupArt10.title = "Samocitati autora i časopisa"


def tableCitationOldness(sci, cit, sci_groups, cit_groups, **kwargs):
    dois_art10 = sci_groups['ygroup_art10']['1978-2010']
    names = [
        # "prosječno starost citata",
        "medijan starost citata",
        "maks. starost citata",
    ]

    reps = [reportCitingBase(sci, cit, sci_groups['ygroup_art10'][k])
            for k in ygroup_sci_names]
    d, h = tools.joinDictReports(reps, names, ygroup_sci_names)

    reps_sciento = [reportCitingBase(sci, cit, sci_groups['ygroup_art10'][k],
                                     cit_groups['jou']['sciento']) for k in
                    ygroup_sci_names]
    dj, hj = tools.joinDictReports(reps_sciento, names, ygroup_sci_names)
    for row in dj:
        row[0] = "  u Scientometrics"

    reps_nonsciento = [reportCitingBase(sci, cit, sci_groups['ygroup_art10'][k],
                                        cit_groups['jou']['ostalo']) for k in
                       ygroup_sci_names]
    dn, hn = tools.joinDictReports(reps_nonsciento, names, ygroup_sci_names)

    for row in dn:
        row[0] = "  u ostalim časopisima"

    d.insert(1, dj[0])
    d.insert(2, dn[0])
    d.append(dj[1])
    d.append(dn[1])
```

```
    return d, h


tableCitationOldness.title = "Starost citata članaka u časopisu " \
                             "Scientometrics u odnosu na razdoblja"


def tableCitationOldnessFocus(sci, cit, sci_groups, cit_groups, **kwargs):
    dois_art10 = sci_groups['ygroup_art10']['1978-2010']
    names = [
        # "prosječno starost citata",
        "medijan starost citata",
        "maks. starost citata",
    ]

    reps = [reportCitingBase(sci, cit, sci_groups['prim_focus_art10'][k]) for k
            in focus_sci_names]
    d, h = tools.joinDictReports(reps, names, focus_sci_names)

    reps_sciento = [
        reportCitingBase(sci, cit, sci_groups['prim_focus_art10'][k],
                         cit_groups['jou']['sciento']) for k in
        focus_sci_names]
    dj, hj = tools.joinDictReports(reps_sciento, names, focus_sci_names)
    for row in dj:
        row[0] = "  u Scientometrics"

    reps_nonsciento = [
        reportCitingBase(sci, cit, sci_groups['prim_focus_art10'][k],
                         cit_groups['jou']['ostalo']) for k in focus_sci_names]
    dn, hn = tools.joinDictReports(reps_nonsciento, names, focus_sci_names)

    for row in dn:
        row[0] = "  u ostalim časopisima"

    d.insert(1, dj[0])
    d.insert(2, dn[0])
    d.append(dj[1])
    d.append(dn[1])

    return d, h


tableCitationOldnessFocus.title = "Starost citata članaka u časopisu " \
                                  "Scientometrics u odnosu na tematiku"

cit_table_spec = [
    (tableCitationsOverview,
     data_paths.t_citations_overview),

    (tableCitedPapersYgroupArt10,
     data_paths.t_cited_papers_ygroup_art10),

    (tableCitedPapersFocusArt10,
     data_paths.t_cited_papers_focus_art10),

    (tableCitedPapersHighlyFocusArt10,
     data_paths.t_cited_papers_highly_focus_art10),

    (tableCitedPapersPrimFocusArt10,
     data_paths.t_cited_papers_prim_focus_art10),

    (tableCitingPapersYgroupArt10,
     data_paths.t_citing_papers_ygroup_art10),

    (tableSelfCitedPapersYgroupArt10,
     data_paths.t_self_cited_papers_ygroup_art10),

    (tableTopCitingJournalsArt10,
     data_paths.t_top_citing_journals_art10),

    (tableCitationOldness,
     data_paths.t_citation_oldness_art10),

    (tableCitationsOverviewNonArt,
     data_paths.t_citations_overview_non_art),
```

