



Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Analysis of bibliographic networks

Daria Maltseva, Vladimir Batagelj

NRU HSE Moscow, IMFM Ljubljana, and IAM UP Koper

**10th International Summer School
“Analysis of Scientific Networks”**

Moscow, July, 15–21, 2019



Outline

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

1 Transforming bibliographic data into networks

- Goals, research questions, and theory
- Bibliographic data:
 - Structure of bibliographic data
 - Bibliographic databases and data collection
 - Networks from bibliographic data
- Problems associated with bibliographic data collection
- Tools for collection and maintenance of bibliographic data
- Conversion to networks

2 Analyzing bibliographic networks

- Preanalysis, boundary, basic statistics
- Citation among authors / journals
- Collaboration among authors
- Co-citation among authors / journals
- Keywords co-occurrence



Goals, research questions, and theory

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

Goals: Study of social and cognitive structure of different scientific fields.

Research questions:

- How do scientists collaborate with each other? How different groups of scientists relate to each other?
→ Co-authorship, co-citation network analysis
- How the certain fields in science develop trough time?
→ Citation network analysis
- What is the topic structure of the scientific filed?
→ Co-occurrence key words analysis

Different levels (authors, institutions, countries) and units (publications, journals, keywords) of analysis.



Goals, research questions, and theory

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Theoretical background:

- The “philosophical” grounds of the field go back to the works of the sociologist R. Merton and the historians of science D. de Sola Price and G. Small.
- E. Garfield - the first scientific citation index - Science Citation Index (SCI) [Garfield, 1972]. Since its creation, the citation analysis has grown into an independent research field [Wilson, 1999, Bar-Ilan, 2008]
- D. Crane [Crane, 1972] building on D.J. de Solla Price's work on citation networks, introduced the notion of “invisible college” (R. Boyle, 17th century) - a core group of scientists who collaborated with each other and generated a disproportionate volume of new ideas - and showed that internal social structure of the scientific community influences the development of the ideas, and **study of informal social and communication structures can bring important results for understanding the modern development of scientific disciplines.**



Bibliographic data: some examples

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Papers

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. Sociological theory, 201-233.

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. Reviews of modern physics, 74(1), 47.

Batagelj V., Ferligoj A., Squazzoni F. (2017) The emergence of a field: a network analysis of research on peer review. Scientometrics, 113(1), 503-532.

Books

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

Burt R. S. Structural holes: The social structure of competition.
-- Harvard university press, 2009.

Batagelj, Vladimir, Andrej Mrvar, and Wouter de Nooy. "Exploratory social network analysis with Pajek." (2008).



Bibliographic data: some examples

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Survey bibliographies:

- 1 Web survey bibliography
- 2 Bibliography of Research Methods Texts
- 3 A survey and annotated bibliography of multiobjective combinatorial optimization
- 4 Community detection in graphs

Book bibliography:

- 1 Handbook of Product Graphs, Second Edition
- 2 Computational Geometry

Bibliography of scientific community:

- 1 Bibliography on Self-Organizing Map (SOM) method
- 2 Computational Geometry Bibliographies
- 3 TUG bibliography archive



Bibliographic databases

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Records from **BiBTeX** (reference management software for formatting lists of references, typically used together with the LaTeX document preparation system).

```
@Article{int:Mizuno1,  
  author =      "S. Mizuno",  
  title =      "An  $\{0(n^3)L\}$  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",  
}
```

BiBTeX → Pajek converter **Bib2Pajek.py**



Bibliographic databases

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Records from **DBLP** (online database database of a computer science bibliography).

```
<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>
```

DBLP XML data to Pajek Converter → Pajek converter **DBLP2Pajek.py**.



Bibliographic databases

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Records from **Zentralblatt Math** (international reviewing service providing reviews and abstracts for articles in pure and applied mathematics).

```
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  $\text{Gr}(1,\mathbb{P}^3)$ .
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{\i}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; 1532-4125
```

ZBml files → Pajek converter **ZBml.py**.



Bibliographic databases

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Records from **Web of Science** (online subscription-based scientific citation indexing service providing a comprehensive citation search).

PT J

AU Elmer, T

Boda, Z

Stadtfeld, C

TI The co-evolution of emotional well-being with weak and strong friendship ties

SO NETWORK SCIENCE

LA English

DT Article

DE social networks; ordered stochastic actor-oriented models [...]

ID ADOLESCENT DEPRESSIVE SYMPTOMS; [...]

AB Social ties are strongly related to well-being. But what characterizes this relationship? [...]

