A Comparative Analysis of Article Recommendation Platforms

Rand Alchokr*, Jacob Krüger*†, Gunter Saake*, Thomas Leich‡
*Otto-von-Guericke University Magdeburg, †Ruhr-University Bochum, ‡Harz University Wernigerode; Germany
Email: {rand.alchokr | saake}@ovgu.de; jacob.krueger@rub.de; tleich@hs-harz.de

***Abstract*—Even though it is a controversial matter, research (e.g., publications, projects, researchers) is regularly evaluated based on some form of scientific impact. Particularly citation counts and metrics building on them (e.g., impact factor, h-index) are established for this purpose, despite missing evidence that they are reasonable and researchers rightfully criticizing their use. Several ideas aim to tackle such problems by proposing to abandon metrics-based evaluations or suggesting new methods that cover other properties, for instance, through Altmetrics or Article Recommendation Platforms (ARPs). ARPs are particularly interesting, since they encourage their community to decide which publications are important, for instance, based on recommendations, post-publication reviews, comments, or discussions. In this paper, we report a comparative analysis of 11 ARPs, which utilize human expertise to assess the quality, correctness, and potential importance of a publication. We compare the different properties, pros, and cons of the ARPs, and discuss the adoption potential for computer science. We find that some of the platforms' features are challenging to understand, but they enforce the trend of involving humans instead of metrics for evaluating research.**

***Index Terms*—quality assessment, peer review, post publication, computer science, recommendation service platforms**

I. Introduction

Assessing the quality of publications (e.g., during peer review, for literature reviews, for grant proposals) is a challenging task for researchers, requiring more and more effort due to the rapidly increasing number of publications. Still, researchers, institutions, and publication venues are judged based on their publications. An interesting direction for providing reliable quality assessments, which emerged from online journals, are *post-publication assessments*, such as publishing the actual reviews (e.g., open reviews), comments, recommendations, discussions, or endorsements [44]. Contrary to traditional metrics and pre-publication peer reviews, such post-publication assessments can provide more detailed insights and can value research that may be interesting, but is out of the reader's own research direction. We refer to platforms that provide any form of post-publication assessment as **Article Recommendation Platforms (ARPs)**, for example, Peeriodicals or FacultyOpinions. These platforms rely on their communities to evaluate the quality and importance of publications, aiming to overcome the limitations of metrics and pre-publication reviews.

Traditionally, metrics (e.g., citation counts, impact factors) and peer reviews are used to evaluate the quality and importance of publications [6], [30], [32]. However, researchers criticize the use of citations, and metrics building upon them, as quality indicators [27], [38]—but these may still be the best option, at least for standard literature [5], [27]. To complement traditional metrics, modern communication channels (e.g., social media, blogs) have been explored to measure the scientific impact of a publication outside of academia, leading to the introduction of Altmetrics [36]. While such alternative metrics may provide a better indication for a publication's public popularity and potential impact [12], [25], their actual value is debated and they are used sparsely. Also, all of such metrics serve as measures for the impact after publishing research.

In contrast, peer review is the standard practice to assure the quality of a piece of work before it is published [45]. Peer reviewing has been greatly discussed in research [4], resulting in various strategies of peer review that are still debated (e.g., unblinded, single blinded, or double blinded). Arguably, the pros and cons of the different strategies remain poorly understood (e.g., regarding biases, fairness, review quality), despite the high regards some researchers hold for some of these strategies [4]. ARPs are an interesting direction to combine traditional metrics with post-publication reviews, providing an additional quality assessment that may help tackle the problems of these two forms of assessments.

