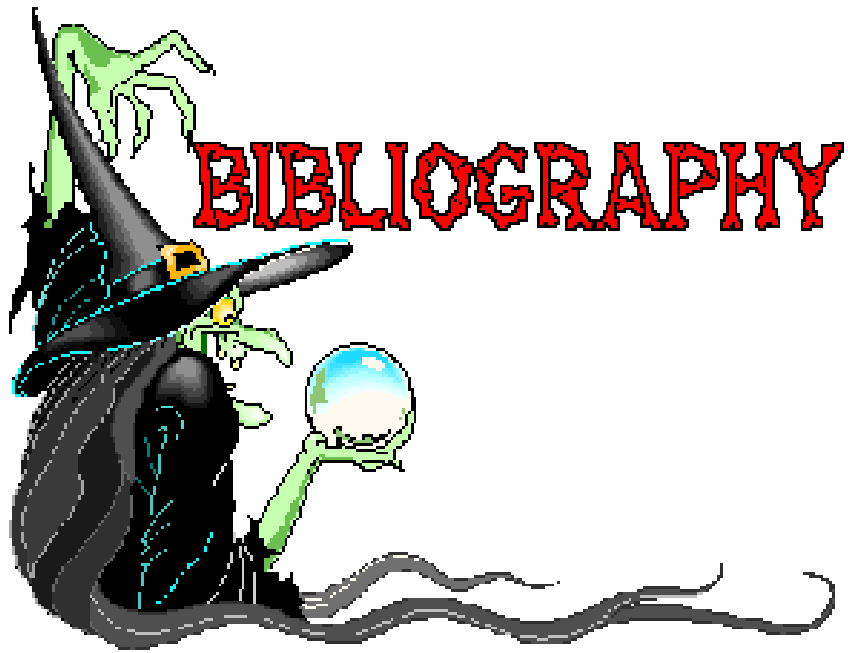
Prepoznavanje enot v bibliografskih podatkih

Vladimir Batagelj
UP FAMNIT Koper in IMFM Ljubljana

1336. in 1337. sredin seminar
Ljubljana, 26. julij in 2. avgust 2023

Kazalo



Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Tekoča različica prosojnic (August 2, 2023 at 04 : 49): PDF
https://github.com/bavla/biblio/

UP FAMNIT, ULj FF, IMFM; triletni (oktober 2022-2025)

1. Poiskati zanimive primere bibliografskih storitev višje stopnje za razne vrste uporabnikov. Razvoj nekaj prototipnih rešitev.
2. Razvoj metod in algoritmov za kakovostno prepoznavanje bibliografskih enot (na osnovi analize bibliografskih omrežij). Ti so osnova za pridobivanje visokokakovostnih bibliografskih podatkov za nadaljnje analize.
3. Nadaljnji razvoj metodologij in algoritmov za analizo bibliografskih omrežij, ki temeljijo na naših preteklih raziskavah (dvovrstna omrežja, deležni pristop, časovna omrežja in časovne količine).

Dosedanje delo na analizi bibliografskih podatkov

Bibliografska podatkovja so bogat vir zanimivih omrežij.

1. SPC 1991, Erdos 2000 [7, 11, 20]
2. Dva članka za IS 2002: omrežja iz besedil [13, 14]
3. Normalizacije: Slovenski časopisi [5], Reuters 11 september [12]
4. Matjaž Zaveršnik, otoki SOM, Geom [21], SPC Patenti [4]
5. Amazon [15]
6. Social networks WoS2Pajek, Viszards [24]
7. FDV, analiza slovenske znanosti [22]
8. Doktorski (Cerinšek, Bodlaj, Praprotnik), projekt GReGAS [8, 17, 18, 16], Španci
9. COST Peere [9]
10. Daša [25, 27, 6, 10, 26, 28]
11. Nataša [29]

Viri bibliografskih podatkov

Gradivo o pripravi bibliografskih podatkov za analize se mi je nabiralo postopoma ob reševanju sprotnih težav. Prvič sem ga uredil za predavanje za doktorske študente v Uppsali leta 2016. Izpopolnjeni različici sva pripravila skupaj z Dašo (Daria Maltseva) za delavnico "Analysis of bibliographic networks" na konferenci NetGlow (Networks in the Global World) v St. Petersburgu, 4-6. julija 2018 in za 10. mednarodno poletno šolo "Analysis of Scientific Networks" v Voronovem pri Moskvi, 15-21. julija 2019.

Bibliografska omrežja

Do bibliografskih podatkov lahko pridemo na razna načine. Na primer iz:

- seznama virov v preglednem članku ali knjigi (ReM, MCO, CoD)
- področnih bibliografij (BibTEX)
- bibliografskih storitev: Web of Science, Scopus, SICRIS, CiteSeer, Zentralblatt MATH, Google Scholar, Crossref, DBLP Bibliography, US patent office, IMDb, in drugih.

Iz njih lahko pridobimo različna dvovrstna omrežja. Ponavadi omrežji dela $\times$ avtorji (**WA**) in dela $\times$ gesla (**WK**).

Pri opisno bogatejših virih dobimo vsaj še omrežje dela $\times$ kode (**WC**) – klasifikacija (npr. MOS) in enovrstno omrežje sklicevanj

delo $\times$ delo (**Ci**). Pri tem dela vsebujejo članke, knjige, poročila, patente, itd. Pristop je uporaben tudi za filme in glasbo.

Poleg tega dobimo še vsaj razbitje po revijah ali založbah in po letih objave ter vektor števila strani.

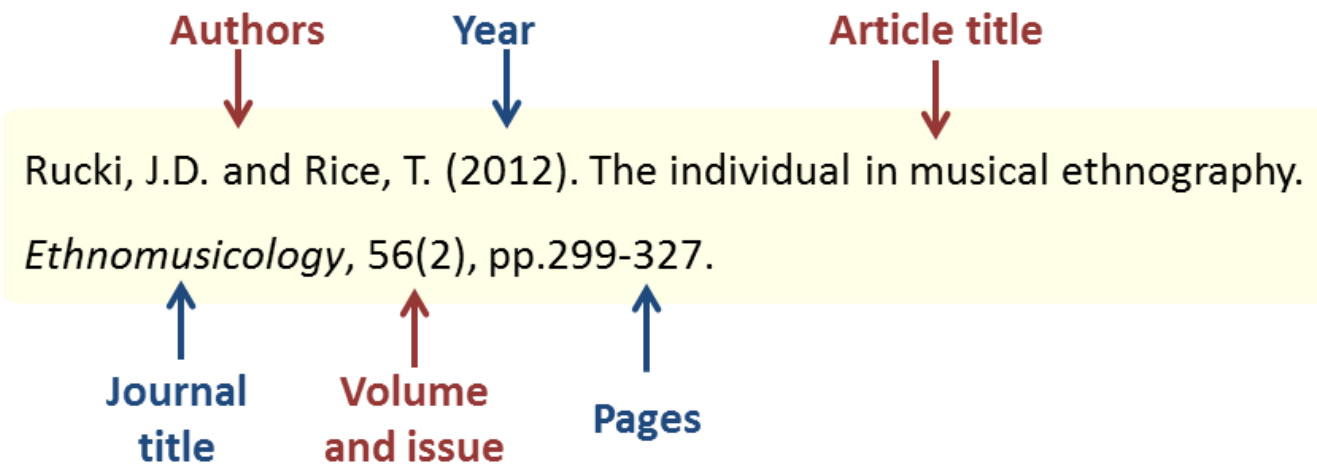
Enote v bibliografskih podatkih

O enotah v bibliografskih podatkih so začeli razmišljati že zgodaj. Med brskanjem po spletu sem naletel na zanimiv vir iz leta 1967 [19].