C1 [Elmer, Timon; Boda, Zsofia; Stadtfeld, Christoph] Swiss Fed Inst Technol, Chair Social Networks, Dept Humanities Social & Polit Sci, Zurich, Switzerland.

RP Elmer, T (reprint author), Swiss Fed Inst Technol, Chair Social Networks, Dept Humanities Social & Polit Sci, Zurich, Switzerland.

EM timon.elmer@gess.ethz.ch; [...]

CR Aharony N, 2011, PERSVASIVE MOB COMPUT, V7, P643, DOI 10.1016/j.pmcj.2011.09.004
Baerveldt C., 2004, CONNECTIONS, V26, P11

Reis H. T., 2000, HDB RES METHODS SOCI

Ripley Ruth M., 2015, MANUAL RSIENA

...



Bibliographic databases

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

NR 83
TC 1
PU CAMBRIDGE UNIV PRESS
PI NEW YORK
PA 32 AVENUE OF THE AMERICAS, NEW YORK, NY 10013-2473 USA
SN 2050-1242
J9 NETW SCI
JI Netw. Sci.
PD SEP
PY 2017
VL 5
IS 3
BP 278
EP 307
DI 10.1017/nws.2017.20
PG 30
SC Social Sciences - Other Topics
GA FFOAM
UT WOS:000408564600003
ER

X-format → WoS-format → (WoSPajek) → Pajek files
Done: RIS → WoS



Structure of bibliographic data

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Structure of bibliographic data

Bibliographic analysis

D. Maltseva,
V. Batagelj

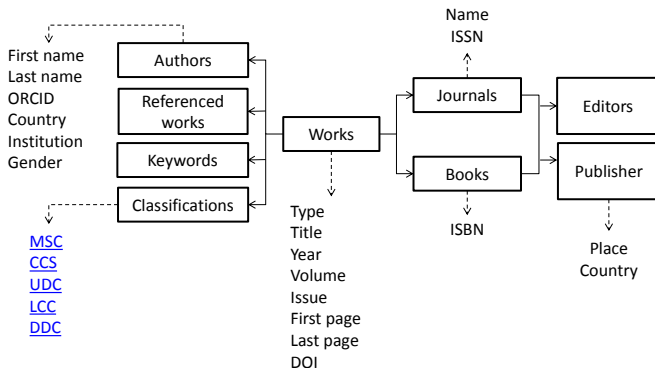
Goals

Bibliographic
Data

Problems
with Data

Tools

Books



Some attributes have time intervals of their activity.

The scheme can be extended to other types of works (video, pictures, data, programs, etc.) see [source](#).

The role of classifications in Internet resource description and discovery



Networks from bibliographic data

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

We can derive some two-mode networks on selected topics from bibliographies from:

- books and survey papers,
- special bibliographies
- bibliographic services
 - Web of Science
 - Scopus
 - SICRIS
 - CiteSeer
 - Zentralblatt MATH
 - Google Scholar
 - DBLP Bibliography

The same approach can be used also for other types of works:

- US patent office
- IMDb



Networks from bibliographic data

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Two-mode networks on selected topics:

- works \times authors (**WA**),
- works \times journals or book publishers (**WJ**);
- works \times keywords **WK**);
- works \times classification (**WC**) - from some data;
- authors \times institutions (**AI**);
- authors \times countries (**AC**).

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.



Networks from bibliographic data

Creating your own bibliographic data base in Excel

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

How to describe a network $\mathbf{N} = (\mathbf{V}, \mathbf{L}, \mathbf{P}, \mathbf{W})$? In principle the answer is simple - we list its components: nodes \mathbf{V} , links \mathbf{L} , node properties \mathbf{P} , and link weights \mathbf{W} .

The simplest way is to describe a network \mathbf{N} by providing (\mathbf{V}, \mathbf{P}) and (\mathbf{L}, \mathbf{W}) in a form of two tables.



Networks from bibliographic data

Creating your own bibliographic data base in Excel

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

As an example, let us describe a part of network determined by the following works:

Generalized blockmodeling, Clustering with relational constraint, Partitioning signed social networks, The Strength of Weak Ties.

There are **nodes of different types** (modes): *persons, papers, books, series, journals, publishers*;

and **different relations** among them: *author_of, editor_of, contained_in, cites, published_by*.

For small bibliographies both tables can be maintained in Excel and exported as text in **CSV** (Comma Separated Values) format.



