Prepoznavanje enot v bibliografskih podatkih

Vladimir Batagelj
UP FAMNIT Koper in IMFM Ljubljana

1336. sredin seminar
Ljubljana, 26. julij 2023

Outline

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si
Tekoča različica prosojnic (July 26, 2023 at 00:06): PDF
https://github.com/bavla/ibm3m/

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. Erdos 2000 [7]
2. Dva članka za IS 2002: omrežja iz besedil [9, **?**]
3. Normalizacije (Slovenski časopisi, Reuters 11 september) [8]
4. Matjaž Zaveršnik, otoki SOM [12] SPC Patenti [2]
5. Amazon
6. Social networks WoS2Pajek, Viszards [13]
7. FDV, analiza slovenske znanosti
8. Doktorski, EvroProj [4, 10]
9. COST Peere [5]
10. Daša [14, 15, 3, 6]
11. Nataša

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 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.

Networks from bibliographic data

From special bibliographies (BibTEX) and bibliographic services (Web of Science, Scopus, SICRIS, CiteSeer, Zentralblatt MATH, Google Scholar, DBLP Bibliography, US patent office, IMDb, and others) we can derive some two-mode networks on selected topics:

works $\times$ authors (**WA**),
works $\times$ keywords (**WK**);

and from some data also the network

works $\times$ classification (**WC**)

and the one-mode citation network

works $\times$ works (**Ci**);

where works include papers, reports, books, patents etc.

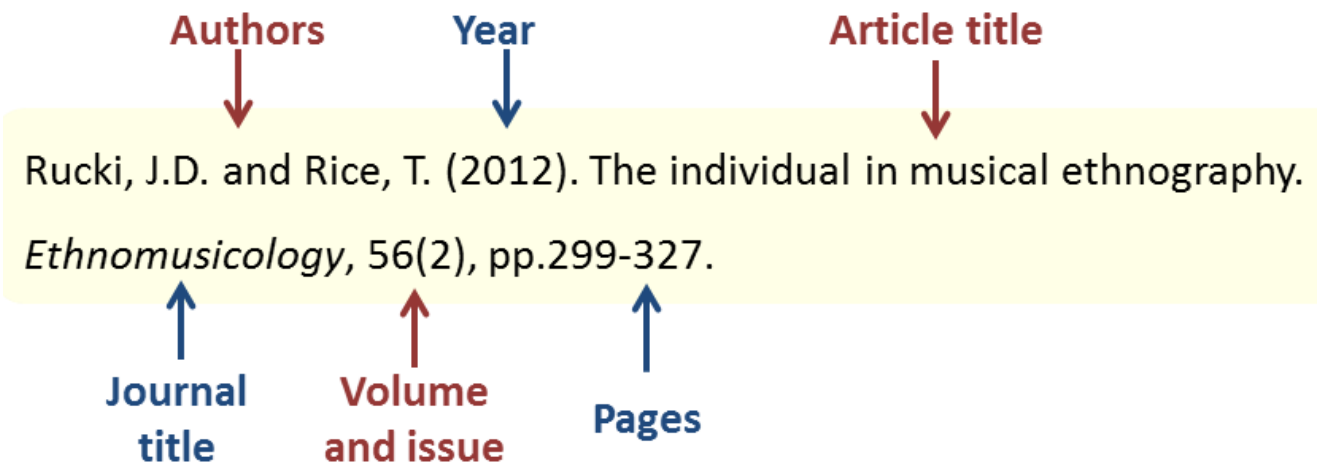
Besides this we get also at least the partition of works by the journal or publisher, the partition of works by the publication year, and the vector of number of pages.

