Analyzing the use of email addresses in scholarly publications

Marc Luwel[*] and Nees Jan van Eck[**]

[*]*luwel@cwts.leidenuniv.nl*
https://orcid.org/0000-0003-3133-1737
Centre for Science and Technology Studies, Leiden University, The Netherlands

[**]*ecknjpvan@cwts.leidenuniv.nl*
https://orcid.org/0000-0001-8448-4521
Centre for Science and Technology Studies, Leiden University, The Netherlands

Due to recent fraud cases followed by massive retractions of papers, the authors' use of institutional versus non-institutional email addresses gained a lot of attention. A database was set up with all email addresses in the by-line of articles and reviews indexed in the Web of Science database.
In the period 2017-2021, the usage by corresponding authors of institutional email addresses is much more prolific in the Anglo-Saxon and Western European countries than in the BRICS countries. In the latter, corresponding authors use nearly as often a non-institutional as an institutional email address. The journals' publishing model does not seem of have a large impact on the type of email address used. Most strikingly is the much more frequent usage of non-institutional email addresses in retracted papers compared to non-retracted papers.

1. Introduction

Since its introduction in 70s of last century, 'electronic mail' became popular and journals started to include the email address of the corresponding authors in the publications' by-line (Wren, Grissom & Conway, 2006). Some journals do not only include the email address of the corresponding author but also the email address of other authors.

In the literature on the use of email address in scholarly publications, a distinction is made between two types of email addresses: institutional and non-institutional. An institutional email address is an address that requires a confirmed identity, such as a staff member of a university, an employee of a firm, or a public servant working in an administration. Non-institutional email addresses are email addresses that can be obtained by anyone from email service providers, such as Gmail, Outlook, Yahoo Mail, and iCloud Mail (Shen, Rousseau & Wang, 2018). In the literature, three topics have been studied:

- The decay of the use of institutional email addresses over time (Wren, Grissom & Conway, 2006);
- The relation between the use of a type of email address and the number of citations publications receive (Rodriguez-Esteban, Vishnyakova & Rinaldi, 2021);
- A possible relationship between the use of non-institutional email addresses and the retraction of publications (Liu & Chen, 2021).

Recently the use of non-institutional and doctored email addresses by paper mills, service providers that deliver paid assistance to write or fabricate papers and often also provide authorship slots and email addresses as well as manipulated peer review work, get considerable attention (Pinna, Clavel & Rocco, 2020; Wise, 2022). Paper mills and rogue peer review networks have led to massive retractions of papers (Kincaid, 2023; Petrou, 2023). Fake email addresses are one of the 'red flags' either to identify or to raise suspicion about paper mill practices (Abalkina & Bishop, 2022; Bishop, 2023). To the best of our knowledge, the analyses of the use of email addresses in publications are either limited to rather small samples or to a few publication years (Liu & Chen, 2021).

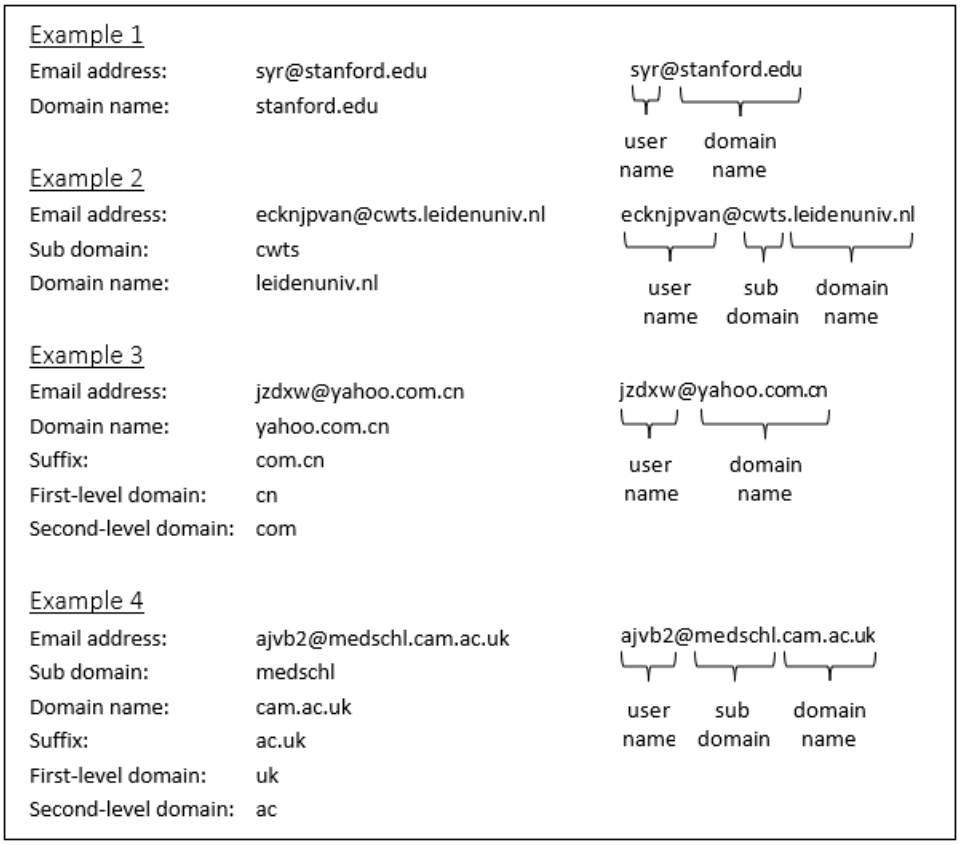
The objective of this study is the construction and analysis of a data set based on the email addresses mentioned in the by-line of publications in the journals indexed in the Web of Science (WoS). The section 'Data and Methods' describes the procedure to extract and enrich the data. In the next section a first exploratory analysis is made of its characteristics. In the conclusions these results are briefly discussed in relation to earlier work on the usage of email addresses and an overview of work in progress around the above-mentioned topics is outlined.

