ELEPHANT
IN THE LAB

HOW TO
SHORT ANALYSIS

How to analyse authorship

Short title	How to analyse authorship
Long title	Methodology for analysing the number of authors per article using meta data from Scopus.
Authors	Martin Schmidt [1], Benedikt Fecher [1], Christian Kobsda [1]
Author affiliation	[1] Knowledge Dimension, Alexander von Humboldt Institute for Internet and Society, Berlin, Germany
Author bios	Martin Schmidt is a doctoral researcher at the Institute of Landscape Systems Analysis within Leibniz Centre for Agricultural Landscape Research and associate researcher at Alexander von Humboldt Institute for Internet and Society. Benedikt Fecher is the programme director of the research programme Knowledge Dimension and heads the Open Science research group at the Alexander von Humboldt Institute for Internet and Society. Christian Kobsda works as the political consultant at the Leibniz Association and is an associate researcher at the Alexander von Humboldt Institute for Internet and Society.
Author social links	Martin Schmidt: ORCID – ResearchGate – Twitter Benedikt Fecher: ORCID – ResearchGate Christian Kobsa: ORCID – ResearchGate
Date published	10 July 2017
DOI	10.5281/zenodo.805718
Cite as (APA)	Schmidt, M., Fecher, B., Kobsda, C. (2017). Methodology for analysing the number of authors per article using meta data from Scopus. *Elephant in the lab*. DOI: 10.5281/zenodo.805718

Methodology

We conducted an *Advanced search* in Scopus – a bibliographic data base with more than 22.000 peer-reviewed journals (Scopus, 2017) – between 27th of February and 7th of March 2017. The

results of the search were defined by an algorithm (see below *Scopus Algorithm* for an example on the subject area *Agricultural and Biological Sciences* – acronym: AGRI) that includes

- a time period of published items beginning in 2010 to 2016 (inclusive)
- the document types (articles or reviews),
- and a quantitative limitation regarding the publication output (articles by the 20 authors with the most Scopus listed articles in every subject area).

The selection of articles ruled out overlaps in order to avoid that articles are analysed for more than one category. Please notice that the results are only representative for the select group of authors. It is possible that articles are reviews and *vice versa* at the same time.

The data was analysed for the 27 subject areas given by Scopus (2017b). Journals decide on their own to which subject areas they belong in the data base, so we had no influence on that and neither does Scopus. A search conducted later than ours, may result in more articles as some articles are added to Scopus after our data collection, also for earlier years. Further, the analyses were performed with original data with no changes expect the ones given in the *R code* (see below).

Statistical analyses were performed with R (R Core Team, 2017) using a word count function (see below *R Code*) and a package for quantitative analysis (Rinker, 2013). The XXXX is a proxy for the four characters code of the Scopus subject areas (Scopus, 2017b). As authors can add suffixes to their last names the word count function might result in slightly inaccurate results, although we already tried to exclude those cases which occurred. There is no guideline for that by Scopus.

Scopus Algorithm

DOCTYPE (re) OR DOCTYPE (ar)

AND SUBJAREA (AGRI)

AND (EXCLUDE (SUBJAREA , "BIOC ") OR EXCLUDE (SUBJAREA , "PHYS ") OR EXCLUDE (SUBJAREA , "MEDI ") OR EXCLUDE (SUBJAREA , "ENVI ") OR EXCLUDE (SUBJAREA , "EART ") OR EXCLUDE (SUBJAREA , "IMMU ") OR EXCLUDE (SUBJAREA , "CHEM ") OR EXCLUDE (SUBJAREA , "VETE ") OR EXCLUDE (SUBJAREA , "PHAR ") OR EXCLUDE (SUBJAREA , "NEUR ") OR EXCLUDE (SUBJAREA , "SOCI ") OR EXCLUDE (SUBJAREA , "ENGI ") OR EXCLUDE (SUBJAREA , "MATE ") OR EXCLUDE (SUBJAREA , "MATH ") OR EXCLUDE (SUBJAREA , "ECON ") OR EXCLUDE (SUBJAREA , "COMP ") OR EXCLUDE (SUBJAREA , "ARTS ") OR EXCLUDE (SUBJAREA , "CENG ") OR EXCLUDE (SUBJAREA , "ENER ") OR EXCLUDE (SUBJAREA , "DECI ") OR EXCLUDE (SUBJAREA , "HEAL ") OR EXCLUDE (SUBJAREA , "NURS ") OR EXCLUDE (SUBJAREA , "PSYC ") OR EXCLUDE (SUBJAREA , "BUSI ") OR EXCLUDE (SUBJAREA , "DENT ") OR EXCLUDE (SUBJAREA , "MULT "))