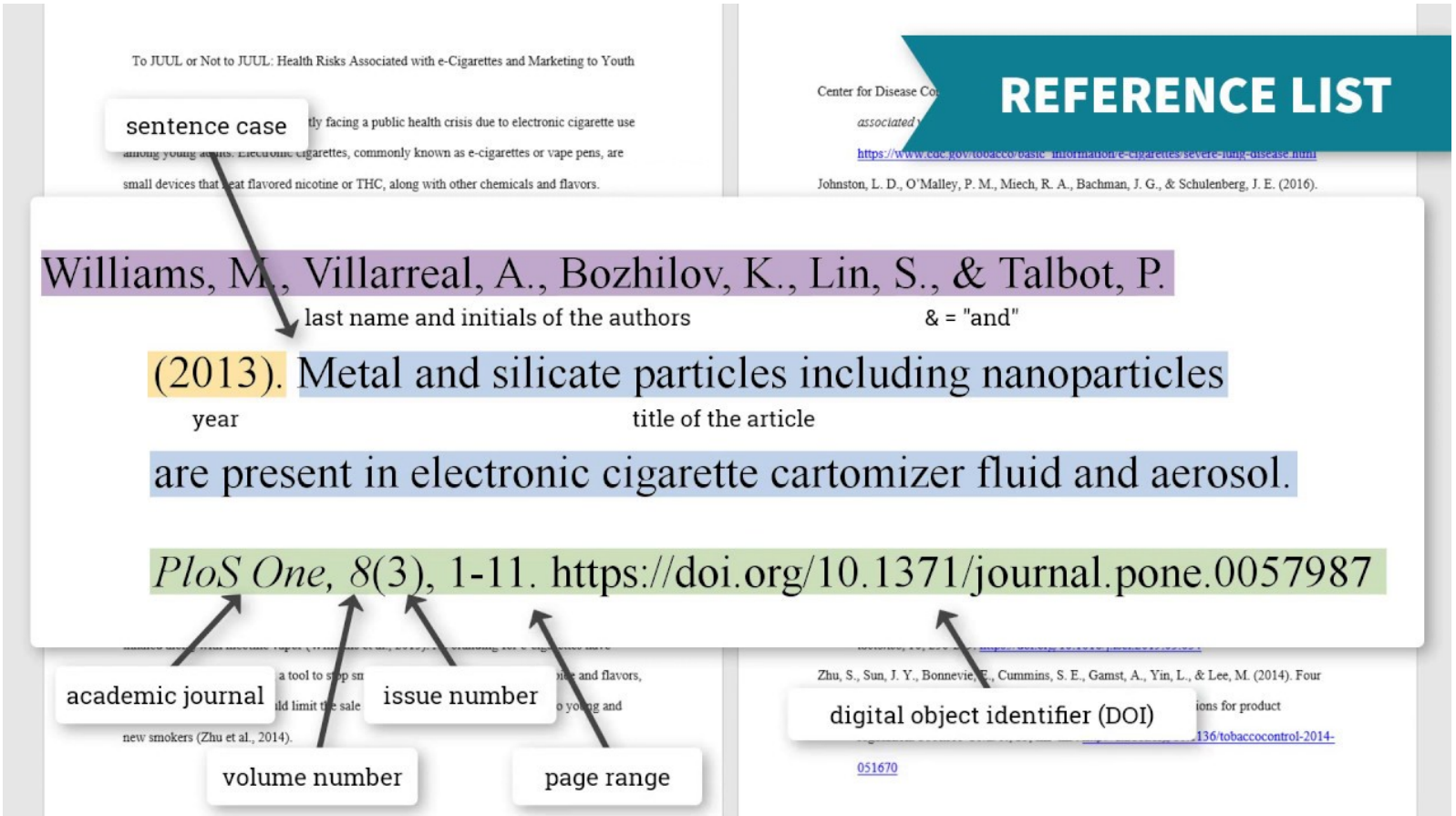
Opis dela v zbirki bibliografskih podatkov je praviloma veliko popolnejši kot je opis dela v seznamu virov.

To poraja težave pri pripravi podatkov za analize, pri katerih želimo upoštevati tudi sklicevanja med deli (omrežje **Ci**).

Opisi del v virih



. . . Opisi del v virih



Slog APA

Book:

Author, A.A. (Year of Publication). *Title of work: Capital letter also for subtitle.* Location: Publisher.

Journal Article from a Database (without a doi):

Author, A.A., & Author, B.B. (Year of Publication). Title of article. *Title of Journal, volume number* (issue number), page range. Retrieved from http://www.someaddress.com/full/url/

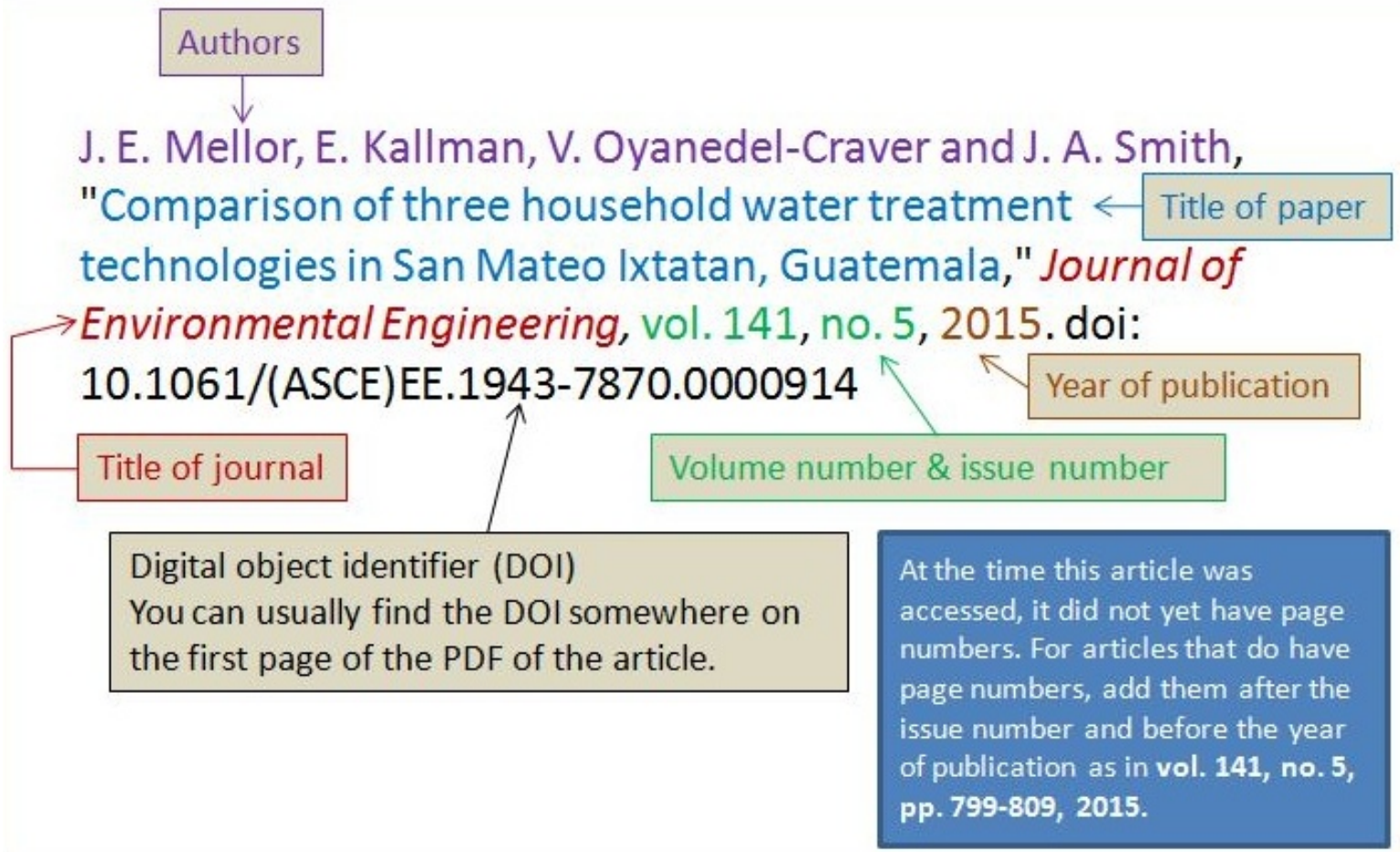
Journal Article from a Database (with a doi):