Types of units

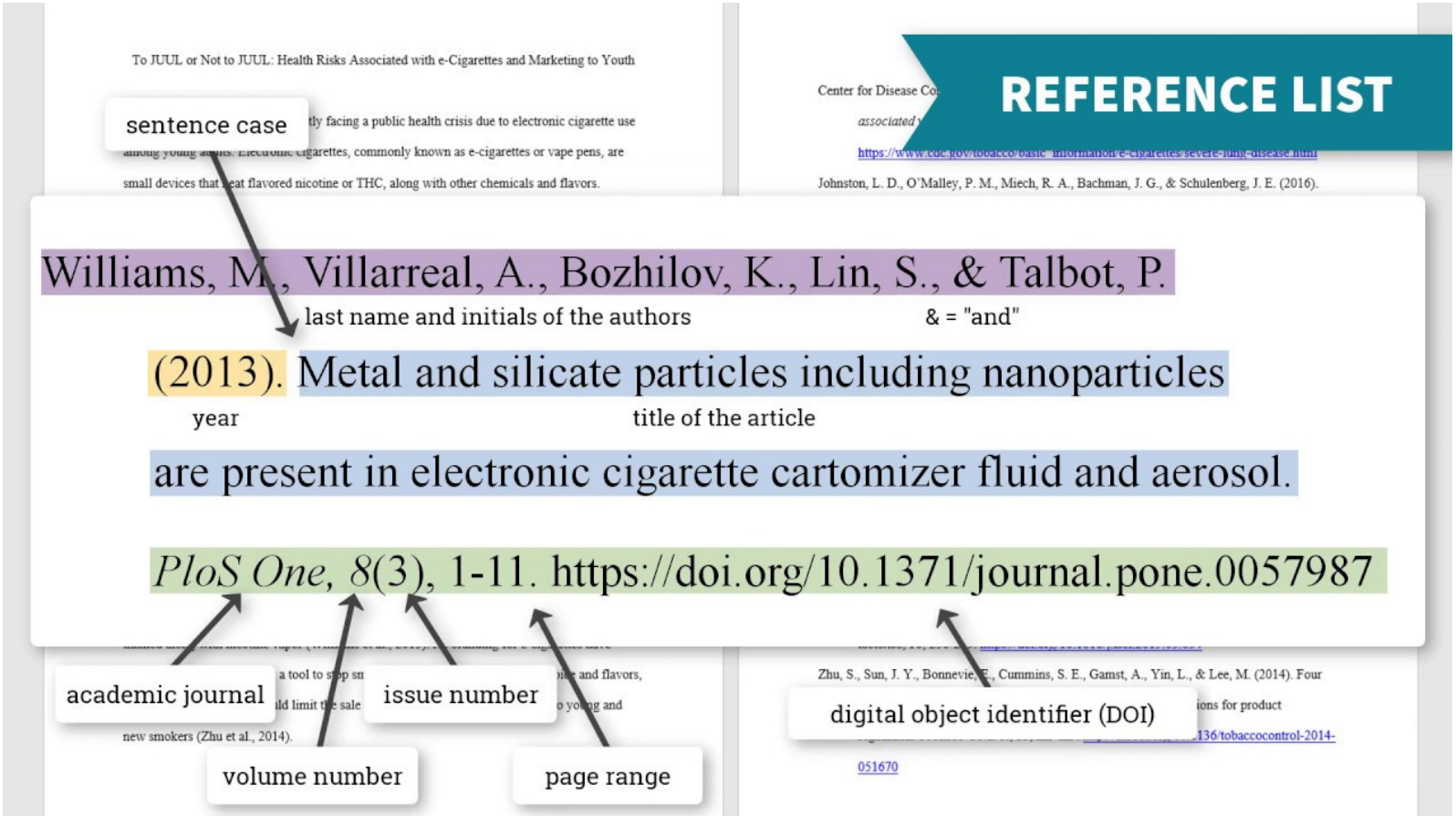
[11]

References


Author(s)
Book Title
Miragliotta, Narelle, Wayne Errington and Nicholas Barry. The Australian Political System in Action.
2nd ed. South Melbourne: Oxford University Press, 2012.
Edition (if applicable)
Place of publication
Publisher
Year of publication


Authors
Year
Article title
Rucki, J.D. and Rice, T. (2012). The individual in musical ethnography.
Ethnomusicology, 56(2), pp.299-327.
Journal title
Volume and issue
Pages

. . . References


To JUUL or Not to JUUL: Health Risks Associated with e-Cigarettes and Marketing to Youth
sentence case
small devices that heat flavored nicotine or THC, along with other chemicals and flavors.
new smokers (Zhu et al., 2014).
REFERENCE LIST
Johnston, L. D., O'Malley, P. M., Miech, R. A., Bachman, J. G., & Schulenberg, J. E. (2016).
Williams, M., Villarreal, A., Bozhilov, K., Lin, S., & Talbot, P.
last name and initials of the authors
& = "and"
(2013). Metal and silicate particles including nanoparticles
year
title of the article
are present in electronic cigarette cartomizer fluid and aerosol.
PloS One, 8(3), 1-11. https://doi.org/10.1371/journal.pone.0057987
academic journal
issue number
volume number
page range
digital object identifier (DOI)
Zhu, S., Sun, J. Y., Bonnevie, E., Cummins, S. E., Gamst, A., Yin, L., & Lee, M. (2014). Four
051670