Networks from bibliographic data

Creating your own bibliographic data base in Excel

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

bibNodes.csv

```
name;mode;country;sex;year;vol;num;fPage;lPage;x;y
"Batagelj, Vladimir";person;SI;m;;;;;809.1;653.7
"Dorean, Patrick";person;US;m;;;;;358.5;679.1
"Ferligoj, Anuška";person;SI;f;;;;;619.5;680.7
"Granovetter, Mark";person;US;m;;;;;145.6;660.5
"Moustaki, Irini";person;UK;f;;;;;783.0;228.0
"Mrvar, Andrej";person;SI;m;;;;;478.0;630.1
"Clustering with relational constraint";paper;;;1982;47;;413;426;684.1;380.1
"The Strength of Weak Ties";paper;;;1973;78;6;1360;1380;111.3;329.4
"Partitioning signed social networks";paper;;;2009;31;1;1;11;408.0;337.8
"Generalized Blockmodeling";book;;;2005;24;;1;385;533.0;445.9
"Psychometrika";journal;;;;;;741.8;086.1
"Social Networks";journal;;;;;;321.4;236.5
"The American Journal of Sociology";journal;;;;;;111.3;168.9
"Structural Analysis in the Social Sciences";series;;;;;;310.4;082.8
"Cambridge University Press";publisher;UK;;;;;;534.3;238.2
"Springer";publisher;US;;;;;;884.6;174.0
```

In large networks, to avoid the empty cells, we split a network to some subnetworks - a collection.



Networks from bibliographic data

Creating your own bibliographic data base in Excel

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

bibLinks.csv

```
from;relation;to
"Batagelj, Vladimir";authorOf;"Generalized Blockmodeling"
"Doreian, Patrick";authorOf;"Generalized Blockmodeling"
"Ferligoj, Anuška";authorOf;"Generalized Blockmodeling"
"Batagelj, Vladimir";authorOf;"Clustering with relational constraint"
"Ferligoj, Anuška";authorOf;"Clustering with relational constraint"
"Granovetter, Mark";authorOf;"The Strength of Weak Ties"
"Granovetter, Mark";editorOf;"Structural Analysis in the Social Sciences"
"Doreian, Patrick";authorOf;"Partitioning signed social networks"
"Mrvar, Andrej";authorOf;"Partitioning signed social networks"
"Moustaki, Irini";editorOf;"Psychometrika"
"Doreian, Patrick";editorOf;"Social Networks"
"Generalized Blockmodeling";containedIn;"Structural Analysis in the Social Sciences"
"Clustering with relational constraint";containedIn;"Psychometrika"
"The Strength of Weak Ties";containedIn;"The American Journal of Sociology"
"Partitioning signed social networks";containedIn;"Social Networks"
"Partitioning signed social networks";cites;"Generalized Blockmodeling"
"Generalized Blockmodeling";cites;"Clustering with relational constraint"
"Structural Analysis in the Social Sciences";publishedBy;"Cambridge University Press"
"Psychometrika";publishedBy;"Springer"
```



Networks from bibliographic data

Factorization and description of large networks

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

To save space and improve the computing efficiency we often replace values of categorical variables with integers. In R this encoding is called a **factorization**.

We enumerate all possible values of a given categorical variable (coding table) and afterwards replace each its value by the corresponding index in the coding table.

This approach is used in most programs dealing with large networks. Unfortunately the coding table is often a kind of meta-data.



Networks from bibliographic data

Factorization and description of large networks

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

CSV2Pajek.R

```
# transforming CSV file to Pajek files
# by Vladimir Batagelj, June 2016
setwd("C:/Users/batagelj/work/Python/graph/SVG/EUSN")
colC <- c(rep("character",4),rep("integer",7)); nas <- c("", "NA", "NaN")
nodes <- read.csv2("bibNodes.csv",encoding='UTF-8',colClasses=colC,na.strings=nas)
n <- nrow(nodes); M <- factor(nodes$mode); S <- factor(nodes$sex)
mod <- levels(M); sx <- levels(S); S <- as.numeric(S); S[is.na(S)] <- 0
links <- read.csv2("bibLinks.csv",encoding='UTF-8',colClasses="character")
F <- factor(links$from,levels=nodes$name,ordered=TRUE)
T <- factor(links$to,levels=nodes$name,ordered=TRUE)
R <- factor(links$relation); rel <- levels(R)
net <- file("bib.net","w"); cat('*vertices ',n,'\n',file=net)
clu <- file("bibMode.clu","w"); sex <- file("bibSex.clu","w")
cat('% ',file=clu); cat('% ',file=sex)
for(i in 1:length(mod)) cat(' ',i,mod[i],file=clu)
cat('\n*vertices ',n,'\n',file=clu)
for(i in 1:length(sx)) cat(' ',i,sx[i],file=sex)
cat('\n*vertices ',n,'\n',file=sex)
for(v in 1:n) {
  cat(v,' ',nodes$name[v],'\n',sep='',file=net);
  cat(M[v],'\n',file=clu); cat(S[v],'\n',file=sex)
}
for(r in 1:length(rel)) cat('*arcs :',r,' ',rel[r],'\n',sep='',file=net)
cat('*arcs\n',file=net)
for(a in 1:nrow(links))
  cat(R[a],': ',F[a],', ',T[a],', 1 1 ',rel[R[a]],'\n',sep='',file=net)
close(net); close(clu); close(sex)
```