In this paper, we present a comparative overview of 11 ARPs. We investigate the services each ARP provides to understand commonalities and differences, based on which we distinguish seven properties that specify the degree of post-publication assessment an ARP enables (e.g., commenting only, actual reviews). For this purpose, we performed a systematic manual analysis of each ARP and its individual services. We use our results to discuss the pros and cons of ARPs, their relation to traditional metrics, and their potential for computer science. *So, our goal is to explore platforms that provide post-publication assessments and link our findings to traditional quality assessments.* More precisely, we contribute the following in this paper:

- We present an overview of 11 recent ARPs and their properties, providing a classification and detailed understanding of the pros and cons they can have compared to traditional quality assessments (i.e., metrics, peer review).
- We explore existing studies that investigate ARPs and their relation to metrics (i.e., citations, Altmetrics), discussing the empirical evidence on using ARPs as a complement.
- We discuss the problems of ARPs, how they could be

978-1-5386-5541-2/18/$31.00 ©2021 IEEE

adopted or improved (particularly for computer science), and what future research is required.
- We publish an open-access repository with lists of the publications we considered for this analysis.[1]

Our results suggest that the services of ARPs are suitable to tackle problems of existing quality assessments. So, our paper helps researchers and developers to understand important properties of ARPs, highlights their shortcomings, and ideally helps improve them in the future.

II. Background and Related Work

ARPs aim to provide an overview of the rapidly evolving publishing landscape, often combining metrics and reviews based on their post-publication assessment mechanisms. In the following, we briefly discuss bibliometrics and Altmetrics as quality indicators after publishing. Moreover, we introduce peer reviewing as a quality assurance mechanism before publishing and relate it to ARPs.

A. Bibliometrics

Traditional metrics are common measures the research community relies on when assessing the scientific impact and quality of a publication [11]. Such metrics are considered to facilitate the examination of large datasets and even the decision making on individuals, institutions, or research grants [19]. For many years, different bibliometrics (particularly citation-based) metrics have been proposed as valuable complements to peer reviewing, for example, to avoid biases [6], [51]. So, comparing metrics to peer reviews has been widely acknowledged as a way of validating the feasibility of these metrics [17], [22].

Arguably, the number of citations, the h-index, and the impact factor are among the most important metrics used for assessing the impact and quality of publications, publishing venues, authors, or research in general. Citations are derived directly from other publications referencing a piece of work, and are assumed to directly reflect on the impact and quality of a publication [1]. Actually, citations imply credibility to the reader and citation links provide a valuable source for identifying related work [41]. Citation-based metrics are intended to measure the total impact of a publication on a research field [35], with several adapted metrics aiming to improve on the simple citation count (e.g., h-index, impact factor) [40].

Despite their potential benefits, bibliometrics have always been criticized in the context of measuring the impact or quality of research, which they do not necessarily capture [19]. Some cons of different bibliometrics include:

- Some bibliometrics, such as impact factors, are influenced by technical issues that are not related to the quality or impact of a publication [35].
- Some citation-based metrics reward low and penalizes high productivity [35].
- The metrics vary between different sources, due to each source's properties and coverage [1], [18], [20], [40].
- The average citation numbers vary heavily across different research areas [1].
- Citations require time to accumulate [14].
- Some metrics can be manipulated, for instance, by using inappropriate self-citations [15], [19].
- Many citation metrics are insensitive to the position of an author on the publication [35].

Despite these cons, studies suggest that using bibliometrics is a helpful complement to mitigate potential biases during traditional peer review [28], [37], [51].

B. Altmetrics

Altmetrics have been introduced in 2010 as a means to assess the impact of a publication based on publicly available interfaces of various online platforms [15]. Basically, these metrics allow researchers to track the impact of publications beyond traditional bibliographic metrics [16]. Although Altmetrics may not accurately represent scientific quality, they can reflect on the spread of research to a broader audience by calculating quantitative values of user interactions on social platforms, for instance, Wikipedia, Twitter, or Facebook. Typically, Altmetrics cover usage statistics, such as the number of downloads, views, or read times—providing a more real-time assessment compared to citation metrics [14], [46].

Based on such benefits, Altmetrics are studied intensively and researchers argue that they can serve as quality indicators or replacements for traditional metrics (e.g., suggesting that Altmetrics correlate to later citations) [14], [29], [33]. Researchers recommend to use both kinds of metrics when assessing the impact of a publication to complement their pros and cons [29]. Some cons of Altmetrics are:

- There is missing evidence and a lack of theory on the benefits of Altmetrics in quality assessments [42].
- Altmetrics can be difficult to access and collect [42], [46].
- Most current Altmetrics are commercialized [29], [46].
- Since they are based on social platforms, Altmetrics have a low coverage, particularly for older publications [46].
- Altmetrics can easily be manipulated, since there is a lack of quality control on the data [29], [46], [52].