APA citation style

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/

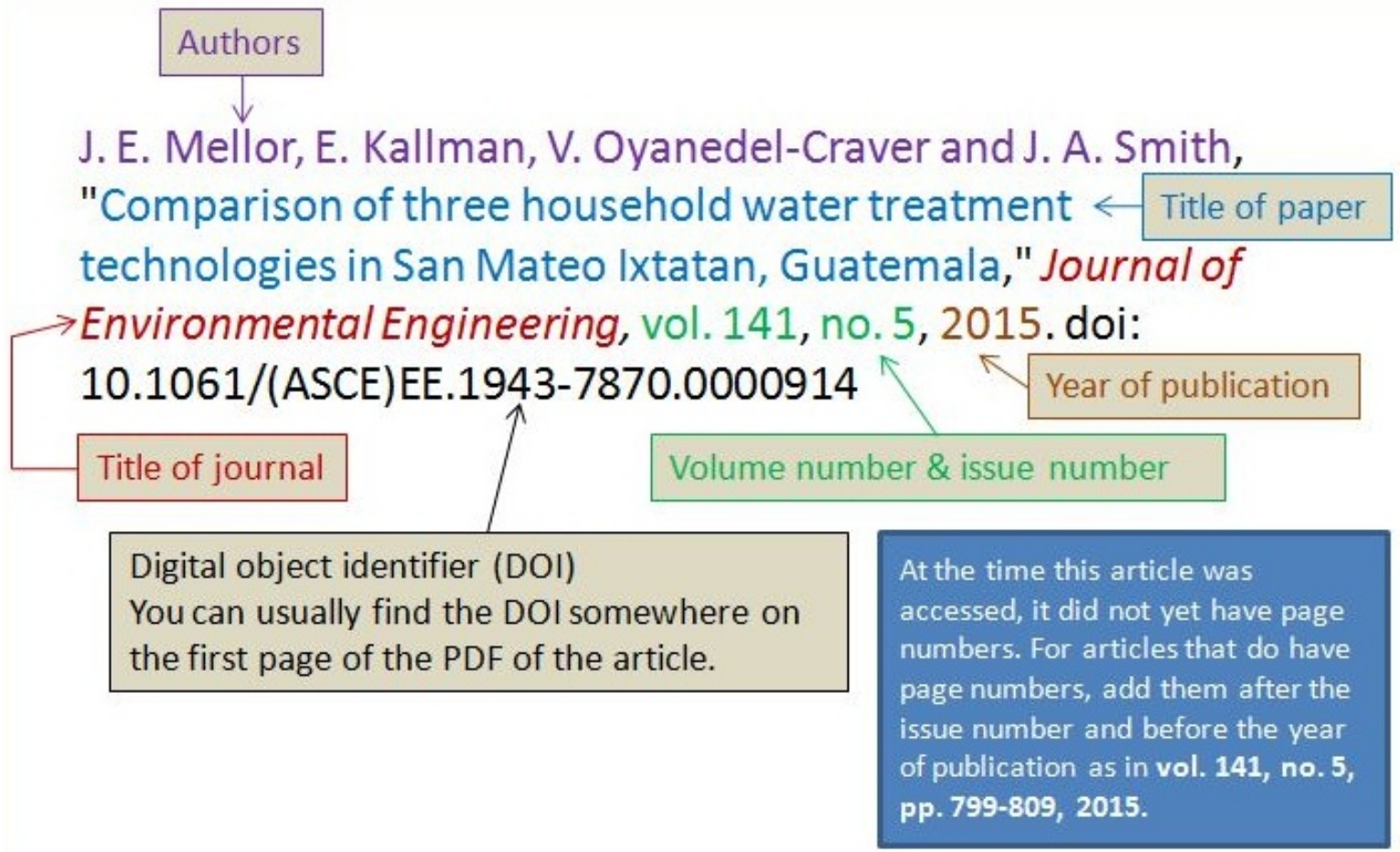
Journal Article from a Database (with a doi):