2. Data and methods

Data was collected from the WoS database. We used the in-house version of the WoS database available at the Centre for Science and Technology Studies (CWTS) at Leiden University. This version includes publications starting from 1980 in the Science Citation Index Expanded, the Social Sciences Citation Index, the Arts & Humanities Citation Index, and the Conference Proceedings Citation Index.

We first collected all email addresses available in the database. In a next step we tried to parse them and to identify the domain name. Parsing the email addresses and identifying the domain name is not a trivial task as is illustrated in Figure 1. Selecting everything after the @-sign is not accurate enough because of subdomains. Selecting the part before and after the last dot is also not accurate enough because of multi-part suffixes. By making use of the public suffix list[1] we were able to correctly identify the domain name of each email address. Out of the 664.7 thousand unique domain names we identified in this way, 11,608 turned out to be associated with more than 10 different email addresses and appeared in more than 100 publications. These 11,608 domain names covered 95% of all publications with an email address.

Figure 1: Parsing of email addresses and identification of domain names.



[1] https://publicsuffix.org/list/public_suffix_list.dat

Using a rule we then determined for each of the 11,608 domain names whether they could be linked to a scholarly organization. The rule-based approach we used to determine this relies on the Organization Enhanced affiliation data that is available in WoS. More specifically, for each combination of a domain name and an affiliated organization, we determined whether i) the number of publications in which this combination occurs is greater than 10, ii) the percentage of publications in which this combination occurs compared to the total number of publications of the affiliated organization is greater than 5%, and iii) the percentage of the publications in which this combination occurs compared to the total number of publications of the domain name is greater than 30%. If all three criteria were met, we linked the domain name to the organization and classified it as an institutional domain name. The idea of this rule-based approach is as follows. If a certain email domain name appears in the publications of authors from many different organizations, this is an indication that it concerns a domain name of an email service provider. If a certain email domain name is only used in publications by authors of the same organization or a limited number of sub organizations, this is an indication that it concerns a domain name of a scholarly organization.

After applying the rule-based approach, we manually validated and corrected the domain name assignments using Website Informer[2], an online tool by Informer Technologies, Inc. that can be used to gather detailed information on domain names. We used the free version of the tool to collect information on the domain name and its owner. Where no or not sufficient information was obtained using the tool, the Microsoft Edge browser was used to further determine the organization associated with a domain name.

First, the robustness of the rule-based assignments of domain names to scholarly organizations was manually tested. These assignments turned out to be very accurate. Based on a 5% random sample, only 7 anomalies in the linked domain names to organizations were detected. We then focused on domain names that were not assigned to an organization. It turned out that the rule-based approach was less accurate in this case. We therefore decided to check all unassigned domain names and tried to assign them manually to a scholarly organization or an email service provider. Only 164 (1.4%) could not be assigned, representing only 0.2% of the total number of email addresses linked to all of the 11.608 domain names taken into account.

After applying the rule-based approach and the manual correction, 10.838 domain names were assigned to a scholarly organization and 606 domain names to an email service provider. Based on these assigned domain names, we finally classified all email addresses in the WoS. Email addresses linked to a domain name of a scholarly organization were classified as institutional and email addresses linked to a domain name of an email service provided were classified as non-institutional. All other email addresses were classified as unknown.

3. Results

3.1. Authorships with email addresses over time

In this section, changes over time in the availability of email addresses in publications in WoS are analyzed. Only publications classified as ‘article’ and ‘review’ and published in the period 2004-2021 were taken into account, representing 26.3 million publications and 136.7 million authorships (i.e., publication-author combinations).