Networks from bibliographic data

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

bib.net

Line breaks \n were added manually

```
*vertices 16
1 "Batagelj, Vladimir"
2 "Doreian, Patrick"
3 "Ferligoj, Anuška"
4 "Granovetter, Mark"
5 "Moustaki, Irini"
6 "Mrvar, Andrej"
7 "Clustering with\nrelational constraint"
8 "The Strength of Weak Ties"
9 "Partitioning signed social networks"
10 "Generalized Blockmodeling"
11 "Psychometrika"
12 "Social Networks"
13 "The American\nJournal of Sociology"
14 "Structural Analysis in\nthe Social Sciences"
15 "Cambridge University Press"
16 "Springer"

*arcs :1 "authorOf"
*arcs :2 "cites"
*arcs :3 "containedIn"
*arcs :4 "editorOf"
*arcs :5 "publishedBy"

1: 1 10 1 1 "authorOf"
1: 2 10 1 1 "authorOf"
1: 3 10 1 1 "authorOf"
1: 1 7 1 1 "authorOf"
1: 3 7 1 1 "authorOf"
1: 4 8 1 1 "authorOf"
4: 4 14 1 1 "editorOf"
1: 2 9 1 1 "authorOf"
1: 6 9 1 1 "authorOf"
4: 5 11 1 1 "editorOf"
4: 2 12 1 1 "editorOf"
3: 10 14 1 1 "containedIn"
3: 7 11 1 1 "containedIn"
3: 8 13 1 1 "containedIn"
3: 9 12 1 1 "containedIn"
2: 9 10 1 1 "cites"
2: 10 7 1 1 "cites"
5: 14 15 1 1 "publishedBy"
5: 11 16 1 1 "publishedBy"
```

bibMode.clu, **bibSex.clu**, **bib.paj**, **bib.ini**



Networks from bibliographic data

Bibliographic analysis

D. Maltseva,
V. Batagelj

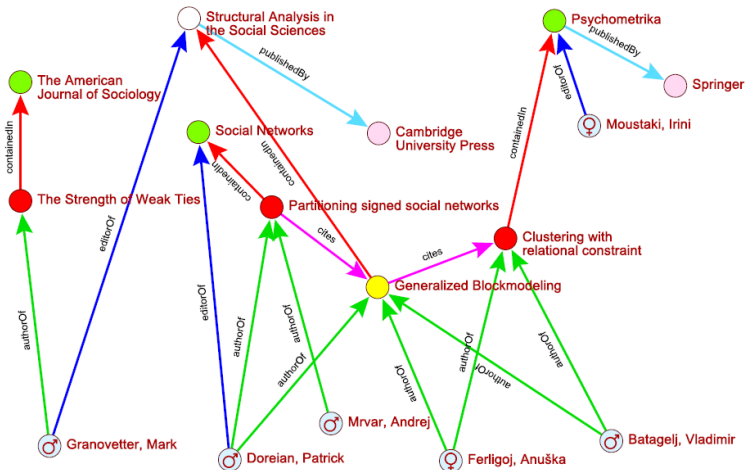
Goals

Bibliographic
Data

Problems
with Data

Tools

Books





Networks from bibliographic data

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

With this approach, only small data sets can be obtained.

Researcher knows and can check each record and make data set consistent.

For small bibliographies, **Text2Pajek** can be used.



Collecting the data from WoS with WoS2Pajek

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

For converting WoS file into networks in Pajek's format a program **WoS2Pajek** was developed (in Python).