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

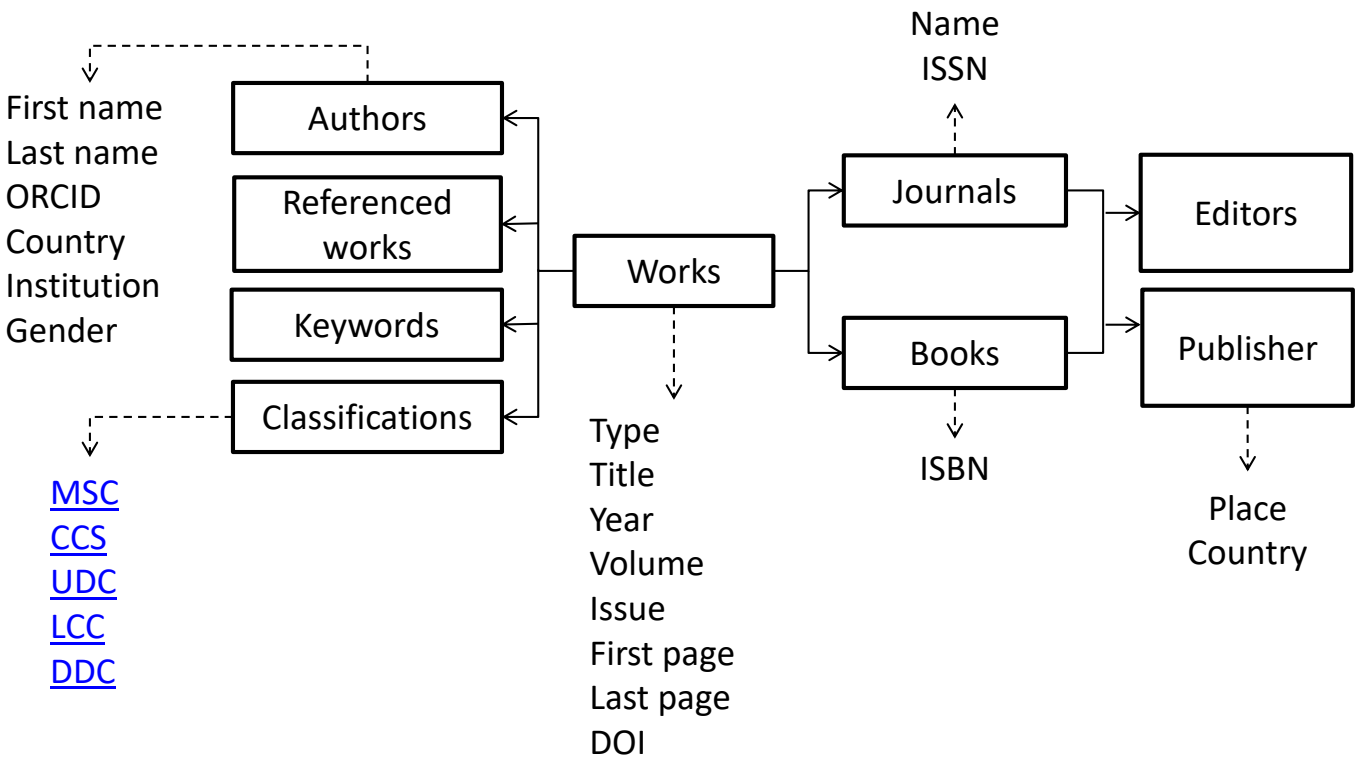
Non-periodical Web Document, Web Page, or Report (Website)