[2] https://website.informer.com

Figure 2 shows the gradual increase of the average number of authors per publication from 4 in 2004 to 6 in 2021. As could be expected, nearly all publications have a corresponding author, referred to in the WoS as the reprint author. During this period, the number of authors with an email address doubled. In 2021, an email address is available in WoS for 1 in 3 authors of a publication.

Figure 2: Average number of authors per publication.

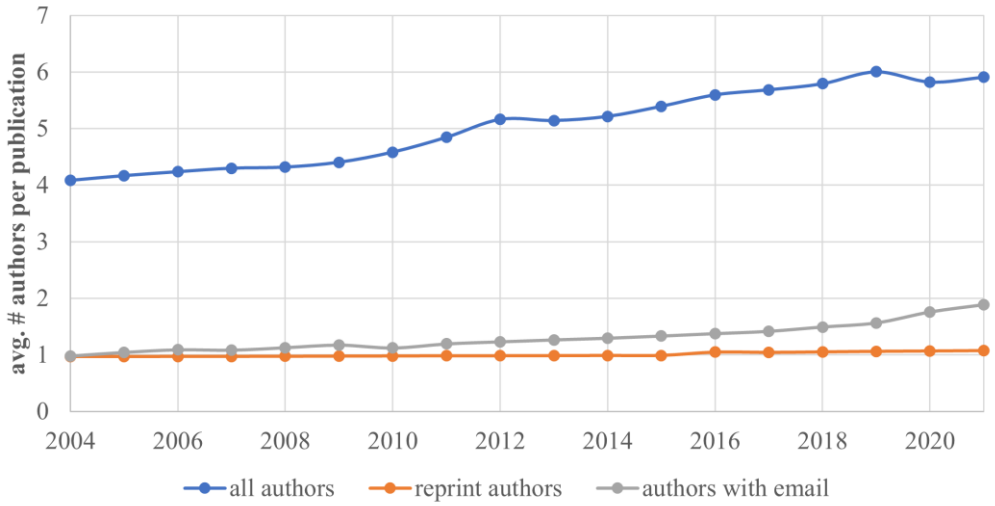
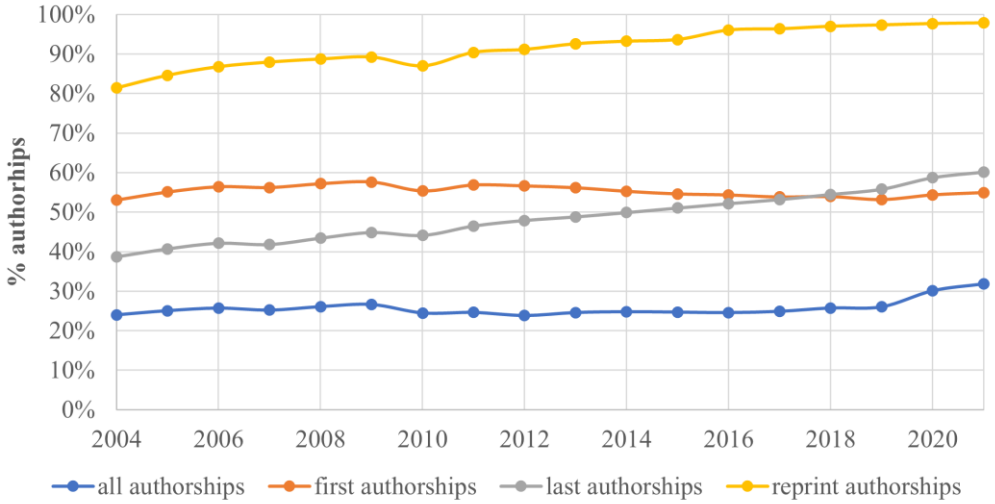


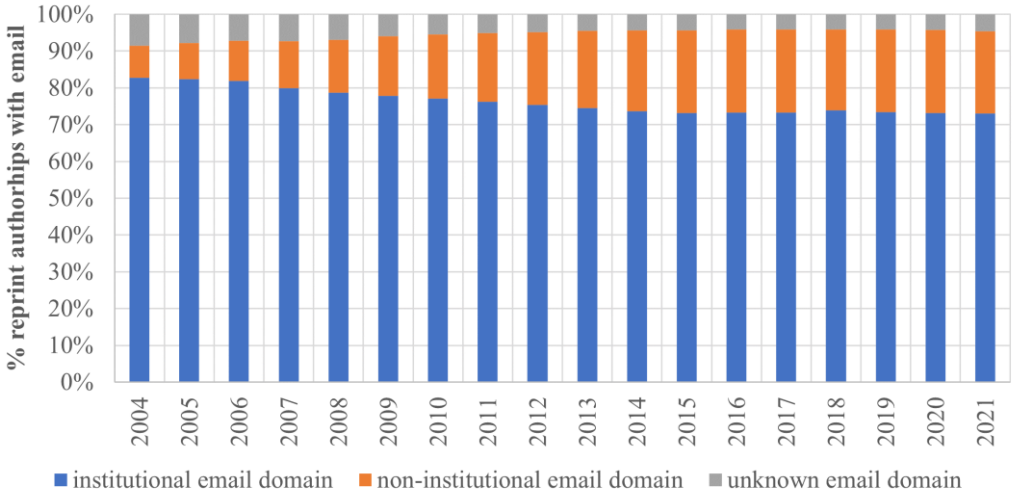
Figure 3 shows for the same period that the percentage of reprint authorships has increased from 80% in 2004 to almost 100% in 2021. In 2021, in 60% of the publications the last authors' email address is available, slightly more than that of the first author. After remaining around 25%, in the last two years the percentage of authorships with an email address has increased to slightly above 30%.

