

Transitioning the Information Retrieval Literature to a Fully Open Access Model

Djoerd Hiemstra¹, Marie-Francine Moens², Raffaele Perego³, Fabrizio Sebastiani³

¹ Radboud University, Nijmegen, The Netherlands

² KU Leuven, Belgium

³ Istituto di Scienza e Tecnologie dell'Informazione,
Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy

*djoerd.hiemstra@ru.nl, marie-francine.moens@cs.kuleuven.be,
raffaele.perego@isti.cnr.it, fabrizio.sebastiani@isti.cnr.it*

Abstract

Almost all of the important literature on Information Retrieval (IR) is published in subscription-based journals and digital libraries. We argue that the lack of open access publishing in IR is seriously hampering progress and inclusiveness of the field. We propose that the IR community starts working on a road map for transitioning the IR literature to a fully, “diamond”, open access model.

1 Introduction

Thirty years ago when Tim Berners-Lee invented the World Wide Web at CERN, his main goal was to enable and promote the free sharing of scientific information. Ironically, while the Web disrupted many industries, most scientific information is not shared freely on-line. Papers in many fields are still mostly published in subscription-based journals and hosted by subscription-based digital libraries. Before the World Wide Web, subscription-based commercial models of publishing provided an important service to the research community: they made sure that research papers were shared as rapidly and widely as possible. The Web changed all that thirty years ago. The Web enables the direct, free, unrestricted availability of information at a fraction of the cost of printed dissemination of results. The needs of the research community are no longer served well by subscription-based (semi-)commercial publishing. By locking papers up behind paywalls, we exclude researchers who work at organizations that cannot afford the subscription fees, thereby limiting our communities to the “global north” and a few new emerging economies, as well as seriously limiting the

potential impact of research results as measured by the number of downloads and citations that research papers get.

Unfortunately, Information Retrieval (IR) – where our passion lies – is an example of a research field that almost exclusively publishes in subscription-based journals and libraries. The following IR journals and proceedings are only available via subscriptions: The Association for Computing Machinery (ACM) publishes the journal Transactions on Information Systems, and the conference proceedings of SIGIR, CIKM, WSDM, CHIIR, ICTIR, and FIRE. Springer publishes the Information Retrieval Journal, and the proceedings of ECIR, SPIRE, and CLEF. Elsevier publishes Information Processing and Management (IP&M), and Wiley publishes the Journal of the Association for Information Science and Technology (JASIST). These publications cover almost the entire field of IR. Maybe the most notable IR proceedings series that *are* published fully open access are the proceedings of the evaluation conferences TREC, NTCIR and CLEF (the CLEF working note papers). The papers of these proceedings are, however, not peer-reviewed.

We believe the current situation of closed access and hybrid open access publishing in IR is seriously hampering progress in the field. In the next section we make a case for a fully, “diamond” open access publishing model, and in Section 3 we describe steps that the IR community should consider in order to transition the IR literature to such a publishing model. Open access publishing is the future of science, and IR as a field still has a long way to go.

2 The Case for a Fully Open Access Model

The term *open access* was coined in 2012 by the Budapest Open Access Initiative¹ to refer to the *free and unrestricted online availability* of peer-reviewed scientific literature achieved by *removing the barriers, especially the price barriers* to its world-wide electronic distribution. In the following subsections we define open access in line with the Budapest Initiative, describe the benefits of open access and give examples of communities that transitioned successfully to a fully open access model.

2.1 Defining full open access

Since the Budapest Open Access Initiative, the discussions on open access (OA) have been blurred by definition questions and curious colouring schemes that distinguish for instance green OA [7], gold OA [13], bronze OA [21], black OA [6], and diamond OA [19]. We do not intend to repeat those discussions in this position paper. What we mean by a Fully Open Access model is the following:

Full Open Access is a model for free and unrestricted online availability of peer-reviewed scientific literature where the author does not pay, the reader does not pay, and the entire mechanism is self-funded, running on the volunteer work by editors, reviewers, technicians, admins, and on micro-donations by friend organizations such as universities and research centers.

¹<https://www.budapestopenaccessinitiative.org/read>

Fully Open Access corresponds to diamond open access in the colouring schemes above. A Fully Open Access publishing model builds on the long tradition of scientists and academics to publish papers, and to review and edit their peers' papers, without asking any direct payment in return.