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

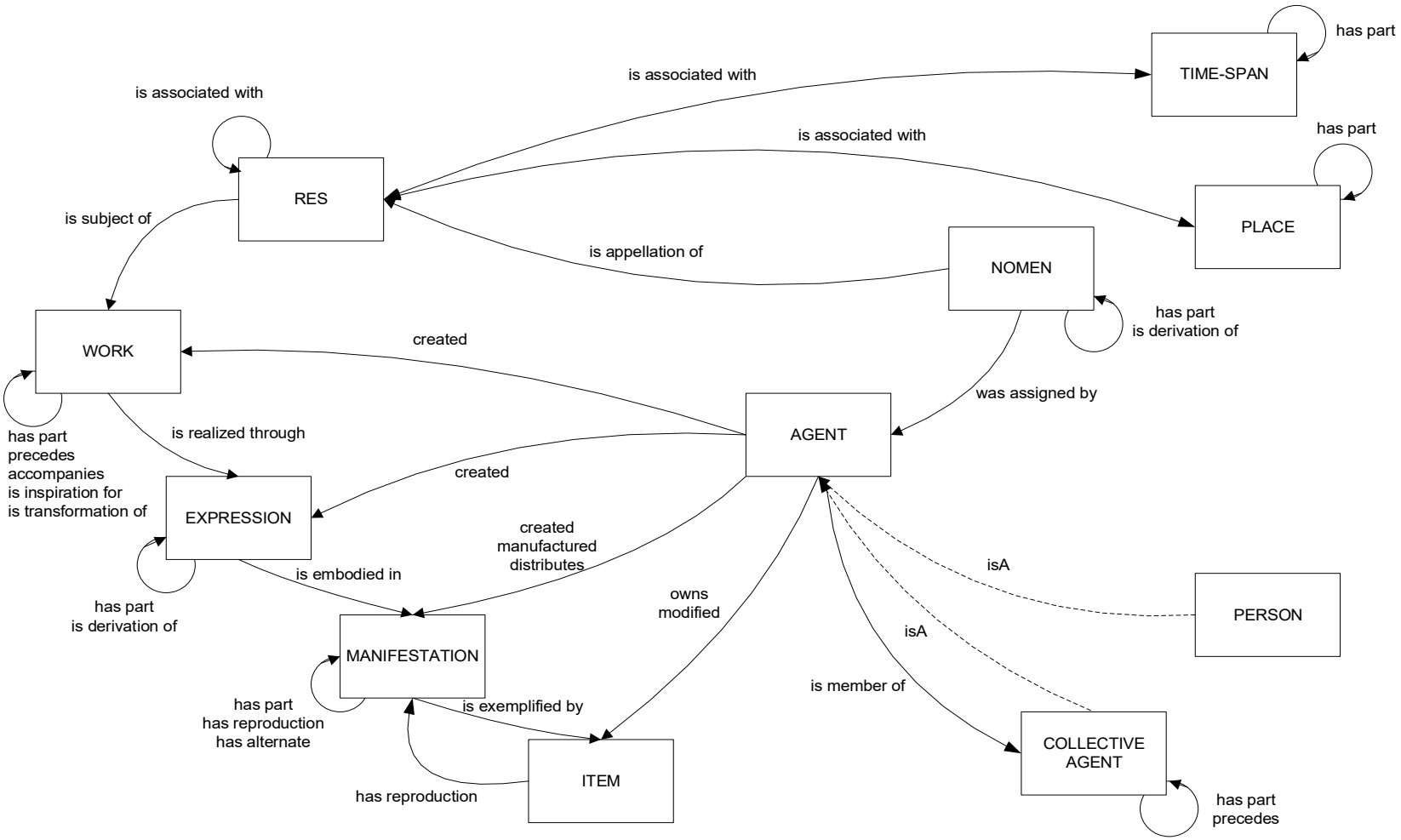
IEEE citation style



Basic types of entities



A scheme of basic types of entities in IFLA-LRM [19]



Records from BiBTEX


```
@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

Records from 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

Records from Zentralblatt


```
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

Record from 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

Problems in producing networks

Most of the source bibliographic data are semi-structured – they are available in the form of records from some data base. Selected fields in the record represent different units: names of people, names of journals, keywords, IDs of works, countries, institutions . . . Unfortunately the names of these units are usually not stored in a standardized way.

Synonymy: Unit names meaning the same. Make a partition. Identify (shrink) equivalent units.

Homonymy: Same unit names having different meanings. Correct the data in your copy of the data base.

When the unit names are extracted from the text the so called *stopwords* are omitted. The equivalence is automatically determined using stemming or lemmatization.
Names: many ways to write the name. Some data bases are trying to standardize the names (DBLP, ZB, ResearcherId). Chinese, 100 names.
Keywords: provided in data or extracted from the text (title, abstract). Key phrases.
There can be also errors (typos) in the data base – correct them in your copy of the data base data.

Analyses

The saved records from a data base can still contain some inconsistencies. Some of them are detected as results of the analyses. The simplest way to deal with them is to correct them in the saved data base file and rerun the creation of `Pajek`'s files and analyses.

To improve the quality of the data some tools for detecting (possible) inconsistencies could be developed.

Check (in `Pajek`) the obtained networks for multiple lines and remove them, if they exist. Remove also the loops from the citation network.

Primeri opisov del

cores

1. Descriptions

Most of the source bibliographic data are semi-structured – they are available in the form of records from some database. Selected fields in the record represent different units: names of people, names of journals, keywords, IDs of works, countries, institutions, etc. Unfortunately, the names of these units are usually not stored in a standardized way.

- detail of description (list of attributes)
- completeness of description (all authors?)
- Citation formats: Different academic styles, guided by different associations: 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 (Russian state standard).

1. Descriptions

source

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.