Figure 3: Availability of email addresses for different types of authorships.



In the period 2004-2021, the use of institutional email addresses by reprint authors has decreased by 10% (Figure 4). In 2021, about 22% use a non-institutional email address. The share of email addresses of reprint authors that could not be linked to one of the two categories has decreased slightly between 2004 and 2010 to about 5% for the more recent years.

Figure 4: Usage of institutional and non-institutional email addresses by reprint authors.



4.2. Usage of institutional and non-institutional email addresses by reprint authors

In this section, the usage of institutional and non-institutional email addresses by reprint authors in the 9.8 million publications from the period 2017-2021 is analyzed in more detail. In this period, an email address is available for 97% of the 10.4 million reprint authorships.From the reprint authorships with an email address, 73% are identified as institutional, 22% as non-institutional, and for 4% of the authorships our approach was unable to determine one of these two categories.

For the non-institutional email addresses used by reprint authors, Table 1 provides an overview of the most frequently used email domain names. This list is dominated by three global players (Google, Yahoo. and Microsoft), and by Chinese and to a lesser extent Russian email service providers.

Table 1: Most frequently used domain names of non-institutional email addresses used by reprint authors.

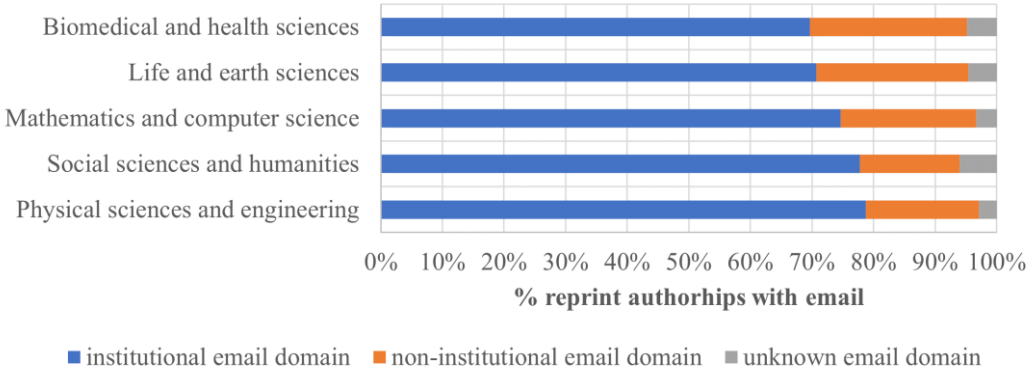
domain name	**owner (country / headquarters)**	**# reprint authorships**
gmail.com	Google LLC (USA)	871,498
163.com	NetEase, Inc. (China)	411,255
126.com	Guangzhou NetEase Computer System Co., Ltd (China)	186,522
yahoo.com	Yahoo! Inc. (USA) *	164,266
hotmail.com	Microsoft Corp. (USA)	156,147
sina.com	Sina (China)	54,320
qq.com	Shenzhen Tencent Computer System Co., Ltd (China)	48,488
mail.ru	VK Company Limited (Russia)	36,536
aliyun.com	Alibaba Cloud Computing Ltd. (China)	18,947
outlook.com	Microsoft Corp. (USA)	18,456
yahoo.fr	Yahoo EMEA LIMITED (Ireland) **	17,611
yandex.ru	Yandex LLC (Russia)	16,660
yahoo.co.in	Verizon Media Inc. (USA)	13,673
yahoo.com.br	Yahoo do Brasil Internet Ltda. (Brasil) **	13,493

rediffmail.com	Rediff.com India Ltd. (India)	11,703
yeah.net	Guangzhou NetEase Computer System Co., Ltd (China)	10,983
foxmail.com	Shenzhen Tencent Computer System Co., Ltd (China)	10,612
sohu.com	Beijing Sohu Internet Information Service Co., Ltd. (China)	6,640
hanmail.net	Kakao Corp. (South Korea)	6,112
yahoo.co.jp	Yahoo Japan Corporation (Japan)	6,028
yahoo.co.uk	Yahoo! Inc. (USA) *	5,980
naver.com	NAVER Corp. (South Korea)	5,652
263.net	net263 co., ltd (China)	5,563
aol.net	Yahoo! Inc. (USA) *	4,386
libero.it	Italiaonline S.p.A. (Italy)	4,082