Author, A.A., & Author, B.B. (Year of Publication). Title of article. *Title of Journal, volume number* (issue number), page range. doi: 000000/00000000000

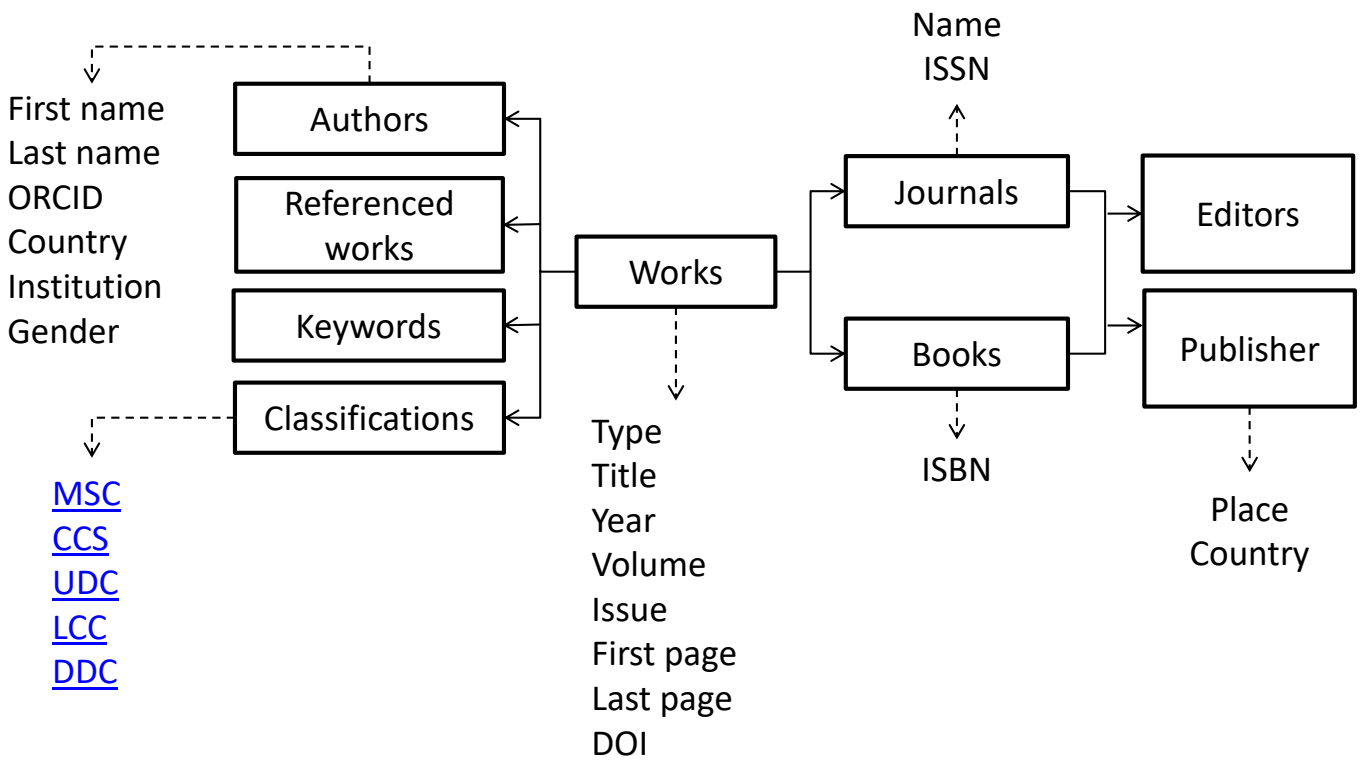
Non-periodical Web Document, Web Page, or Report (Website)

Author, A.A. & Author, B.B. (Date of publication). *Title of document.* Retrieved from http://www.someaddress.com/full/url/