```
    (tableCitationOldnessFocus,
     data_paths.t_citation_oldness_art10_focus),
]


import matplotlib.gridspec as gridspec
from matplotlib import pyplot as plt


def figureCitedDistro(sci, sci_groups, sci_info, **kwargs):
    cutoff = sci_info["highly_cited_cutoff"]

    v = [sci[doi]['citation_count'] for doi in sci_groups['primary']]
    hv = [c for c in v if c > cutoff]
    assert len(hv) == len(sci_groups["highly_cited_art10"])
    tot = sum(v)
    hc_per = round((sum(hv) / tot) * 100, 2)
    hc_n_per = round((len(hv) / len(v)) * 100, 2)
    text = "n rad={} ({}%)\nn cit={} ({}%)".format(len(hv), hc_n_per, sum(hv),
                                                   hc_per)

    fs = charttools.getFigSize("11.5cm")
    fs = fs[0], fs[1] + 1
    fig = plt.figure(figsize=fs)
    gs = gridspec.GridSpec(2, 1, height_ratios=[1, 3])
    ax0 = plt.subplot(gs[0])
    charttools.boxplot(ax0, v)
    ax0.axvline(cutoff)
    plt.setp(ax0.get_xticklabels(), visible=False)
    plt.setp(ax0.get_yticklabels(), visible=False)

    ax1 = plt.subplot(gs[1], sharex=ax0)
    ax1.hist(v, 50, color=charttools.cs_van_gogh[0])
    ax1.set_xlim(-5, 305)
    ax1.axvline(sci_info["highly_cited_cutoff"])
    ax1.text(50, 1000, text)
    plt.tight_layout(h_pad=0.2)
    fig.subplots_adjust(0.12, 0.08, 0.94, 0.97)

    return fig


figureCitedDistro.title = "Distribucija citata po citiranim člancima i visiko citirani radovi"


def figureCitingJournalsDistroArt10(cit, sci_groups, **kwargs):
    a10 = sci_groups['primary']
    jous = defaultdict(int)
    for cid, item in cit.items():
        cites = set(item["cites"]) & a10
        if cites:
            jous[item["publication_title"]] += len(cites)
    del (jous['Scientometrics'])

    cit_dist = list(jous.values())
    qr = tools.quantileReport(cit_dist)
    tmo = qr["top extreme whisker"]
    top = []
    bottom = []
    for n in cit_dist:
        if n > tmo:
            top.append(n)
        else:
            bottom.append(n)

    tot_jou = len(cit_dist)
    tot_cit = sum(cit_dist)
    n_top = len(top)
    cit_top = sum(top)
    cit_top_per = round(cit_top / tot_cit * 100, 2)
    n_top_per = round(n_top / tot_jou * 100, 2)
    n_bottom = len(bottom)
    cit_bottom = sum(bottom)
    cit_bottom_per = round(cit_bottom / tot_cit * 100, 2)
    n_bottom_per = round(n_bottom / tot_jou * 100, 2)
```

```
    top_text = "Visoko citirajući: {} ({}%) ; {} ({}%)".format(
        n_top, n_top_per, cit_top, cit_top_per)
    bottom_text = "Nisko citirajući: {} ({}%) ; {} ({}%)".format(
        n_bottom, n_bottom_per, cit_bottom,cit_bottom_per)

    all_text = "Citirajući časopisi: n čas={} ; n cit={}".format(tot_jou,
                                                                 tot_cit)

    fs = charttools.getFigSize("11.5cm")
    fs = (fs[0], fs[1] + 1.5)
    fig, axes = plt.subplots(3, 1, figsize=fs)

    axes[0].boxplot(cit_dist, vert=False, whis=3)
    axes[0].set_title(all_text)
    axes[0].set_yticklabels('')

    axes[1].boxplot(top, vert=False, whis=3)
    axes[1].set_title(top_text)
    axes[1].set_yticklabels('')