* owned by Apollo Global Management (90%) and Verizon Communications (10%)
** (partially) owned by Verizon Communications

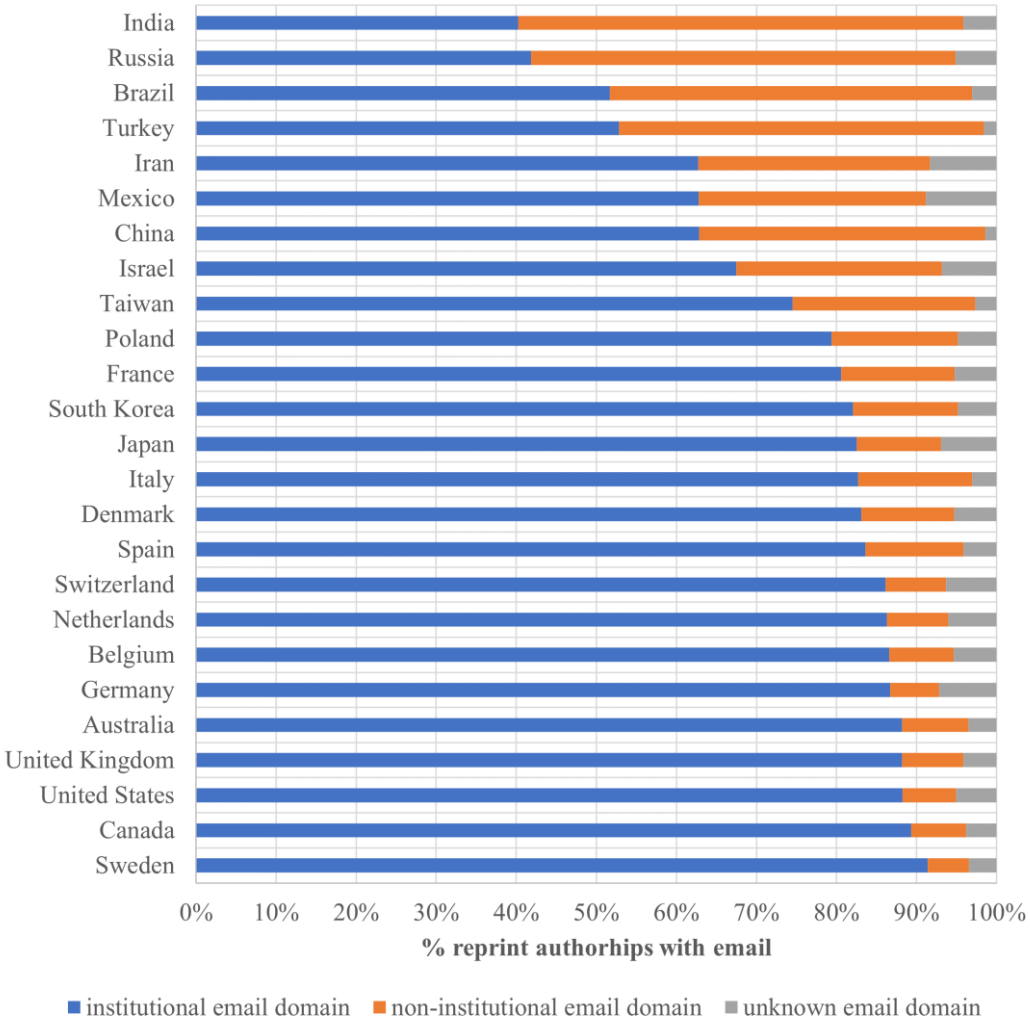
Figure 5 shows the differences between disciplines. About 79% of reprint authors in the physical and engineering sciences use an institutional email address, while in the biomedical and health sciences this is about 70%.

Figure 5: Usage of institutional and non-institutional email addresses by reprint authors for each field.



The difference in the usage of institutional and non-institutional email addresses by reprint authors is even larger between countries (Figure 6). Of the top 25 countries with the most reprint authorships, the share of institutional email addresses varies from about 90% for countries like Sweden, Canada, the United States, and the United Kingdom to about 40% for India and Russia. This enormous difference cannot be explained by the share of email addresses that could not be assigned to a scholarly organization or an email service provider: for all these countries this information is missing for about 5% of the reprint authorships with an email address.

Figure 6: Usage of institutional and non-institutional email addresses by reprint authors for the top 25 countries with most reprint authorships.