The procedure is the following:

- 1 export data from WoS
- 2 combine files into WoS file
- 3 run WoS2Pajek
- 4 get the collection of networks



WoS2Pajek installation

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

- 1 Install **Python version 2.7**
- 2 Download **WoS2pajek15** (the latest version)
- 3 Download **Montylingua package** (montylingua-2.1) and copy it into directory in Python:
`C:\Python27\Lib\site-packages -> montylingua-2.1`
- 4 Add the environment variable:
Дополнительные параметры системы -
Переменные среды - Системные переменные -
Создать - Имя переменной: MONTYLINGUA
Значение переменной:
`C:\Python27\Lib\site-packages\montylingua-2.1\python`
OK - OK - OK



WoS2Pajek installation

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

- 5 Put the folder "resources"(WoS2Pajek goes with it) to the folder where WoS2Pajek is located.

- 6 Run Python:

```
import sys; wdir = r'C:\WoS2Pajek'
sys.path.append(wdir)
MLdir = r'C:\Python27\Lib\site-packages
\montylingua-2.1\python'
sys.path.append(MLdir)
import WoS2Pajek15
...
Module Wos2Pajek imported.
```

- 7 Run WoS2Pajek15

Edit with IDLE
Run module



Names of works in WoS2Pajek

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

The usual *ISI name* of a work (field CR)

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

has the following structure

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

All its elements are in upper case.

In WoS the same work can have different ISI names. To improve the precision the program WoS2Pajek supports also *short names* (similar to the names used in HISTCITE output). They have the format:

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

For example: LEFKOVIT_L(1985)70:585

From the last names with prefixes VAN, DE, ... the space is deleted.

Unusual names start with character * or \$.



Names of works in WoS2Pajek

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

In the CR field other forms of ISI names and several errors and inconsistencies can be found:

NEWMAN MEJ, 2004, PHYS REV E 2, V69, ARTN 066133
PALLA G, 2005, NATURE, V435, P814, DOI 10.1038/nature03607
PAPIN JA, 2004, TRENDS BIOCHEM SCI, V29, P641, DOI
10.1016/j.tibs.2004.10.001
DOLCINI MM, 2005, J ADOLESCENT HEALTH, V36, UNSP 267.E6-15
EVANS JD, 2001, GENOME BIOL, V2, UNSP RESEARCH0001
NEWMAN MEJ, 2001, IN PRESS COMPLEX NETUNSP 215239
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
BRICKER PD, 1968, OCT M PSYCH SOC ST L : BRICKER

We decided to treat in short names the ARTN and UNSP values as BP values.
We also remove the DOI parts. There are also irregular names in AU field:

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

The user can correct the typing errors and nonuniformities on the WoS file.



Program WoS2Pajek

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

WoS2Pajek produces the following files:

- citation network: works \times works, **Ci**;
- authorship (two-mode) network: works \times authors, **WA**; for works without complete description only the first author is known;
- keywords (two-mode) network: works \times keywords, **WK**; only for works with complete description;
- journals (two-mode) network: works \times journals, **WJ**; field J9;
- partition of works by the publication year, **year**;
- partition of works, **DC** – complete description (1) / ISI name only (0);
- vector number of pages, **np**; PG or EP – BP + 1.



Program WoS2Pajek

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

The keywords are obtained from the fields TI (title), ID, DE and AB (abstract). From the text the **stopwords** are removed and a list of words is produced. The words are lemmatized using **MontyLingua** package.

In future versions additional networks can be derived: works \times discipline, works \times countries, ...

In version 0.7 a GUI support (based on Tkinter) for specifying the program parameters was implemented.

Similar package of programs was produced also by **Loet Leydesdorff**.



Program WoS2Pajek

Bibliographic analysis

D. Maltseva,
V. Batagelj

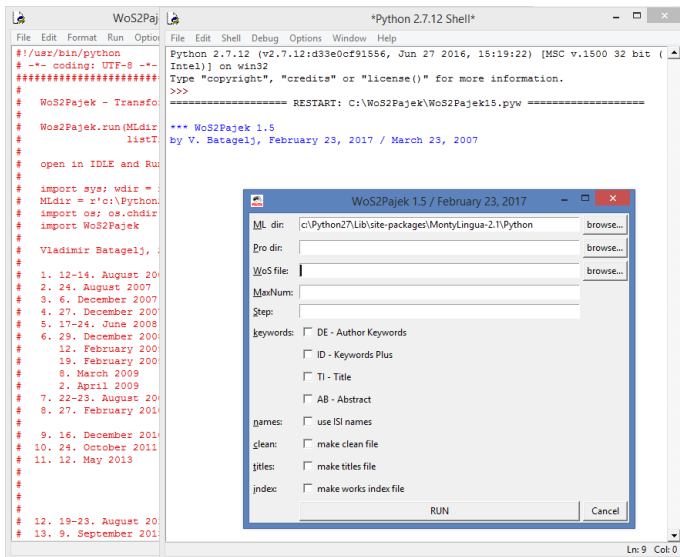
Goals

Bibliographic
Data

Problems
with Data

Tools

Books



D. Maltseva, V. Batagelj

Bibliographic analysis



Program WoS2Pajek

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

The current version of WoS2Pajek requires 7 parameters to be given by the user:

- MontyLingua directory: path to the directory in which the MontyLingua package is installed (put it also in the PATH env-variable);
- project directory: where the output files are saved;
- WoS file;
- maxnum – estimate of the number of all vertices (number of records + number of cited Works) – 30* number of records;
- step – prints info about each k*step record as a trace; step = 0 – no trace.
- use ISI name / short name;
- make a clean WoS file without duplicates;
- boolean list [DE, ID, TI, AB] specifying which fields are sources of keywords.



Collecting the data from Web of Science

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

When preparing for exploration of selected topic we first search the WoS for records matching selected keys. WoS allows storing the hits on a file, but only 500 at once.

For the list of hits we can in the **Citation report**, using the option **View citing articles**, get also the list of all works citing the works from the first list. We also save these records and join all the files into a single file to which we apply WoS2Pajek.

Using Pajek we identify the important only cited works – out-degree = 0 and large in-degree. We search for them in WoS and try to add their descriptions.

In this way we produce the final source data set for our analysis.



R code for WoS files

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

To compose a complete WoS data file, the followig code can be used:

```
getwd()
files <- dir()
files
if (is.element(el = "CompleteWoSDataBase.txt", set = files) == TRUE)
file.remove("CompleteWoSDataBase.txt")
for (i in 1:length(files)){
  tmp <- readLines(files[i])
  cat(tmp, file = "CompleteWoSDataBase.txt", sep = "\n", append = TRUE)
}
```



Types on DC file

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

When we combine partial files with saved records from WoS into a single file required by the program WoS2Pajek we can include into this file some additional lines:

Comments have the form

```
** comment
```

Besides this we can specify different types of input records using the lines of the form

```
*T n
```

where n is a type number (1, 2, ...). Since the same record can appear in different parts of the file its class is determined as the set of all corresponding types transformed in integer. For example:
 $\{3, 1\} \rightarrow 2^2 + 2^0 = 5$.



Networks from bibliographic data

The general procedure of transformation

Bibliographic analysis

D. Maltseva,
V. Batagelj

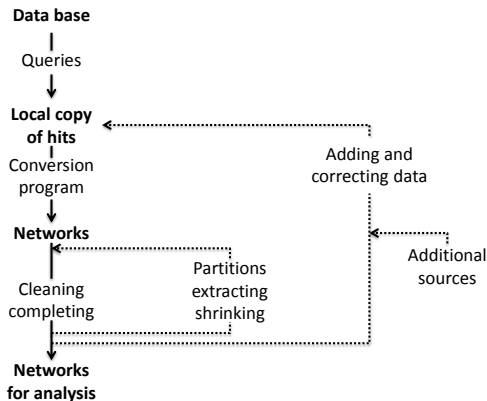
Goals

Bibliographic
Data

Problems
with Data

Tools

Books





Problems with bibliographic data

Problem 1: Descriptions

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

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, 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 relevant entities are included - authors)



Problems with bibliographic data

Problem 1: Descriptions

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Citation formats:

Different academic styles, guided by different associations:

- **MLA** (Modern Language Association of America)

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

- **APA** (American Psychological Association)

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

- **Chicago** (University of Chicago Press)

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



Problems with bibliographic data

Problem 1: Descriptions

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

- **GOST** (Russian state standard)

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

- **AMA** (American Medical Association)

White, H. (2012). Identity and Control. Princeton: Princeton University Press.

- **SCE** (Council of Science Editors)

White, Harrison. 2008. Identity and Control. 2nd ed. Princeton: Princeton University Press. p 456.



Problems with bibliographic data

Problem 1: Descriptions

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Bibliographic formats:

- **BibTeX** (LaTeX style),

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

- **EndNote** (Clarivate Analytics),

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

- **RIS** (Research Information Systems style), etc.

```
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 -
```



Problems with bibliographic data

Problem 1: Descriptions

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

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

- Names of authors; sometimes not complete (et al.)
- Title
- Publication year (date)

Papers:

- Journal
- Volume
- Issue
- Pages

Books:

- Publisher (Company, Place)

WASSERMAN S, 1994, SOCILA NETWORK ANAL

Wasserman, S., & Faust, K. (1994). Social network analysis:

Methods and applications (Vol. 8). Cambridge university press.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited.