    axes[2].boxplot(cit_dist, 0, '', vert=False, whis=3)
    axes[2].set_title(bottom_text)
    axes[2].set_yticklabels('')

    fig.subplots_adjust(0.04, 0.1, 0.94, 0.9, 0.2, 0.6)

    return fig


figureCitingJournalsDistroArt10.title = "Distribucija citata na članke u " \
                                        "časopisu *Scientometrics* po citirajućim časopisima i
visoko citirajući časopisi"

cit_chart_spec = [
    (figureCitedDistro, data_paths.f_cit_distribution_art10),
    (figureCitingJournalsDistroArt10,
     data_paths.f_citing_jou_distribution_art10),
]

import igraph

cdict = {
    "methods": charttools.cs_van_gogh[1],
    "theory": charttools.cs_van_gogh[2],
    "applied": charttools.cs_van_gogh[0],
}


def getCocitPapers(cit, dois):
    cocit = defaultdict(set)
    # get all cocited pairs and items that cocited them for a group of dois
    for id, item in cit.items():
        cites = set(item['cites']) & dois
        if len(cites) > 1:
            for pair in combinations(cites, 2):
                cocit[frozenset(pair)].add(id)
    return cocit


def getGraphCocitPapers(sci, cit, dois):
    cocit = getCocitPapers(cit, dois)
    nodes = []
    edges = []
    for id in dois:
        item = sci[id]
        node = {
            "name": id,
            "weight": item["citation_count"],
            "size": item["citation_count"] / 4 + 10,
            "color": cdict[item["prim_focus"][0]]
        }
        nodes.append(node)
    for s, items in cocit.items():
        if len(items) < 10:
            continue
        e = list(s)
        edge = {
```

```
            "source": e[0],
            "target": e[1],
            "weight": len(items),
            "width": len(items) / 4 + 1,
            "color": "black"
        }
        edges.append(edge)
    return igraph.Graph.DictList(nodes, edges)


def graphCocitHighlyCitedArts10(sci, cit, sci_groups, **kwargs):
    g = getGraphCocitPapers(sci, cit, sci_groups['highly_cited_art10'])
    aut_coords = tools.loadJson(data_paths.fname_highly_cit_art_layout)
    lf = [aut_coords[n["name"]] for n in g.vs]
    igraph.plot(g,
                data_paths.f_graph_highly_cited_art10,
                (2000, 2000),
                layout=lf
    )


graphCocitHighlyCitedArts10.title = "Mreža kocitata visoko citiranih članaka objavljenih u
časopisu *Scientometrics* koji su kocitirani barem 10 puta"