2.2 A more inclusive IR community

Access to scientific information is a human right. Article 27 of the United Nations declaration of human rights states that *"Everyone has the right to freely participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits."*² Before the World Wide Web, subscription-based scientific publishing models were good models for sharing scientific advancement. But with the Web in place, subscription-based models limit the world-wide sharing of information more than they enable it, thereby effectively violating Article 27 of the United Nations declaration.

Subscription-based journals contribute to the North-South divide, i.e., to the difference in the socio-economic status of the earth's global north as compared to the global south [24]. Fair opportunities for knowledge access are not granted to the researchers from the global south and from other resource-poor countries because they lack the financial resources to read our literature [1]. As a result, some of them cannot fully participate in the IR community or lack the information and expertise to even meaningfully engage with it. Guillermo Paraje et al. [20] found that less than 2% of the publications they investigated (based on the Science Citation Index and the Social Sciences Citation Index from 1992 to 2001) were written by researchers from low-income countries, and that rate was still declining. Only one-fifth of those papers originated from sub-Saharan Africa.

The impact of a fully open access publishing model on IR has benefits far beyond our research community. Access to scientific literature is an important driver of innovative start-up companies and research and development in the private sector [4]. It will advance possibilities for citizen science [8], and enable more coverage of our work in the press and on social networks [28]. A fully open access model for the IR literature will benefit the public interest and society as a whole.

2.3 More IR citations and downloads

Having our papers freely accessible online results in more citations and more impact for our research results. Already in 2001, Steve Lawrence [16] found that computer science publications that are freely available on the Web, received on average 2.6 to 4.5 times as many citations as publications that were not freely available. These results are confirmed for several other disciplines by, e.g., Antelman [3], Davis et al. [10], and Donovan [11], who found that across a variety of disciplines, open-access articles have a greater research impact in citations and downloads than articles that are not freely available. Even commercial publishers like Springer [12] report that in hybrid open access journals, the open access articles are cited and accessed more often.

²<https://www.un.org/en/universal-declaration-human-rights/>

2.4 Research communities that thrive by full open access

From the very beginning of the World Wide Web, some journals have made the transition to a fully open access publishing model. The earliest example we could find is the Journal of Clinical Investigation (JCI). Its editor, Ajit Varki [27], writes in the editorial of 1 January 1996: *The nonprofit nature of the JCI allows consideration of a truly novel solution – not to charge anyone at all!* Let's have a look at some best-practices closer to home: The communities of Computational Linguistics, Computing Systems and Machine Learning.

Open access in computational linguistics

The Association for Computational Linguistics (ACL) publishes all its journals, conference and workshop papers fully open access in the ACL Anthology. The ACL Anthology was launched in 2002 and is run and maintained completely by community volunteers, coordinated by the ACL Anthology editor. Hosting and bandwidth is provided by the National University of Singapore and Saarland University free of charge [14]. The minimal costs that still have to be incurred for publishing are covered by the ACL membership fees and conference registration fees.

Open access in computing systems

The USENIX Advanced Computing Systems Association removed its paywalls in 2008, providing conference proceedings free of charge [15]. USENIX too relies on volunteers for running its publication service. USENIX uses sponsor support and revenue from memberships and conferences to cover its publication costs. Authors of papers in USENIX proceedings retain the copyright of their work. USENIX also used external partners, like PeerJ Computer Science, to enable open access peer-reviewed publishing for USENIX members and USENIX conference delegates.

Open access in machine learning

Important conferences in machine learning, such as NeurIPS, ICML, and ICLR all have their publications available online. The move to open access in machine learning started in the spring of 2000, when the Journal of Machine Learning Research (JMLR) was founded. The JMLR is completely run by volunteers. Its editorial board has members with the usual job descriptions such as the editor-in-chief and the managing editor, but also production staff members and a web master. Interestingly, the initial editorial board of JMLR consisted of forty members of the former editorial board of the Machine Learning Journal published by Kluwer (now Springer), who resigned to support open access publishing in machine learning. Their resignation letter was published by SIGIR Forum [5] and convinced some IR researchers to only volunteer for open access journals [17].

Discussion

The three research communities above, Computational Linguistics, Computing Systems, and Machine Learning have grown more in size than the IR community in the past 30 years. Of course, the growth of other communities cannot be explained completely by their adoption of

open access, but we do believe that the effects of increasing downloads, increasing citations and a more inclusive community accumulate over the years. It is a worrying sign that the IR community has not really grown substantially after web search became *the* killer application on the World Wide Web in the end of the 1990's. We have missed good opportunities as a community, and will continue to miss them if we continue on the current path of subscription-based publishing.