AND (LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2010))

AND (LIMIT-TO (AU-ID , "*NAME*" *AUTHOR-ID*) OR LIMIT-TO (AU-ID , "*NAME*" *AUTHOR-ID*) OR LIMIT-TO (AU-ID , "*NAME*" *AUTHOR-ID*) [... *and so on*]

R Code

```
library(ggplot2)
library(grid)
library(gridExtra)
library(qdap)
library(extrafont)

# Import data (XXXX)
XXXX <- read.csv("~/XXXX.csv", stringsAsFactors=F)

 # Delete name suffixes for comma separated counting
XXXX$Authors <- mgsub(c("Jr.","Sr.","I,","II,","III,","IV,","V,"),"", XXXX$Authors)

# Count number of authors by comma separation
AuthCountXXXX <- sapply(gregexpr(",", XXXX$Authors), length) + 1

 # Create dataframe
AuthCountPerYearXXXX <- data.frame(AuthCountXXXX)
rm(AuthCountXXXX)

 # Add year
AuthCountPerYearXXXX$YearXXXX <- as.numeric(XXXX$Year)

# Model for slope in AuthCount
fitXXXX <- lm(AuthCountPerYearXXXX$AuthCount ~ AuthCountPerYearXXXX$Year)

# Plot graph
 pPlot <- ggplot() + theme_bw() +
        theme(axis.text.x=element_text(colour="darkgrey"),
                axis.text.y=element_text(colour="darkgrey"),
                axis.ticks = element_line(colour = "lightgrey"),
                panel.background = element_rect(fill = NA),
                panel.grid.major.y = element_line(colour = "lightgrey"),
                panel.grid.major.x = element_blank(),
                panel.grid.minor  = element_blank(),
                panel.border = element_rect(colour = "lightgrey")) +
        coord_cartesian(ylim=c(-1,25)) + ylab("") + xlab("") +
        scale_x_continuous(breaks=c(2010,2011,2012,2013,2014,2015,2016)) +
        geom_boxplot(data=AuthCountPerYearAGRI,
        aes(x=YearAGRI, y=AuthCountAGRI, group=YearAGRI), na.rm = T) +
        stat_smooth(data=AuthCountPerYearAGRI,
        aes(x=YearAGRI, y=AuthCountAGRI), method = "lm", col = "#e4a50d")

        xaxis <- textGrob(label = "Year", hjust = 2, vjust =26, gp=gpar(fontfamily ="Lato-Regular"))
        yaxis <- textGrob(label = "Number of authors per article", hjust =-0.4, vjust=2.5,
                rot = 90, gp=gpar(fontfamily ="Lato-Regular"))
pAGRI <- grid.arrange(pPlot, right = xaxis, left = yaxis)
```

Acknowledgements

We want to thank Claus Dalchow and Jana Rumler for reviewing earlier drafts of our approach and methodology.

References

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rinker, T. W. (2013). qdap: Quantitative Discourse Analysis Package. 2.2.5. University at Buffalo. Buffalo, New York. http://github.com/trinker/qdap

Scopus. (2017a). *Content*. Elsevier. https://www.elsevier.com/solutions/scopus/content/ . Retrieved 2017-04-24.

Scopus. (2017b). *Access and use Support Center*. Elsevier. https://service.elsevier.com/app/answers/detail/a_id/11236/supporthub/scopus/#tips. Retrieved 2017-06-11.