cit_graph_spec = [
    (graphCocitHighlyCitedArts10, data_paths.f_graph_highly_cited_art10)
]
```

ŽIVOTOPIS

Krešimir Zauder rođen je 1980-e godine u Zagrebu. 1998 godine završava Tehničku školu Ruđera Boškovića sa zvanjem tehničara za naočalnu optiku. 2006. godine završava diplomski studij informacijskih znanosti smjer bibliotekarstvo, i engleskog jezika i književnosti. Diplomski rad "Organizacija znanja u elektroničkoj sredini" prikazuje neke od njegovih glavnih interesa.

2008. godine zapošljava se u Nacionalnoj i Sveučilišnoj knjižnici u Zagrebu kao znanstveni novak na projekt dr. sc. Maje Jokić "Izrada modela vrednovanja znanstvenog rada u RH za sva znanstvena područja". 2012. godine zajedno s projektom prelazi u Institut za društvena istraživanja u Zagrebu gdje radi i danas.

Njegov glavni interes je računalna obrada podataka (u društvenim znanostima) i to posebno algoritamskim pristupom u kojem smislu Krešimir često koristi programski jezik Python za koji je stručnjak, a posjeduje i velik broj drugih računalnih vještina.

BIBLIOGRAFIJA ZNANSTVENIH RADOVA

Knjige

1. Jokić, Maja; Zauder, Krešimir; Letina, Srebrenka. *Karakteristike hrvatske nacionalne i međunarodne znanstvene produkcije u društveno-humanističkim znanostima i umjetničkom području za razdoblje 1991-2005*. Zagreb : Institut za društvena istraživanja u Zagrebu, 2012.

Radovi u časopisima

1. Jokić, Maja; Zauder, Krešimir. Bibliometrijska analiza časopisa Sociologija sela : Sociologija i prostor u razdoblju 1963.- 2012. // *Sociologija i prostor*. 51 (2013) , 196(2); 331-349.
2. Letina, Srebrenka; Zauder, Krešimir; Jokić, Maja. Produktivnost hrvatskih psihologa: Scientometrijska analiza mreže suradnji na radovima indeksiranim u bazi WoS 1991-2010. // *Suvremena Psihologija*. 15 (2012) , 1; 97-117.
3. Jokić, Maja; Zauder, Krešimir; Letina, Srebrenka. Croatian scholarly productivity 1991–2005 measured by journals indexed in Web of Science. // *Scientometrics*. 83 (2010) , 2; 375-395.

Znanstveni radovi u zbornicima skupova s međunarodnom recenzijom

1. Zauder, Krešimir; Pečarić, Đilda; Tuđman, Miroslav. Sources for Scientific Frustrations: Productivity and Citation Data // *Central European Conference on Information and Intelligent Systems* / Tihomir Hunjak, Sandra Lovrenčić, Igor Tomičić (ur.). Varaždin : Faculty of Organization and Informatics, 2011. 235-241.
2. Jokić, Maja; Zauder, Krešimir. *Nabava/dostupnost znanstvene i stručne literature – stanje u sveučilišnim knjižnicama s naglaskom na fakultetske knjižnice Sveučilišta u Zagrebu*. *Zbornik radova s 10. okruglog stola Slobodan pristup informacijama* / Belan-Šimić, Alemka; Horvat Aleksandra, Hrvatsko knjižničarsko društvo, Zagreb, 2011, str. 25-31.
3. Machala, Lobel; Zauder, Krešimir. Catalogue 2.0 and Bibliography 2.0: Collaboratively created structured resource lists and their aggregation // *Zbornik radova 2. međunarodne znanstvene konferencije "The Future of Information Sciences: INFuture2009 – Digital Resources and Knowledge Sharing"* / Seljan, Sanja ; Stančić, Hrvoje (ur.). Zagreb : Odsjek za informacijske znanosti, Filozofski fakultet, 2009.
4. Zauder, Krešimir; Lasić-Lazić, Jadranka; Banek Zorica, Mihaela. Collaborative Tagging Supported Knowledge Discovery *// Information Technology Interfaces* / Luzar-Stiffler, Vesna. Huljuz Dobric, V. (ur.). Zagreb : Srce, 2007. 437-442.
5. Banek Zorica, Mihaela; Špiranec, Sonja; Zauder, Krešimir. Collaborative Tagging: Providing User Created Organizational Structure for Web 2.0 // *Zbornik radova 1. međunarodne znanstvene konferencije "The Future of Information Sciences : INFuture2007 – Digital Information and Heritage"* / Seljan, Sanja ; Stančić, Hrvoje (ur.). Zagreb : Odsjek za informacijske znanosti, Filozofski fakultet, 2007. 193-203.
6. Banek Zorica, Mihaela; Špiranec, Sonja; Zauder, Krešimir. Where are the library and information professionals in e-learning *// Innovation for the European Era* / Epelboin, Yves. Desnos, Jean-François (ur.). Grenoble : EUNIS, 2007.

Znanstveni radovi u zbornicima skupova

1. Zauder, Krešimir. Web 2.0: mrežno suradničko elektroničko okruženje // *Mogućnost suradnje u okruženju globalne informacijske infrastrukture* / Willer, Mirna (ur.). Zagreb : Hrvatsko Knjižničarsko društvo, 2008. 43-49.