Looking at the publications' language, major differences are also found between the use of institutional and non-institutional email addresses by reprint authors (Figure 7). For publications in the Russian and Portuguese languages, the use of non-institutional email addresses is dominating.

Almost as pronounced as the differences between countries are those between publishers (Figure 8). Among the top 20 publishers, we see three American and one British learned society where 85% or more of the reprint authorships have an institutional email address. On the other end of the spectrum, in publications from the Hindawi Publishing Group, acquired by John Wiley & Sons in 2021, reprint authors are equally likely to use institutional as non-institutional email addresses.

Figure 7: Usage of institutional and non-institutional email addresses by reprint authors for the top 10 languages with most reprint authorships.

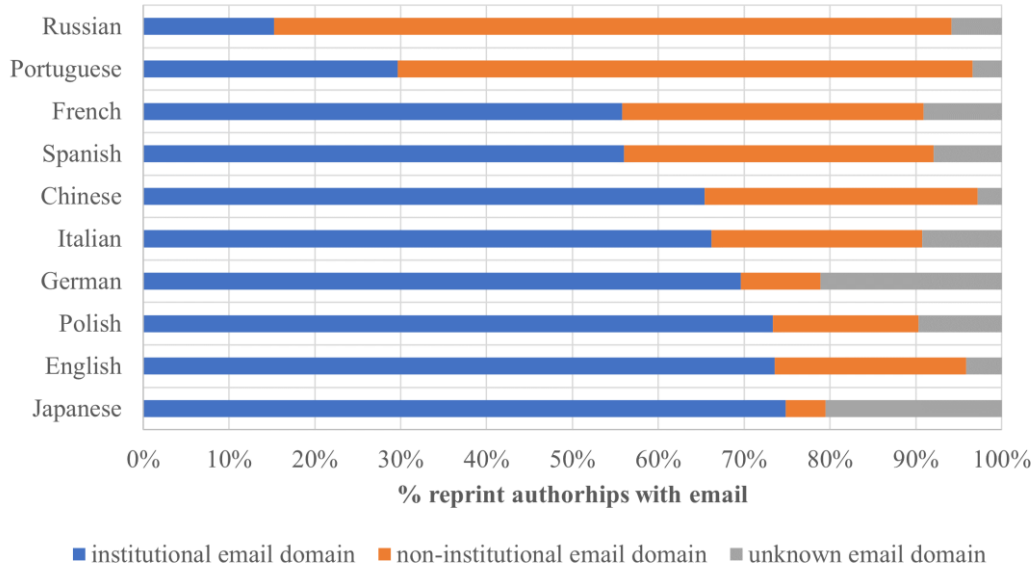


Figure 8: Usage of institutional and non-institutional email addresses by reprint authors for the top 20 publishers with most reprint authorships.