Slog IEEE

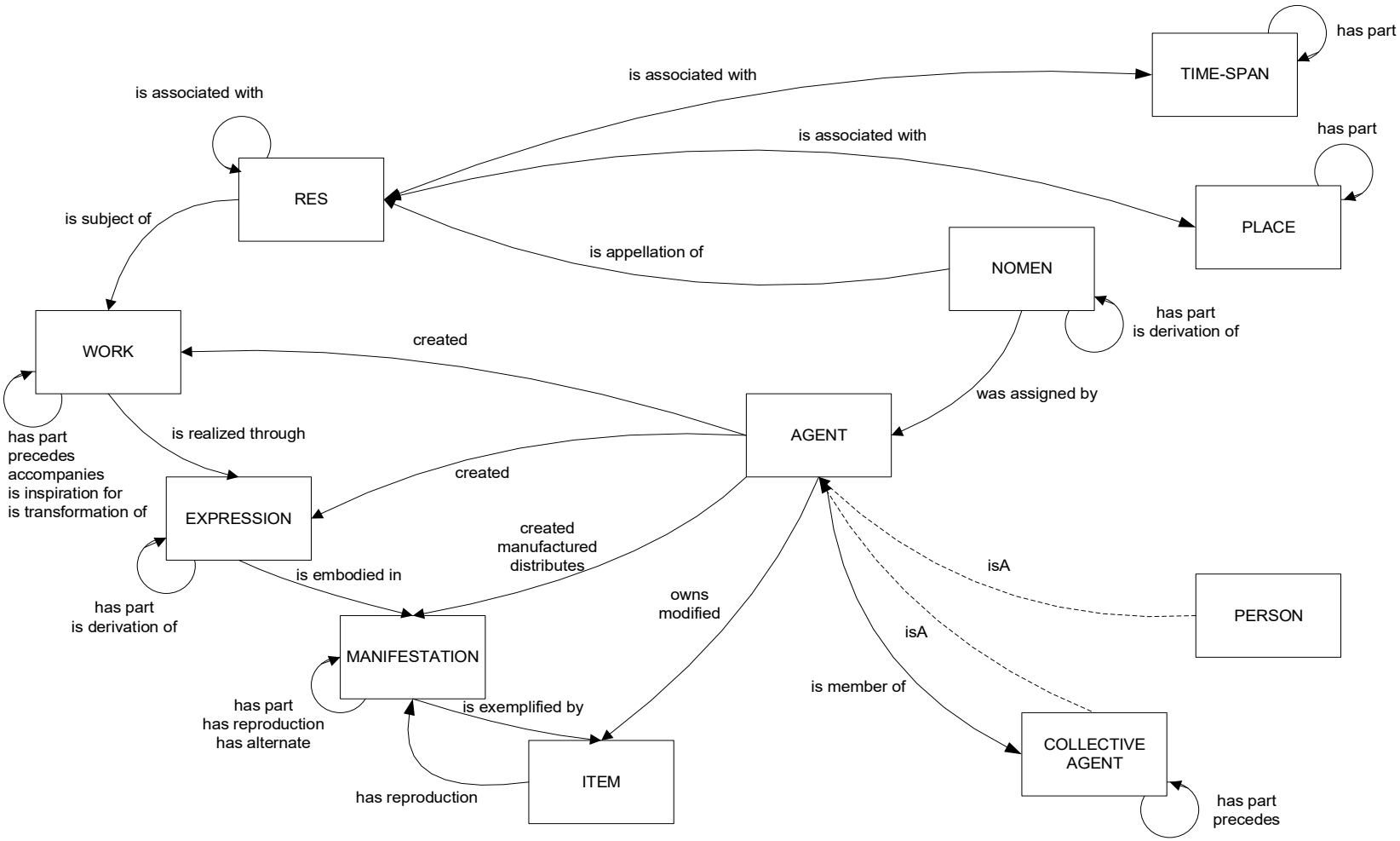

Authors
J. E. Mellor, E. Kallman, V. Oyanedel-Craver and J. A. Smith,
"Comparison of three household water treatment
technologies in San Mateo Ixtatan, Guatemala," Journal of
Environmental Engineering, vol. 141, no. 5, 2015. doi:
10.1061/(ASCE)EE.1943-7870.0000914
Title of paper
Year of publication
Volume number & issue number
Title of journal
Digital object identifier (DOI)
You can usually find the DOI somewhere on the first page of the PDF of the article.
At the time this article was accessed, it did not yet have page numbers. For articles that do have page numbers, add them after the issue number and before the year of publication as in vol. 141, no. 5, pp. 799-809, 2015.

Vrste enot



Nekatere lastnosti se spreminjajo skozi čas. Dodatne lastnosti: e-mail, URL, itd. Seznam lahko dopolnimo z deli drugih vrst (video, glasba, slike, podatki, programi, spletna gradiva, itd.).

Osnovne vrste enot po IFLA-LRM [31]



Zapis v BiBTEXu


```
@Article{int:Mizuno1,
  author =        "S. Mizuno",
  title =         "An ${O(n^{3}L)}$ algorithm using a sequence for
                  linear complementarity problems",
  journal =       "Journal of the Operations Research Society of Japan",
  volume =        "33",
  year =          "1990",
  pages =         "66--75",
}
@InCollection{int:Vorst1,
  author =        "{J. G. G. van de} Vorst",
  title =         "An attempt to use parallel computing in large scale
                  optimisation",
  booktitle =     "Logistics, Where Ends Have to Meet~: Proceedings of
                  the Shell Conference on Logistics in Apeldoorn, The
                  Netherlands, November 1988",
  editor =        "{C. F. H. van} Rijn",
  year =          "1989",
  pages =         "112--119",
  publisher =     "Pergamon Press",
  address =       "Oxford, United Kingdom",
}
```

Bib2Pajek.py

Zapis iz DBLP


```
<article mdate="2004-01-15" key="journals/arscom/BeinekeGL97">
<author>Lowell W. Beineke</author>
<author>Wayne Goddard</author>
<author>Marc J. Lipman</author>
<title>Graphs with Maximum Edge-Integrity.</title>
<year>1997</year>
<volume>46</volume>
<journal>Ars Comb.</journal>
<url>db/journals/arscom/arscom46.html#BeinekeGL97</url>
</article>

<inproceedings mdate="2004-12-09" key="conf/sigcse/BermanD96">
<author>A. Michael Berman</author>
<author>Robert C. Duvall</author>
<title>Thinking about binary trees in an object-oriented world.</title>
<pages>185-189</pages>
<year>1996</year>
<crossref>conf/sigcse/1996</crossref>
<booktitle>SIGCSE</booktitle>
<ee>http://doi.acm.org/10.1145/236536</ee>
<url>db/conf/sigcse/sigcse1996.html#BermanD96</url>
</inproceedings>
```

DBLP2Pajek.py

Zapis iz zbMATH


```
an  00549739
ai  gross.mark-d
is  ISSN 0025-5874; ISSN 1432-1823
au  Gross, Mark
py  1993
cc  *14M15 14J15
ti  Surfaces of bidegree $(3,n)$ in $\text{Gr}(1,\bbfP\sp 3)$.
ut  congruence; family of lines
so  Math. Z. 212, No.1, 73-106 (1993).
an  01488230
ai  tiras.yuecel; harmanci.abdullah; -
is  ISSN 0092-7872; ISSN 1532-4125
au  T{\i}ra\D{s}, Y\"ucel; Harmanc{\i}, Abdullah; Smith, P.F.
py  2000
cc  *13A15 13C05
ti  Some remarks on dense submodules of multiplication modules.
ut  multiplication module; dense submodule
so  Commun. Algebra 28, No.5, 2291-2296 (2000).
se  00000057  Communications in Algebra  Commun. Algebra 0092-7872; 1532
```

ZBml.py

Zapis iz Web of Science


```
PT J
AU Dipple, H
   Evans, B
TI The Leicestershire Huntington's disease support group: a social network
   analysis
SO HEALTH & SOCIAL CARE IN THE COMMUNITY
LA English
DT Article
C1 Rehabil Serv, Troon Way Business Ctr, Leicester LE4 9HA, Leics, England.
RP Dipple, H, Rehabil Serv, Troon Way Business Ctr, Sandringham
   Suite,Humberstone Lane, Leicester LE4 9HA, Leics, England.
CR BORGATTI SP, 1992, UCINET 4 VERSION 1 0
   FOLSTEIN S, 1989, HUNTINGTONS DIS DISO
   SCOTT J, 1991, SOCIAL NETWORK ANAL
NR 3
TC 3
PU BLACKWELL SCIENCE LTD
PI OXFORD
PA P O BOX 88, OSNEY MEAD, OXFORD OX2 ONE, OXON, ENGLAND
SN 0966-0410
J9 HEALTH SOC CARE COMMUNITY
JI Health Soc. Care Community
PD JUL
PY 1998
VL 6
IS 4
BP 286
EP 289
PG 4
SC Public, Environmental & Occupational Health; Social Work
GA 105UP
UT ISI:000075092200008
ER
```

WoS2Pajek

Pretvorbe opisov

RIS → WoS


```
ris <- readLines("Additional_RIS_Data.txt")
T <- gsub("^JO","SO",
     gsub("^AD","PI",
     gsub("^SP","BP",
     gsub("^A1","AU",
     gsub("^T1","TI",
     gsub("^Y1","PY",
     gsub("^PB","PU",
     gsub("^KW","DE",
     gsub("^TY (\\S)(\\S*)","PT \\1",
     gsub(" -","",ris[nchar(ris)>0]))))))))))
Encoding(T) <- "UTF-8"
writeLines(T,"AdditionalRIS.WoS")
```

X-format → WoS-format → (WoS2Pajek) → Pajek-ova omrežja

Problemi pri pripravi bibliografskih podatkov

Večina izvornih bibliografskih podatkov je polstrukturiranih — na voljo so v obliki zapisov iz neke baze podatkov. Izbrana polja v zapisu predstavljajo različne enote: imena ljudi, imena revij, ključne besede, ID del, države, ustanove . . . Te enote določajo množice vozlišč bibliografskih omrežij. Natančno prepoznavanje posameznih enot je ključnega pomena za izgradnjo kakovostnih omrežij.