1. Descriptions

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
```

1. Descriptions

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  -
```

1. Descriptions

Graph products

A typical description in bibliographies from books and (survey) papers contains the following elements:

1. Names of authors; sometimes not complete (et al.)

   `WASSERMAN S, 1994, SOCIAL NETWORK ANAL`

   for

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

2. Title
3. Publication year (date)

for papers:

1. Journal
2. Volume
3. Issue
4. Pages

for books: - Publisher (Company, Place) ¿ examples

Problem 2: Different cultures

in different disciplines

1. number of co-authors; PhD-candidates supposed to publish as the only author, credit
2. writing of names (initial/full name, first name first/last)
3. the order of first and last name (French, Spanish, Arabian names, etc., names with prefixes).
4. some journals have special rules about abbreviations of journal names

```
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).
```

Examples of diverse citation practices

1. Vol, Issue, pages
2. Paper number
3. Citation without paper title

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

– pages

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

– volume, issue, first page

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

– volume, first page, year

Problem 4: Non-Latin alphabets

Cyrillic to Latin (Unicode, automatic transcription)
Transliteration
citeNet.R - the file making citation network
Depends on the language into which you transliterate the Russian name.

```
English: Pyotr Ilyich Tchaikovsky
German: Pjotr Iljitsch Tschaikowski
French: Piotr Ilitch Tchaïkovski
Spanish: Piotr Ilich Chaikovski
Italian: Pëtr Il'ič Čajkovskij
Slovenian: Peter Iljič Čajkovski
```

Russian

cores

cores

cores

cores

ECML PKDD Discovery Challenge - Recommending Given Names

cores

Doerfel, S.; Jäschke, R., Stumme, G. (2012), Publication Analysis of the Formal Concept Analysis Community. In F. Domenach; D.I. Ignatov, J. Poelmans, ed., 'ICFCA 2012', Springer, Berlin/Heidelberg, pp. 77-95.
Therefore, we employed the normalization steps described in [16] with an additional removal of diacritics (e.g., 'á' and 'ä' were replaced by 'a'). We used different heuristics, e.g., the Levenshtein distance, to nd errors in author names and titles. All references without authors (often encountered for cited web pages) were removed from the dataset. Since many publications were cited as different editions or prior to their publication ('to appear'), we normalized the publication year by dating back different editions to the earliest mentioned date of publication. For example, the collected papers of Charles S. Peirce [47] were cited with different publication years (1931, 1935, 1953, 1958, 1966) which we normalized to 1931.

J. Voss, A. Hotho, and R. Jäschke. Mapping bibliographic records with bibliographic hash keys. In R. Kuhlen, editor, Information: Droge, Ware oder Com- mons?, Proceedings of the ISI. Hochschulverband Informationswissenschaft, Verlag Werner Hulsbusch, 2009.

Introduction

individualno
pripadnost skupini
Vzhodnoslovansko osebno ime Lastno ime Občno ime
Kassel datasets
[17] [16] [1] [18]

Acknowledgments

This work is supported in part by the Slovenian Research Agency (research program P1-0294 and research projects J5-2557, J1-2481 and J5- 4596), and prepared within the framework of the COST action CA21163 (HiTEc).

References I

 Ulrichsweb, 2022.

 Vladimir Batagelj.
Efficient algorithms for citation network analysis, 14 Sep 2003.

 Vladimir Batagelj.
On fractional approach to analysis of linked networks.
Scientometrics, 123(2):621–633, 2020.

 Vladimir Batagelj and Monika Cerinšek.
On bibliographic networks.
Scientometrics, 96(3):845–864, 2013.

 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.

 Vladimir Batagelj and Daria Maltseva.
Temporal bibliographic networks.
J. Informetr., 14(1):Article No. 101006, 2020.

References II


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.


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.


Monika Cerinšek and Vladimir Batagelj.
Network analysis of Zentralblatt MATH data.
Scientometrics, 102(1):977–1001, 2015.


Ann T. Curran and Henriette D. Avram.
The identification of data elements in bibliographic records, 1967.

References III


B. Jones.
Computational geometry database, 2002.


Hugo Liu.
Montylingua: A free, commonsense-enriched natural language understander for english (version 2.1), 2004.


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.


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.


John R. Talburt.
Entity Resolution and Information Quality.
Morgan Kaufmann, 2011.

References IV

Bert TePaske-King and Norman Richert.
The identification of authors in the mathematical reviews database.
Issues Sci. Technol. Librariansh., (31), 2001.

M. Windham.
Unstructured Data Analysis: Entity Resolution and Regular Expressions in SAS.
SAS Institute, 2019.

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