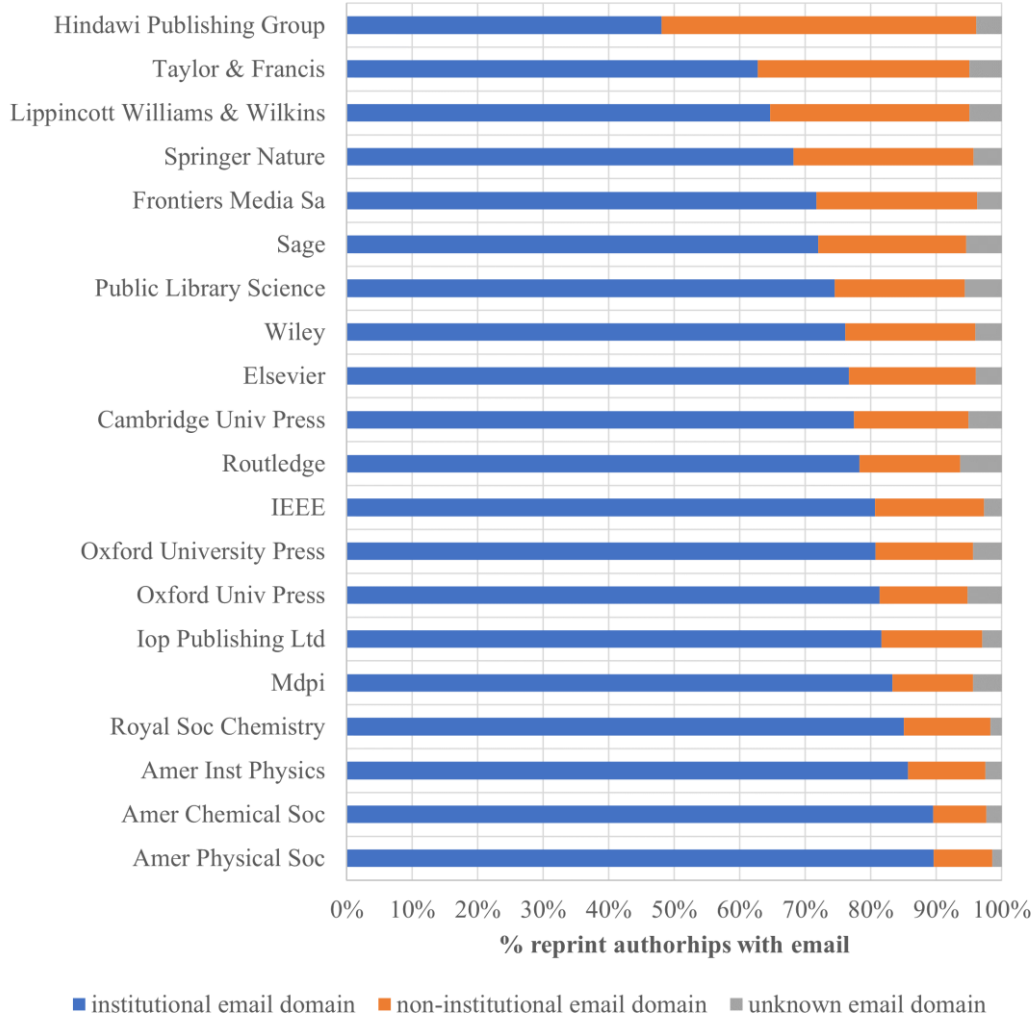


Figure 9 shows that between gold open access journals and journals that use another publishing model there is almost no difference in the share of reprint authorships with an institutional (70-74%) versus a non-institutional email address (21-25%).

Figure 9: Usage of institutional and non-institutional email addresses by reprint authors in gold open access and non-gold open access journals.

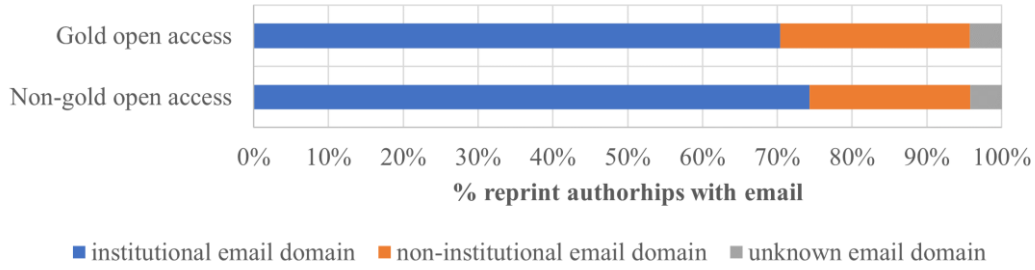
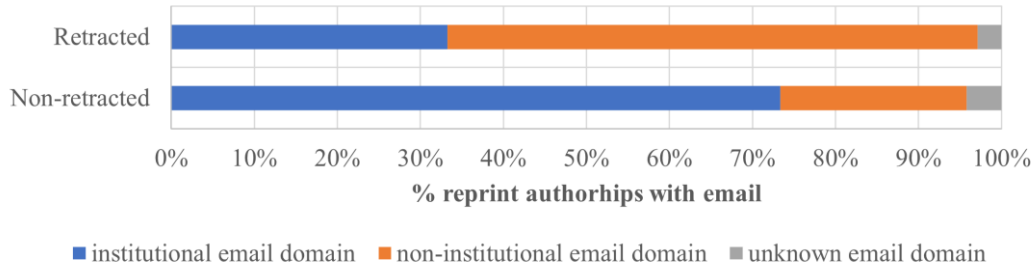


Figure 10 sheds some light on the ongoing controversy about the use of institutional versus non-institutional email addresses in retracted papers. For retracted papers, the share of non-institutional email addresses used by reprint authors is more than twice as large as for non-retracted publications.

Figure 10: Usage of institutional and non-institutional email addresses by reprint authors in retracted and non-retracted publications.



5. Conclusion

A database was set up with all email addresses included in the by-line of articles and reviews indexed in the WoS. For about 95% of the authorships with an email address, the associated domain name could be linked to either a scholarly organization or an email service provider.

Email addresses have gradually become available for the publications in the WoS. Today, an email address is available for one in three authors. Striking is the increase of more than 10% in the use of non-institutional email addresses by corresponding authors in the period 2004-2021. A similar trend was observed by Kozak et al. (2015) in their analysis of the usage of email addresses in the 1,000 most-cited and the 1,000 least-cited articles published in 2000, 2005 and 2010.