Na žalost imena teh enot običajno niso shranjena v standardizirani obliki in se nekatera pojavljajo v seznamih. Pri prepoznavanju posameznih enot se pojavita dve težavi

Sinonimija / sopomenke: Različni imeni določata isto enoto.
Homonimija / dvoumnost: Več enot z enakim imenom.

Problem 1. opisi del

- natančnost opisa (seznam upoštevanih lastnosti del)
- popolnost opisa (ali so vnešeni vsi podatki?)
- **Slogi opisa virov**: obstaja več slogov podprtih od raznih strokovnih združenj: MLA (Modern Language Association of America), APA (American Psychological Association), Chicago (University of Chicago Press), AMA (American Medical Association), SCE (Council of Science Editors), GOST (Ruski državni standard) / AMSBIB.
- večina bibliografskih podatkovnih baz podpira vsaj en znakovni način izpisa izbranih zapisov. Pri izpopolnjevanju podatkovja, ki ga nameravamo uporabiti v analizah, je večkrat potrebno pretvarjati med različnimi oblikami zapisov.

. . . Problem 1. opisi del

članek

APA:

White, H. (2008). Identity and Control: How Social Formations Emerge (Second Edition). PRINCETON; OXFORD: Princeton University Press.

MLA:

White, Harrison C. Identity and Control: How Social Formations Emerge (Second Edition). Princeton University Press, 2008.

Chicago:

White, Harrison C. Identity and Control: How Social Formations Emerge (Second Edition). PRINCETON; OXFORD: Princeton University Press, 2008.

GOST:

White H. C. Identity and control. { Princeton University Press, 2008.

. . . Problem 1. opisi del

Graph products

Opis del v seznamu virov iz knjige ali preglednega članka običajno vsebuje naslednje sestavine:

1. Imena avtorjev; večkrat nepopolna (et al., samo začetnice imen)

   `WASSERMAN S, 1994, SOCIAL NETWORK ANAL`

   za

   `Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8). Cambridge university pres`

2. Naslov
3. Leto ali datum objave

za članke še:

1. Revija
2. Knjiga? Letnik?
3. Zvezek
4. Strani

in za knjige: (izdaja,) izdajatelj (podjetje, kraj)

. . . Problem 1. opisi del

Bibliographic data formats:
BibTex (LaTeX style):

```
@book{white2012identity,
  title={Identity and control},
  author={White, Harrison C},
  year={2012},
  publisher={Princeton University Press}
}
```

EndNote (Clarivate Analytics):

```
%0 Book
%T Identity and control
%A White, Harrison C
%D 2012
%I Princeton University Press
```

. . . Problem 1. opisi del

RIS (Research Information Systems style):

```
TY  - BOOK
TI  - Identity and Control
AU  - White, Harrison C.
AB  - <p>In this completely revised edition ...</p>
PB  - Princeton University Press
PY  - 2008
SN  - 9780691137155
T1  - How Social Formations Emerge (Second Edition)
UR  - http://www.jstor.org/stable/j.ctt1r2fg1
ER  -
```

Problem 2: Različne kulture

na posameznih področjih

1. število soavtorjev; doktorand mora imeti samostojne objave, credit, Higgs boson, single-top-quark
2. zapis imen avtorjev (začetnice/polna imena)
3. vrstni red sestavin avtorjevega imena (Francozi, Španci, Arabci, Rusi, itd., dodatki de, van, ibn, Prof, Dr, itd.).
4. nekatere revije imajo posebna pravila o krajšanju imen revij

```
Bon G., 1896, CROWD STUDY POPULAR
Le Bon G, 1897, CROWD STUDY POPULAR
LeBon  G., 1960, CROWD STUDY POPULAR
Lebon G., 2011, PSIHOLOGIJA NARODOV
Le Bon Gustave, 1930, CROWD STUDY POPULAR
Gustave Le Bon, 1982, PSYCHOL MASSEN
```

```
Newman, M. E. (2001). Scientific collaboration networks.
II. Shortest paths, weighted networks, and centrality.
Physical review E, 64(1), 016132.
```

```
M.E.J. Newman, preceding paper, Phys. Rev. E 64, 016131 (2001).
```

Problem 2: Različne kulture

Različni opisi virov

1. Običajni opis AU (PY). TI. JI, BP-EP

   ```
   Freeman, L. C., & White, D. R. (1993). Using Galois lattices to
   represent network data. Sociological methodology, 127-146.
   ```

2. Številka članka AU (PY). TI. JI, VL(IS), BP

   ```
   Newman, M. E. (2001). Scientific collaboration networks.
   II. Shortest paths, weighted networks, and centrality.
   Physical review E, 64(1), 016132.
   ```

3. Brez naslova članka AU, JI VL, IS (PY)

   ```
   P. Erdos and A. Renyi, Publ. Math. Inst. Hung. Acad. Sci.
   5, 17 (1960).
   ```

Problem 3: Zapis podatkov

Velike/male črke – pretvorba v izbrano obliko.

ASCII / Unicode

Cyrillic to Latin (Unicode, programska pretvorba v ASCII)
Prevedba je odvisna od jezika, kateremu se ime prilagaja. ISO 9

```
Пётр Ильич Чайковский

angleško: Pyotr Ilyich Tchaikovsky
nemško: Pjotr Iljitsch Tschaikowski
francosko: Piotr Ilitch Tchaïkovski
španski: Piotr Ilich Chaikovski
italijansko: Pëtr Il'ič Čajkovskij
slovensko: Peter Iljič Čajkovski
```

Russian

Problem 4. Nepopolni podatki

- **m**anjkajoči podatki: nekatere sestavine opisa manjkajo ali so nepopolni (dodatni avtorji, naslov, letnik, letnik, zvezek, strani itd.). Podatke ročno dopolnimo.
- **M**anjkajoči podatki: manjkajo opis del pomembnih za obravnavano temo, ki v pridobljenem podatkovju niso zajeta (v začetnem obdobju področja se je uporabljalo drugo izrazje). Če imamo omrežje sklicevanj, jih lahko odkrijemo z njegovo analizo in dodamo polne opise našemu podatkovju.

Za male bibliografije, kjer lahko pregledamo, sprejmemo in "popravimo" vsak opis dela zadošča tabelarični pristop (Excel).

V bazi podatkov so lahko tudi napake (tipkarske napake) – popravimo jih v svojem izvodu podatkov.