Sociological Theory, 201-233.

White, H. C. (2008). Identity and control: How social formations emerge (Second Edition).

PRINCETON; OXFORD: Princeton University Press.



Problems with bibliographic data

Problem 2: Different cultures

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Many coauthors:

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Khalek, S. A., Abdelalim, A. A., ... & AbouZeid, O. S. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1), 1-29.

Source



Problems with bibliographic data

Problem 2: Different cultures

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

- Writing of names (initial/full name, first name first/last).
- The order of first and last name (French, Spanish, Arabian names etc., names with prefixes).
- Special rules about abbreviations of journal names (examples: **WoS, All That JAS, Science and Engineering Journals**).



Problems with bibliographic data

Problem 2: Different cultures

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

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

GRANOVET.MS, 1973, AM J SOCIOL, V78, P1360
GRANOVETTER MS, 1973, AMER JOUR SOCIOL, V78, P1360

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).



Problems with bibliographic data

Problem 2: Different cultures

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Examples of diverse citation practices:

- Vol, Issue, Pages
- Paper number
- Citation without paper title

AU (PY). TI. JI, BP-EP

Freeman, L. C., & White, D. R. (1993). Using Galois lattices to represent network data.

Sociological methodology, 127-146.

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.

AU, JI VL, IS (PY)

P. Erdos and A. Renyi, Publ. Math. Inst. Hung.

Acad. Sci. 5, 17 (1960).



Problems with bibliographic data

Problem 3: Non-Latin alphabets

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Some names can be written in several languages – the procedure of author disambiguation is needed.

Cyrillic to Latin (Unicode, automatic transcription).

R, stringi library:

```
> tail(N)
[1] "ГОМЗИН А"      "НЕДУМОВ Я"      "IVANOV I"      "АСТРАХАНЦЕВ Н"
[5] "ТРИПУТИНА В"   "МАКАГОНОВА Н"
> tail(R)
[1] "GOMZIN A"      "NEDUMOV A"      "IVANOV I"      "ASTRAHANCEV N"
[5] "TRIPUTINA V"   "MAKAGONOVA N"
```

Problems with character "Ъ":

```
> N[44]
[1] "ЗОРЬКИНА К"
> R[44]
[1] "ZOR'KINA K"
> utf8ToInt(R[44])
[1] 90 79 82 697 75 73 78 65 32 75
> T <- sapply(R,function(w)gsub(intToUtf8(697),"",w),USE.NAMES=FALSE)
> T[44]
[1] "ZOR'KINA K"
> utf8ToInt(T[44])
[1] 90 79 82 39 75 73 78 65 32 75
```



Problems with bibliographic data

Problem 3: Non-Latin alphabets

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Transliteration (**different approaches**)

Пётр Ильич Чайковский

English: Pyotr Ilyich Tchaikovsky

German: Pjotr Iljitsch Tschaikowski

French: Piotr Ilitch Tchaïkovski

Spanish: Piotr Ilich Chaikovski

Italian: Peter Ilyich Tchaikovsky

Slovenian: Peter Iljič Čajkovski



Problems with bibliographic data

Problem 4: Entity identification/resolution

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Homonymy (ambiguity)

Same names for different units

People:

Lorenzo Bartolini from the movie **Letters to Juliet** – many persons with the same name.

Smith, John W. - publications of the author(s) with this name spanned from 1868 to 2007.

“Multiple personalities”:

Harzing, A. W. (2015). Health warning: might contain multiple personalities — the problem of homonyms in Thomson Reuters Essential Science Indicators

"Three Zhang, four Li" effect (there are at least 623 different mathematicians with the name Zhang, Li in the MathSciNet Database)

Works:

ANONYMO (2015)



Problems with bibliographic data

Problem 4: Entity identification/resolution

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Synonymy

Different names for same units

People:

"Krivoshe\u005C\u005C i n, Leonid Evgen\u005C\prime evich" (using the TeX codes)

Mathematical Reviews Database – 20 name variations for this author.

Otfried Cheong (formerly **Otfried Schwarzkopf**): German computational geometer working in South Korea at KAIST.

Michel Marie Deza (formerly **Mikhail Efimovich Tytkin**): a Soviet and French mathematician (combinatorics, discrete geometry and graph theory).