2.5 Alternative open access models that do not work

Many alternative open access models lead to journals and digital libraries having part of their papers freely available, sometimes called hybrid open access. Such models are still unfair for researchers of institutions that cannot afford the fees for accessing the remaining papers. We also do not want alternative open access models in which the cost is simply shifted from one player to the other, for instance models in which the authors pay instead of the readers. Requiring authors to pay also effectively excludes authors from our community that cannot afford the article processing fees. Let's enumerate some of the open access approaches that do not work:

- Relying on author's personal home pages ("green" open access) results in a substantial amount of papers that are not available freely, for instance because the authors do not bother, because authors switch jobs and leave academia, or because they die.
- International agreements like Springer Compact, which allows researchers from among others Austria, the Netherlands, and the UK to publish open access in Springer's Information Retrieval Journal, lead to papers of researchers from countries that do not have a solid negotiation position to be left out.
- Open choice publishing ("gold" open access), which allows authors of papers of subscription-based journals like IP&M³ or proceedings like SIGIR⁴ and ECIR⁵ to pay a per-article cost for making their paper open access also naturally leads to authors being excluded that cannot afford those fees.
- For profit Open Access Journals like PLoS ONE⁶ charge authors high paper processing fees, thereby excluding many researchers whose institutions cannot afford those costs.
- Open access by "open table of contents" (open TOC) as provided by the ACM⁷ digital library limits access to the selection of papers that have an open TOC, and it further restricts access to those papers because the paywalled papers – but not the open TOCs – are indexed by the main scientific search engines like Google Scholar and ScienceDirect.

³Elsevier charges \$2670 excluding taxes for an IP&M paper, see: <https://www.elsevier.com/journals/information-processing-and-management/0306-4573/open-access-options>

⁴ACM charges \$700 to \$1,300 for ACM members and \$900 to \$1,700 for non-members to publish a SIGIR paper open access.

⁵Springer charges €38 per page for ECIR papers which corresponds to €532 for a 14 page full paper, see: <https://www.springer.com/gp/computer-science/lncs/open-access-publishing-in-computer-proceedings>

⁶PLoS ONE has a questionable acceptance rate of about 69% [23] and charges \$1695 per paper, see: <https://www.plos.org/publication-fees>

⁷ACM allows readers to access some papers via special links that act like a backdoor to their paywalled digital library, see: <https://www.acm.org/publications/openaccess#opentoc>

We believe the efforts above are – rather successful – attempts of the traditional publishers to keep the status-quo, or worse, to profit even more from the current transition period by charging both authors to publish papers and charging libraries and readers to access the papers. The revenues of major commercial publishers are enormous [9] and all this profit is paid by players in the value chain of research (i.e., authors and/or readers and/or their institutions), which stifles research [22]. It is time for the IR community to seriously discuss and implement change.

3 Important Steps Towards Open Access

Rather than proposing a solution, the main goal of this opinion paper is to involve the IR community in devising the best solution. We hope this paper starts a discussion on open access publishing that is taken up inside our institutions, on social media, and at IR conference gatherings, both during the coffee breaks as in conference panels. In the following sections we suggest several more concrete actions.

3.1 The future of the ACM: a community funded model

Of the publishers of IR literature mentioned in the introduction, the ACM is in the best position to adopt a fully open access model, because it already considers “member-driven and volunteer-led” activities as its core values.⁸ Like the ACL and USENIX, the ACM is a scholarly and professional computing society, serving its members (instead of for instance company shareholders). Furthermore, a switch to a fully open access publishing model would affect more publications in IR than the switch of any other publisher.

Unfortunately, the ACM has done disappointingly little to move towards an open access publishing model. At its best, ACM calls its publishing “clopen”, somewhere between closed and open [25]. At its worst, the ACM publicly denounces open access by signing the letter of the Association of American Publishers (AAP) to the US administration in which it “*expressed their strong opposition to a proposed Administration policy that would mandate immediate free distribution of peer-reviewed journal articles reporting on federally funded research*” [18]. The ACM is on route to sustain revenue of its (profitable) digital library at the cost of its members, and to implement a hybrid open access model that shifts most of the costs from its readers to its authors [2, 26].