Limiting the analysis to the period 2017-2021, where the email address of more than 97% of the reprint authorships is available, three results stand out:

- The usage of institutional email addresses is much more prolific in the Anglo-Saxon and Western European countries than in the BRICS countries. In the latter the corresponding authors use nearly as often a non-institutional as an institutional e-mail

address. These results are in line with those obtained by Shen, Rousseau, & Wang (2018) and Rousseau (2018) for the period 2008-2012.
- The journals' publishing model does not seem of have a large impact on the type of email used.
- Most striking is the large difference in the use of non-institutional email addresses by corresponding authors of retracted versus non-retracted papers. The usage of institutional versus non-institutional email addresses by corresponding authors and authors of retracted publications has been studied by several authors (see Liu & Chen (2021) and references therein). Compared to the results presented in this paper, these studies are limited to case studies or small samples of retracted publications.

In ongoing work, a more comprehensive analysis of the relationship between the usage of institutional versus non-institutional email addresses and retractions will be made, considering the authors' country, the research field, and the reason for the retraction. As is often overlooked, not all retractions are due to fraud or other misconduct by one or more authors.

Another unsettled research question is the relationship between the use of the two types of email addresses and the number of citations received. According to Kozak et al. (2015), there is no influence on the number of citations, contrary to Shen, Rousseau, & Wang (2018) who conclude that publications with institutional email addresses tend to be cited more. Using the database an elaborated analysis of citation rates is planned.

Open science practices

The classification of email domain names created and used in this paper has been made available in Zenodo (Luwel and Van Eck, 2023). The data used to create this classification and the data underlying the analysis of this paper were obtained from the WoS database produced by Clarivate Analytics. Due to license restrictions, this data cannot be made openly available. The source code used in the analysis of this paper is available in the following GitHub repository: https://github.com/neesjanvaneck/WoS-email-address-analysis.

Acknowledgments

We would like to thank Bram van den Boomen for his contribution to the parsing and identification of domain names in email addresses.

Author contributions

Marc Luwel: Conceptualization, Data curation, Investigation, Formal analysis, Methodology, Writing—original draft. Nees Jan van Eck: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing—review & editing.

Competing interests

The authors have no competing interest.

Funding information

The authors did not receive any funding for this research.

References

Abalkina, A., & Bishop, D.V.M. (2022, September 5). Paper mills: A novel form of publishing malpractice affecting psychology. *PsyArXiv*. https://doi.org/10.31234/osf.io/2yf8z

Bishop, D.V.M. (2023, February 6). Red flags for paper mills need to go beyond the level of individual articles: a case study of Hindawi special issues. *PsyArXiv*. https://doi.org/10.31234/osf.io/6mbgv

Kincaid, E. (2023). Wiley and Hindawi to retract 1,200 more papers for compromised peer review. *Retraction Watch*. https://retractionwatch.com/2023/04/05/wiley-and-hindawi-to-retract-1200-more-papers-for-compromised-peer-review/

Kozak, M., Iefremova, O., Szkoła, J., & Sas, D. (2015). Do researchers provide public or institutional E-mail accounts as correspondence E-mails in scientific articles? *Journal of the Association for Information Science and Technology*, *66*(10), 2149-2154. https://doi.org/10.1002/asi.23401

Liu, X., & Chen, X. (2021). Authors' noninstitutional emails and their correlation with retraction. *Journal of the Association for Information Science and Technology*, *72*(4), 473-477. https://doi.org/10.1002/asi.24419

Luwel, M., & Van Eck, N.J. (2023). Classification of domain names of scholarly email addresses [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.7851620.

Pinna, N., Clavel, G., & Roco, M. C. (2020). The Journal of Nanoparticle Research victim of an organized rogue editor network! *Journal of Nanoparticle Research*, *22*, 376. https://doi.org/10.1007/s11051-020-05094-0

Rodriguez-Esteban, R., Vishnyakova, D., & Rinaldi, F. (2022). Revisiting the decay of scientific email addresses. *Journal of the Association for Information Science and Technology*, *73*(1), 136-139. https://doi.org/10.1002/asi.24545

Rousseau, R. (2018). Institutional versus commercial email addresses: which one to use in your publications? *LSE Impact Blog*. https://blogs.lse.ac.uk/impactofsocialsciences/2018/06/21/institutional-versus-commercial-email-addresses-which-one-to-use-in-your-publications.

Shen, S., Rousseau, R., & Wang, D. (2018). Do papers with an institutional e-mail address receive more citations than those with a non-institutional one? *Scientometrics*, *115*, 1039-1050. https://doi.org/10.1007/s11192-018-2691-0

Wise, N. (2022). What is going on in Hindawi special issues? *BishopBlog*. http://deevybee.blogspot.com/2022/10/what-is-going-on-in-hindawi-special.html

Wren, J.D., Grissom, J.E., & Conway, T. (2006). Email decay rates among corresponding authors in MEDLINE. *EMBO Reports*, *7*(2), 122-127. https://doi.org/10.1038/sj.embor.7400631