Borštnik, N. S. Mankoč; Mankoč Borštnik, N.; Mankoč-Borštnik, Norma; Mankoč Borštnik, Norma Susana; Mankoc-Borstnik, N.S.; and Mankoč Borštnik, N.S. = same author.

Journals:

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 = same journal



Problems with bibliographic data

Problem 4: Entity identification/resolution

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Author names

- Some data bases are trying to standardize the names (DBLP, ZB, ResearcherId).
MathSciNet; **Orcid** - Enter author name in Search field
Scopus; **eLibrary** - Click on author's name and take the number after "authorid"

- Variations in the first names: Sort (last name, first name)

<https://orcid.org/0000-0002-0240-9446>

https://elibrary.ru/author_items.asp?authorid=155240

- Multi-alphabet (names written in different languages) - convert names to selected alphabet or use "dictionary".

AMS approach – look for details.



Problems with bibliographic data

Problem 4: Entity identification/resolution

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Journal titles and Books

- Form a key from initials of journal name and sort (key, journal name)
- International Standard Serial Number **ISSN**; International Standard Book Number **ISBN**.

Keywords

Provided in data or extracted from the text (title, abstract). Key phrases – use of dictionary.

- Errors (typos) in the data base – correct them in your copy of the data base data.
- Problem of equivalent keywords: form, words, language – stemming, lemmatization, dictionary.



Problems with bibliographic data

Problem 4: Entity identification/resolution

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Description of equivalences - a condensed dictionary

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	Де́за, Мише́л
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	



Problems with bibliographic data

Problem 4: Entity identification/resolution

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

How to deal with synonymy and homonymy:

- **Normalization** - at data entry
- **Standardization** - use standards whenever possible (**ORCID**, **DOI**, **ISBN**, **ISSN**, standard abbreviations **JAS**, **LTWA**, **WoS**, **Caltech**)
- **Dictionaries**
When the unit names are extracted from the text the so called stopwords are removed.
The equivalence is automatically determined using stemming or **lemmatization** – replacement by the canonical forms of words.
 - **lemmatization lists** (dictionaries)
 - **keywords** - keyword recommendations
- **Synonymy**: sort labels of units, manually/visually identify equivalent units, create partition, (shrink) equivalent units.
- **Homonymy**: correct the data in your copy of the data base.



Problems with bibliographic data

Problem 4: Entity identification/resolution

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

ISI names: The usual ISI name of a work (field CR in WoS)

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

has the following structure

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

In WoS the same work can have different ISI names. To improve the precision the program WoS2Pajek supports also short names. They have the format.

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

For example:

LEFKOVIT L(1985)70:585

From the last names with prefixes VAN, DE, . . . the space is deleted.

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

The best/final solution is to enter data in bibliographic data base in standardized way resolving homonyms.



Problems with bibliographic data

Problem 5: Incomplete data. Boundary Problem

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Incomplete data

- **missing data:** some parts of the units (works) descriptions which are not presented (additional authors, title, year, volume, issue, pages, publishers, etc.).
→ To add important missing parts of data manually
- **Missing data:** units (works) important for the studied topic which are not presented in the data set at all. In early phases different terminology was used.
→ To search for them - look at the most cited works from the references and include them into the analysis

For small bibliographies where we can inspect, accept and "correct" each entry. The "Excel table" approach is sufficient.

- when extracting subset of data
- preliminary citation network analysis; manually completing the important data



Problems with bibliographic data

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Some iterations are usually needed before the data set is “complete”!



Conversion from RIS to WoS

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic Data

Problems with Data

Tools

Books

```
ris <- readLines("Additional_RIS_DataBase.txt")
T <- ris[nchar(ris)>0]
writeLines(T,"RISC.txt")
```

```
T <- gsub("^J0","S0",
  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")
```



Tools for collection and maintenance of bibliographic data

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

Bibliographic tools:

- **JabRef** - open source bibliography reference manager
Bibliographic management tools
- **Bibliographic Conversion Tools**
- **Computer Science and Engineering: Bibliographic Tools**
- **12 Best Free Online Bibliography And Citation Tool**
- **Bibliographic Tools**
- **Bibliographic Software Overview**
- **Compare Some of the Popular Bibliographic Software Tools**
- **Bibexcel**
- **Text2Pajek**



Books and papers on bibliometrics

Bibliographic analysis

D. Maltseva,
V. Batagelj

Goals

Bibliographic
Data

Problems
with Data

Tools

Books

List of books and papers