Our opinion paper is a call to all ACM members to use their influence to discuss with the ACM the need to transition to a fully open access publishing model. To push the discussion in the right direction, SIGIR may already allocate part of its budgets to cover the cost of opening up the IR literature. We should not forget that for most of its history, SIGIR budgeted the costs of *printed* proceedings for its conferences. It is time to budget such publication costs again to cover open access publishing. We realize that open access publishing does not come for free, but as a community, we are able to bear open access publishing costs without charging our readers or our authors, possibly at the expense of other community activities [29]. SIGIR may urge the ACM to consider alternative business models for society publishers [30] instead of article publication charges.

⁸<https://www.acm.org/about-acm/about-the-acm-organization>

3.2 The future of for-profit publishers like Springer

After the ACM, Springer is the second biggest publisher of the IR literature, publishing the Information Retrieval Journal as well as the proceedings of ECIR, SPIRE, and CLEF. Like many publishers, Springer has moved small steps towards open access. For instance, Springer made the proceedings of recent ECIR editions accessible during the conference at no extra costs, but closed access to them after the conference finished.

ECIR 2020 made an important step this year, by offering Springer to pay the LNCS series open access fee of €30 per page. For the ECIR proceedings, this amounts to a total cost of almost €50,000 for the full proceedings. Such costs are normally too much for the community to bear (ECIR is organized under the auspices of the British Computer Society), but could be paid this year when ECIR 2020 went online because of the global coronavirus pandemic lock-down that made organizing costs drop dramatically.

We are committed to keeping the ECIR proceedings open access for the 2021 edition and the years to come.⁹ Possible approaches are: Finding a sponsor to cover the Springer open access publishing costs; Negotiating a substantial discount from Springer; Raising the conference registration fees; or Finding alternative publishers, such as the BCS Workshop Proceedings (that published the ECIR proceedings before 2002), the CEUR Workshop Proceedings, or arXiv (possibly as an overlay journal). ECIR is now a CORE A conference.¹⁰ Open access proceedings would probably reinforce this status, due to open publications being more cited. Short time effects of changing publisher are less clear. Any solution should be such that indexing systems that currently index the ECIR Proceedings keep indexing it.

3.3 New fully open access initiatives

Alternatives like arXiv host a large number of publications, most of which do not have the peer reviewing and quality control of the IR literature in place. It is up to the IR community to think of alternatives like the following:

- *An IR Anthology*, modelled on the ACL Anthology, would provide an alternative to the commercial publishers like Elsevier and Springer. It may also host the “open table of contents” of the ACM digital library, at least as long as the ACM’s transition to a fully open access publishing model is not yet complete. SIGCHI provides an example of what this may look like for SIGIR.¹¹
- *An external open access partner*, modelled on the USENIX–PeerJ partnership where the community provides quality control such as peer reviewing and the partner provides professional publishing of the proceedings including the assignment of ISBN and ISSN numbers, DOIs for each paper, and indexing by the ACM digital library, DBLP, Scopus, the Web of Science databases, Google Scholar, etc.
- *Founding a new open access journal in IR*, modelled on the Journal of Machine Learning Research (JMLR) where community members provide the usual volunteers like peer reviewers and editors, but also production staff and a web master.

⁹The authors are the general chairs and program chairs of ECIR 2021.

¹⁰<http://www.core.edu.au/>

¹¹<https://sigchi.org/conferences/conference-proceedings/>

4 Conclusion

Transitioning the IR literature to a fully open access publishing model is of the utmost importance for progress in the field. Open access papers get more citations and downloads, and they bring researchers to the IR community that are now excluded because they cannot afford the costs of subscriptions or the costs of article processing fees. We discussed communities in Computational Linguistics, Computing Systems, and Machine Learning that transitioned to open access before us, and noted that they have grown much bigger than IR since. At this point in time, IR finds itself in a position where its main publisher, the ACM, seems unwilling to embrace full open access soon, even though it fully aligns with ACM's member-driven and volunteer-led activities. The other publishers' commercial approaches make it unlikely that they will support this transition, so our main hope for immediate change are initiatives that found new journals or proceedings series. This opinion paper intends to kick-off a discussion that involves the IR community at large in devising the best solution.

Acknowledgments Many thanks to João Magalhães for insightful discussions and comments that let to last-minute changes of this paper.