Problem 5. Prepoznavanje enot

imena avtorjev

Dvoumnost: "Multiple personalities": Harzing, A. W. (2015). Lorenzo Bartolini iz Letters to Juliet.

Obstaja veliko načinov pisanja imena. Nekatere baze bibliografskih podatkov določijo imenu ustreznega avtorja pri vnosu podatkov in mu pripešejo enolično oznako (DBLP, ZB, ResearcherId). "Three Zhang, four Li" Kitajci 100 priimkov. V bazi MathSciNet je vsaj 557 (različnih?) matematikov z imenom Zhang, Li.

V zbMATH se objave avtorja Smith, John W. pojavljajo od leta1868 do leta 2007.

MathSciNet; Orcid - vnesi ime avtorja v iskalno okence; Scopus; eLibrary - klikni na ime avtorja.

```
https://orcid.org/0000-0002-0240-9446
https://elibrary.ru/author_items.asp?authorid=155240
```

Pristop AMS, DBLP.

. . . Problem 5. Prepoznavanje enot

imena avtorjev

V podatkovju zbMATH: Borštnik, N. S. Mankoč; Mankoč Borštnik, N.; Mankoč-Borštnik, Norma; Mankoč Borštnik, Norma Susana; Mankoc-Borstnik, N.S.; Mankoč Borštnik, N.S..

V zbMATH je na primer O'Regan, Donal zapisan kot oregan.donal (kanonsko ime, ZBunified) in drugič o'regan.d. Včasih ni uporabljano kanonsko ime. V teh primerih naš program predela polno ime O'Regan, Donal v o'regan.d.
Podobno velja za avtorja Pečarić, Josip E, ki ima kanonsko ime pecaric.josip-e. Kadar ni uporabljeno njegovo kanonsko ime, dobimo pecaric.j in pecaric.j-e iz dveh polnih imen Pečarić, J. in Pečarić, J. E.

Za določitev različic v osebnih imenih uredimo imena glede na priimke.

Pri avtorjih z imeni zapisanimi v različnih abecedah prevedemo zapise v izbrano abecedo ali uporabimo "slovar" enakovrednih zapisov.

. . . Problem 5. Prepoznavanje enot

imena avtorjev

V BibTEXovi "Computational Geometry Database" (Jones 2002) ima isti avtor sedem različnih imenČ R.S. Drysdale, Robert L. Drysdale, Robert L. Scot Drysdale, R.L. Drysdale, S. Drysdale, R. Drysdale, in R.L.S. Drysdale.

`Krivoshe\u\i n, Leonid Evgen\cprime evich` ima v Mathematical Reviews Database 20 različnih zapisov imena.

Poznavanje področja je potrebno, da vemo, da sta Otfried Schwarzkopf in Otfried Cheong ali Mikhail Efimovich Tylkin in Michel Marie Deza ista oseba.

Vir problemov so tudi zapisi vzhodnoslovanskih imen. Najbrž sta Krachkovskij, A. P. in Krachkovskii, A. P. ista oseba.

`ANONYMO(2015)`

. . . Problem 5. Prepoznavanje enot

imena avtorjev

id	canon	alter1	alter2	alter3	alter4	alter5
ORCID:0000-0002-0240-9446	Batagelj, Vladimir	Batagelj, Vlado	Batagelj, V	MR:32440	Scopus: 56037441100	
Scopus: 35615877200	Batagelj, Valentin	Batagelj, V				
ORCID:0000-0003-4467-7075	Cheong, Otfried	Schwarz kopf, Otfried	Cheong, O	Schwarzkopf, O	Scopus: 57191986875	
MR:57370	Deza, Michel-Marie	Deza, MM	Deza, M	Deza, Mikhail	Scopus: 7003745115	
MR:57370	Deza, Michel-Marie	Tylkin, Mikhail Efimovich	Tylkin, ME	Тылкин, Михаил Ефимович	Тылкин, ME	Деза, Мишел
ORCID:0000-0002-4294-9017	Zweig, Katharina Anna	Zweig, KA	Zweig, K	Lehmann, Katharina Anna	Lehmann, K	Scopus 592816 2000
eLib:696348	Maltseva, Daria	Мальцева, Дарья Васильевна	Мальцева, ДВ	Мальцева, Д	Maltseva, DV	
eLib:155240	Maltseva, Diana	Мальцева, Диана Васильевна	Мальцева, ДВ	Мальцева, Д	Maltseva, DV	

Problem 5. Prepoznavanje enot

imena del

Imena ISI: pojavljajo se v polju CR v WoS

```
LEFKOVITCH LP, 1985, THEOR APPL GENET, V70, P585
```

in imajo naslednjo zgradbo

```
AU + ', ' + PY + ', ' + SO[:20] + ', V' + VL + ', P' + BP
```

V WoSu ima lahko isto delo več imen ISI. Razlike so tudi v zapisih imen.

```
GRANOVET.MS, 1973, AM J SOCIOL, V78, P1360
GRANOVETTER M, 1983, SOCIOLOGICAL THEORY, V1, P203
BORGATTI SP, 2002, UGINET WINDOWS SOFTW
BORGATTI S, 1999, UCINET V USERS GUIDE
CANTANZARO M, 2005, PHYS REV E, V71, UNSP 027103
CANTAZARO M, 2005, PHYS REV E, V71, UNSP 056104
CATANZARO M, 2005, PHYS REV E 2, V71, ARTN 056104
PALLA G, 2005, NATURE, V435, P814, DOI 10.1038/nature03607
```

. . . Problem 5. Prepoznavanje enot

imena del

Nepravilna imena se pojavljajo tudi v polju `AU`:

```
AU BENSON, , C
   KULHAVY, , W
AU SCHONEMA.PH
```

Za povečanje natančnosti smo v programu WoS2Pajek uvedli **kratka imena** v obliki (HistCite Garfield):

```
LastNm[:8] + ' ' + FirstNm[0] + '(' + PY + ')' + VL + ':' + BP
```

Na primer: `LEFKOVIT L(1985)70:585`.

Iz priimkov z VAN, DE, itd. so izločeni presledki. Nenavadnim imenom je spredaj dodan znak $*$ ali $.

V kratkih imenih vzamemo vrednosti `ARTN` in `UNSP` kot vrednost `BP`.

Najenostavneje je nepravilnosti odpraviti v našem izvodu izvornih podatkov.

Problem 5. Prepoznavanje enot

imena revij in knjig

V WoSovih opisih virov najdemo na primer naslednja imena revij:
NUCLEIC ACIDS RES, NUCL ACIDS RES, NUCLEIC ACIDS RES S, NUCLEIC ACIDS RES S2, NUCL ACID RES, NUCL ACIDS RES S2, NUCL ACIDS S SER, NUCL ACIDS RES S, NUCL AC RES, NUCLEIC ACIDS RES S1, Nucleic Acids Res, NUCL ACIDS RES S1
in
Q J R MET SOC, Q J R METEOROL SOC, Q J ROY METEOR SO S1, Q J ROY METEOR SOC, Q J ROY METEOR SOC B, QUART J ROY METEOR S, QUART J ROY METEOROL, QUART J ROY METEOROL SOC, QUART J ROYAL METEOR.

Ali pripadajo eni sami reviji, ali večim?

. . . Problem 5. Prepoznavanje enot

imena revij in knjig

Ustvarimo ključe iz prvih (dveh) črk imena revij in uredimo pare (ključ, revija). Po "čiščenju" seznama revij v podatkovju zbMATH se je število revij zmanjšalo z 3158 na 2665.

Podatki o knjigah. Pomagamo si lahko z International Standard Serial Number ISSN; Digital Object Identifier DOI; International Standard Book Number ISBN.

Pri prepoznavanju revij si lahko pomagamo z Global Serials Directory [3] in Journal Abbreviation Sources [2] ter drugimi viri na spletu.

Naslovi avtorjev nimajo ustaljene zgradbe. Zato WoS2Pajek enot ustanova in država še ne podpira.

. . . Problem 5. Prepoznavanje enot

gesla

Včasih so gesla (ključne besede) navedena v podatkih. Sicer jih lahko poskusimo pridobiti iz besedila (naslov, povzetek). Pri določanju gesel v besedilu izpustimo *nepomembne besede* (stopwords). Sopomenke prepoznamo z geslenjem (lemmatization) ali/in uporabo slovarjev. Večbesedna gesla.

Na primer izrazi 'function', 'map', 'mapping', and 'transformation' so v matematiki enakovredni. Podobno je, kadar imamo opravka z več jeziki. To lahko razrešimo z uporabo slovarjev.

Drug vir sopomenk so slovnična pravila jezika. Tako se na primer v angleščini 'go' pojavlja v besedilu kot 'go', 'goes', 'gone', 'going' in 'went'. Problem razrešujemo z geslenjem, ki ga podpirajo knjižnice za analizo besedil, kot sta NLTK (Bird et al. 2009; Perkins 2010) in MontyLingua (Liu 2004).
Github Bavla/biblio

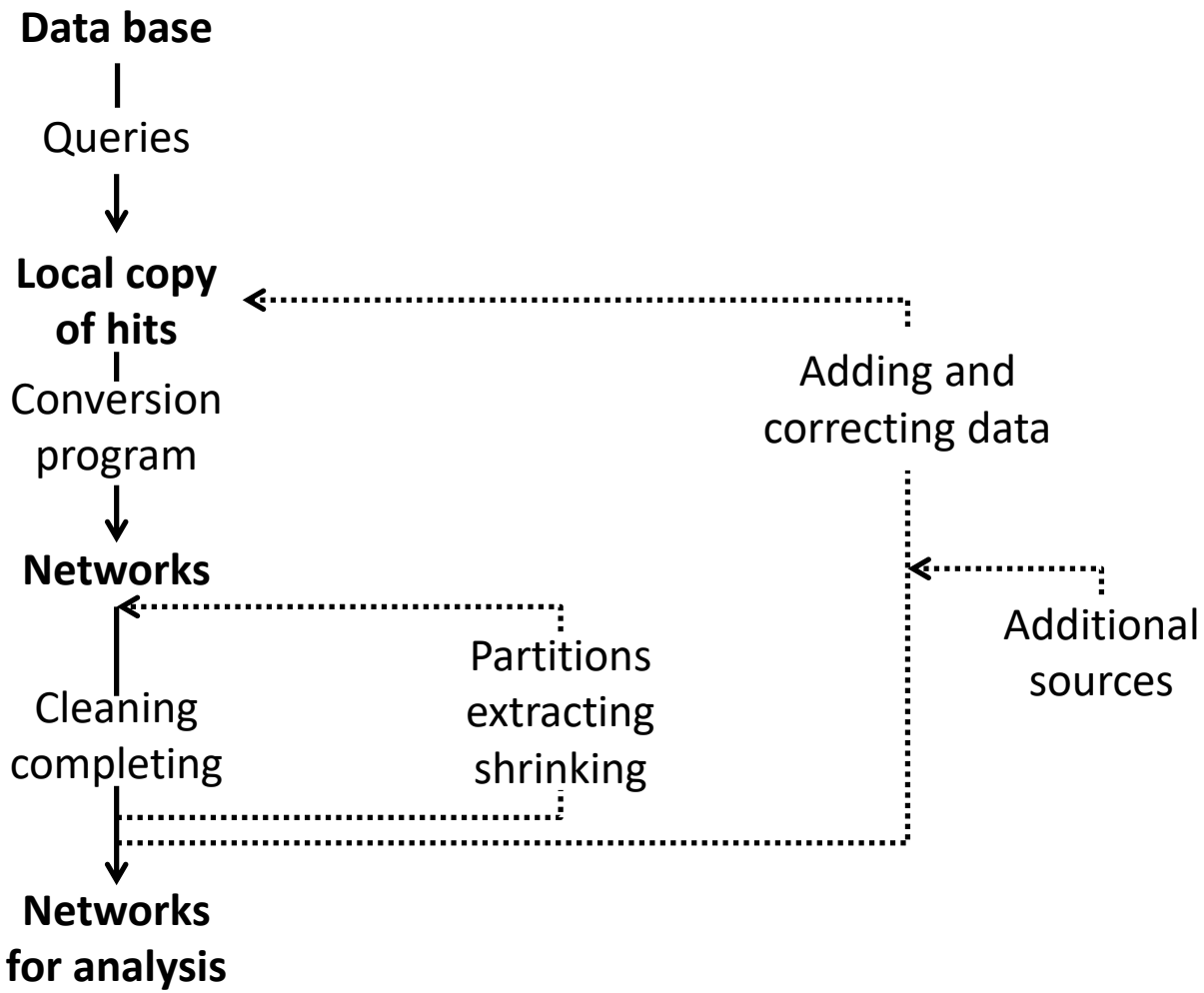
Prečiščevanje in izpopolnjevanje bibliografskih podatkov

Načeloma je večino nepravilnosti mogoče odkriti in popraviti, a je to zelo zamudno opravilo, še posebej za obsežnejša omrežja. Pri tem je treba tehtati med razpoložljivim časom in kakovostjo pridobljenih podatkov.

Vse enote niso enako pomembne pri analizi danega podatkovja. Zato se je izkazala kot sprejemljiva pot, pri kateri nekajkrat ponovimo osnovne analize, ki jih nato upoštevamo pri dodatnem čiščenju podatkov. Neodkrite nepravilnosti na koncu obravnavamo kot šum, ki ne vpliva (naj ne bi vplival) bistveno na izide analiz. Seveda je na koncu vselej potrebno skrbno preveriti rezultate, če se ni kaka nepravilnost kljub vsemu "dvignila na površje".

Pri izpopolnjevanju podatkovja je posebej koristna analiza omrežja sklicevanj. Dela razdelimo na *zadetke* in *preostala*. Za zadetke imamo polne opise, za preostale pa le skop opis vira. To moramo upoštevati pri določanju mej omrežij. Pri analizi med preostalimi deli najdemo taka, na katera je veliko sklicov. Smiselno bi jih bilo spremeniti v zadetke – *izpopolnjevanje*.

. . . Prečiščevanje in izpopolnjevanje bibliografskih podatkov



Zaključki

- Dopolniti program WoS2Pajek z enotami država in (morda) ustanova.
- Za določitev znanstvene discipline dela bi lahko kot prvi približek uporabili razvrstitev revije, v kateri je bilo objavljeno.
- Dopolniti postopek prečiščevanja in izpopolnjevanja s postopki določanja vprašljivih enot in oceno kakovosti prepoznavanja.
- Podpora za podatke iz Scopusa.
- Povezave s semiotiko (Peirce)!?