References

- [1] Joanna Adcock and Edward Fottre. The north-south information highway: case studies of publication access among health researchers in resource-poor countries. *Global Health Action*, 1(1), 2008.
 - [2] Rick Anderson. ACM's new open access agreements: A Q&A with Scott Delman. *Scholarly Kitchen*, February 2020.
 - [3] Kristin Antelman. Do open access articles have a greater research impact? *College and Research Libraries*, 65:372–283, 2004.
 - [4] Peter Arzberger et al. Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3, 2004.
 - [5] Chris Atkeson et al. Editorial board of the Kluwer journal, machine learning: Resignation letter. *SIGIR Forum*, 35(2), 2001.
 - [6] Bo-Christer Björk. Gold, green, and black open access. *Learned Publishing*, 30(2):173–175, 2017.
 - [7] Bo-Christer Björk, Mikael Laakso, Patrik Welling, and Patrik Paetau. Anatomy of green open access. *Journal of the Association for Information Science & Technology (JASIST)*, 65(2), 2014.
 - [8] Rick Bonney et al. Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11), 2009.
 - [9] Stephen Buranyi. Is the staggeringly profitable business of scientific publishing bad for science? *The Guardian*, June 2017.
 - [10] Philip Davis et al. Open access publishing, article downloads, and citations: randomised controlled trial. *BMJ*, 337, 2008.
-

-
- [11] Jamer Donovan and Carol Watson. Citation advantage of open access legal scholarship. *Law Library Journal*, 103(4), 2011.
 - [12] Hélène Draux, Mithu Lucraft, and John Walker. Assessing the open access effect for hybrid journals. *Springer Nature White Paper*, 2018.
 - [13] Yassine Gargouri, Vincent Larivière, Yves Gingras, Les Carr, and Stevan Harnad. Green and gold open access percentages and growth, by discipline. *arXiv*, 1206.3664, 2012.
 - [14] Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. The ACL Anthology: Current state and future directions. In *Proceedings of the 1st ACL Workshop for NLP Open Source Software (NLP-OSS)*, 2018.
 - [15] Michael B. Jones. Open public access to all USENIX conference proceedings. *;login:*, 33(2), April 2008.
 - [16] Steve Lawrence. Online or invisible? *Nature*, 411(6837), 2001.
 - [17] David Lewis. On less restrictive access to archival research literature. *SIGIR Forum*, 35(2), 2001.
 - [18] John McKay and Cara Duckworth. Coalition of 135+ scientific research and publishing organizations sends letter to administration. *Association of American Publishers Newsroom*, December 2019.
 - [19] Stephanie Normand. Is diamond open access the future of open access? *Innovations in Scholarly Communications*, 3(2), 2018.
 - [20] Guillermo Paraje, Ritu Sadana, and Ghassan Karam. Increasing international gaps in health-related publications. *Science*, 308(5724), 2005.
 - [21] Heather Piwowar et al. The state of OA: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, 6:e4375, 2018.
 - [22] Jason Schmitt. Paywalls block scientific progress: Research should be open to everyone. *The Guardian*, March 2019.
 - [23] Valerie Spezi et al. Open-access mega-journals: The future of scholarly communication or academic dumping ground? a review. *Journal of Documentation*, 73(2), 2017.
 - [24] Jonathan Tennant et al. The academic, economic and societal impacts of open access: an evidence-based review. *F1000Res*, 5(632), 2016.
 - [25] Moshe Y. Vardi. Open, closed, or clopen access? *Communications of the ACM*, 52(7):5, 2009.
 - [26] Moshe Y. Vardi. Open access and ACM. *Communications of the ACM*, 61(4):7, 2018.
 - [27] Ajit P. Varki. The times they are still a’changing: keeping up with the times. *Journal of Clinical Investigation (JCI)*, 97(1), 1996.
 - [28] Xianwen Wang, Chen Liu, Wenli Mao, and Zhichao Fang. The open access advantage considering citation, article usage and social media attention. *Scientometrics*, 103(2), 2015.
 - [29] Julie R. Williamson. ACM, SIGCHI, and the economics of open access publishing. *SIGCHI Blog*, January 2020.
 - [30] Alicia Wise and Lorraine Estelle. How society publishers can accelerate their transition to open access and align with Plan S. *Learned Publishing*, 33(1):14–27, 2020.
-