Podpora

Delo je delno podprla Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije ARIS (raziskovalni program P1-0294 in raziskovalni projekti J5-2557, J1-2481 ter J5-4596) in je bilo pripravljeno v okviru COST action CA21163 (HiTEc).

Viri I

Names of persons: national usages for entry in catalogues. – 4th revised and enlarged edition.
UBCIM publications; new series, vol. 16. München: K.G. Saur, 1996.
URL: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/pubs/names-of-persons_1996.pdf.

Journal abbreviation sources, 2022.
URL: https://www.abbreviations.com/jas.php.

Ulrichsweb, 2022.
URL: http://ulrichsweb.serialssolutions.com/.

Vladimir Batagelj.
Efficient algorithms for citation network analysis, 14 Sep 2003.
URL: https://arxiv.org/abs/cs/0309023,
doi:https://doi.org/10.48550/arXiv.cs/0309023.

Vladimir Batagelj.
Pajek, 2001.
Dagstuhl seminar.
URL:
http://vlado.fmf.uni-lj.si/pub/networks/doc/dagstuhl/kazalo.htm.

Viri II


Vladimir Batagelj.
On fractional approach to analysis of linked networks.
Scientometrics, 123(2):621–633, 2020.
doi:10.1007/s11192-020-03383-y.


Vladimir Batagelj.
Some mathematics of network analysis, January 1991.
Network seminar, Department of Sociology, University of Pittsburgh.
URL:
http://vlado.fmf.uni-lj.si/pub/networks/data./cite/report.pdf.


Vladimir Batagelj and Monika Cerinšek.
On bibliographic networks.
Scientometrics, 96(3):845–864, 2013.
doi:10.1007/s11192-012-0940-1.


Vladimir Batagelj, Anuška Ferligoj, and Flaminio Squazzoni.
The emergence of a field: A network analysis of research on peer review.
Scientometrics, 113(1):503–532, 2017.
doi:10.1007/s11192-017-2522-8.

Viri III


Vladimir Batagelj and Daria Maltseva.
Temporal bibliographic networks.
J. Informetr., 14(1):Article No. 101006, 2020.
doi:{10.1016/j.joi.2020.101006}.


Vladimir Batagelj and Andrej Mrvar.
Some analyses of erdos collaboration graph.
Social Networks, 22:173–186, 2000.


Vladimir Batagelj and Andrej Mrvar.
Density based approaches to network analysis: Analysis of reuters terror news network.
In *Workshop on Link Analysis for Detecting Complex Behavior (LinkKDD2003)*, August 27, 2003.
URL:
http://www.cs.cmu.edu/~dunja/LinkKDD2003/papers/Batagelj.pdf.

Viri IV

Vladimir Batagelj, Andrej Mrvar, and Matjaž Zaveršnik.
Network analysis of dictionaries.
In *Proceedings B of the 5th International Multi-Conference IS'2002 / Language Technologies*. Ljubljana, 2002.
URL: http://nl.ijs.si/isjt02/zbornik/sdjt02-24abatagelj.pdf.

Vladimir Batagelj, Andrej Mrvar, and Matjaž Zaveršnik.
Network analysis of texts.
In *Proceedings B of the 5th International Multi-Conference IS'2002 / Language Technologies*. Ljubljana, 2002.
URL: http://nl.ijs.si/isjt02/zbornik/sdjt02-24bbatagelj.pdf.

Vladimir Batagelj, Kejžar Nataša, and Simona Korenjak-Černe.
Analysis of the customers' choice networks: An application on amazon books and cds data.
Metodološki zvezki/Advances in Methodology and Statistics, 4(2):191–204, 2007.

Viri V


Vladimir Batagelj and Selena Praprotnik.
An algebraic approach to temporal network analysis based on temporal quantities.
Soc. Netw. Anal. Min., 6(1):28:1–28:22, 2016.
doi:10.1007/s13278-016-0330-4.


Jernej Bodlaj and Vladimir Batagelj.
Network analysis of publications on topological indices from the web of science.
Molecular informatics, 33(8):514–535, 2014.
doi:10.1002/minf.201400014.


Monika Cerinšek and Vladimir Batagelj.
Network analysis of Zentralblatt MATH data.
Scientometrics, 102(1):977–1001, 2015.
doi:10.1007/s11192-014-1419-z.


Ann T. Curran and Henriette D. Avram.
The identification of data elements in bibliographic records, 1967.
URL: https://apps.dtic.mil/sti/citations/AD0666447,
doi:AD0666447.

Viri VI

Jerry Grossman.
The Erdös number project, 2022.
URL: https://sites.google.com/oakland.edu/grossman/home/the-erdoes-number-project.

B. Jones.
Computational geometry database, 2002.
URL: ftp://ftp.cs.usask.ca/pub/geometry/.

Luka Kronegger, Franc Mali, Anuška Ferligoj, and Patrick Doreian.
Collaboration structures in Slovenian scientific communities.
Scientometrics, 90(2):631–647, 2012.
doi:10.1007/s11192-011-0493-8.

Viri VII

Michael Ley and Patrick Reuther.
Maintaining an online bibliographical database: The problem of data quality.

In Gilbert Ritschard and Chabane Djeraba, editors, *Extraction et gestion des connaissances (EGC'2006), Actes des sixièmes journées Extraction et Gestion des Connaissances, Lille, France, 17-20 janvier 2006, 2 Volumes*, volume RNTI-E-6 of *Revue des Nouvelles Technologies de l'Information*, pages 5–10. Cépaduès-Éditions, 2006.
URL: http://dblp.uni-trier.de/papers/EGC06_ML_PR.pdf.

Hugo Liu.
Montylingua: A free, commonsense-enriched natural language understander for english (version 2.1), 2004.
URL: http://alumni.media.mit.edu/~hugo/montylingua/.

Daria Maltseva and Vladimir Batagelj.
Social network analysis as a field of invasions: Bibliographic approach to study SNA development.
Scientometrics, 121(2):1085–1128, 2019.
doi:10.1007/s11192-019-03193-x.

Viri VIII


Daria Maltseva and Vladimir Batagelj.
imetrics: the development of the discipline with many names.
Scientometrics, 125(1):313–359, 2020.
doi:10.1007/s11192-020-03604-4.


Daria Maltseva and Vladimir Batagelj.
Towards a systematic description of the field using keywords analysis: Main topics in social networks.
Scientometrics, 123(1):357–382, 2020.
doi:10.1007/s11192-020-03365-0.


Daria Maltseva and Vladimir Batagelj.
Journals publishing social network analysis.
Scientometrics, 126(1):3593–3620, 2021.
doi:10.1007/s11192-021-03889-z.


Nataliya Matveeva, Vladimir Batagelj, and Anuška Ferligoj.
Scientific collaboration of post-soviet countries: the effects of different network normalizations.
Scientometrics, 128(1):4219–4242, 2023.

Viri IX

Bert TePaske-King and Norman Richert.
The identification of authors in the mathematical reviews database.
Issues Sci. Technol. Librariansh., (31), Sep. 2001.
URL: https://journals.library.ualberta.ca/istl/index.php/istl/article/view/1861, doi:10.29173/istl1861.

Maja Žumer.
Ifla library reference model (ifla lrm): Harmonisation of the frbr family.
Knowl. Org., 45(4):310–318, 2018.
doi:10.5771/0943-7444-2018-4-310.