These cons highlight the problems of using any form of metric as sole indicator for the quality or impact of a publication, which will arguably be incomplete and potentially skewed.

C. Peer Review

Reviews by academic peers are an essential part of publishing research, representing an important quality assurance mechanism [43]. A peer review should involve a neutral researcher who is knowledgeable enough to read a publication and 1) verify its soundness and originality, 2) validate its results, 3) assure its quality, and 4) assess its appropriateness for a venue [39]. Many different strategies for peer reviewing have been proposed (e.g., double blind) with an increasing trend towards open peer review [24]. Orthogonal to such strategies, we distinguish two types of reviews in this paper.

Pre-Publication Reviews are those commissioned by a publication venue to decide whether to accept a submitted publication. Reviewers agree to review a submitted publication and provide feedback to the authors. The feedback implies a decision

[1] https://doi.org/10.5281/zenodo.5267169

towards acceptance or rejection, and is intended to help the researchers improve their work.

Post-Publication Reviews are an emerging trend that can refer to various different models. Namely, this type of review may be:

- A peer-review strategy for journals that can be an additional service or even a replacement for pre-publication reviews. In such models, publications are published online before a peer review was conducted, and are then reviewed by formally invited or volunteering reviewers.
- A platform to provide feedback on publications, for instance, through discussions, comments, recommendations, or actual reviews. Pre-prints are usually published similarly to the first type of models, but they are independent of the actual publication venue, which is why most dedicated pre-print platforms (e.g., ArXiv) fit into this second type.

We focus on the second type of post-publication reviews, which is driven by qualitative feedback of interested researchers. Such types of reviews actually provide informative insights and help other researchers understand the quality and importance of a publication, making such platforms ARPs.

III. Methodology

In this section, we describe the goal of our study as well as how we elicited and analyzed the data to achieve that goal.

A. Goal and Research Objectives

Our goal was to gather information on ARPs and their key services. To this end, we defined three research objectives:

RO_1 *Compare ARPs and their properties.*
The individual services and properties of ARPs are diverse and not well-investigated in research. We aim to provide a conceptual framework of ARPs based on a comparative analysis of 11 ARPs.

RO_2 *Study correlations between ARPs and traditional metrics.*
We studied the existing literature regarding empirical data on correlations between the assessments on ARPs and bibliometrics as well as Altmetrics. So, we aim to provide an understanding on the impact of ARPs themselves.

RO_3 *Discuss potential obstacles of establishing ARPs.*
Using our previous findings, we discuss the pros and cons of ARPs. We focus particularly on potential improvements and ways to establish ARPs in a research community.

Our results can help researchers understand, scope new, or define improvements for existing ARPs.

B. Collecting Data from the Literature

To obtain a first understanding of ARPs and elicit candidates for our analysis, we employed a systematic literature review [21]. Note that we focused on qualitatively analyzing the publications, which is why we do not report he typical statistics of such reviews. Next, we summarize the individual steps we employed.

Search Strategy. We performed an automated search on Scopus,[2] which covers various publishers and allows to download collections of the returned results. Based on our research objectives, we focused on publications related to quality assessments, post-publication reviews, and recommendation platforms. After testing multiple search strings, we decided to apply the following search string:

```
TITLE-ABS-KEY(article OR publication
OR scholarly OR "scientific paper")
AND ("recommendation system" OR f1000
OR "post peer review") AND quality
AND (LIMIT-TO(SUBJAREA,"COMP")) AND
(LIMIT-TO(LANGUAGE,"English"))
```

We did not apply any restrictions on the publishing date, and recovered all types of publications listed in Scopus.

Selection Criteria. We selected publications for our study that fulfilled the following inclusion criteria (IC):

IC_1 Written in English.

IC_2 Concerned with publication quality, quality assessments and metrics, or post-publication peer review.

IC_3 Belongs to the computer science research area.

Note that we aimed to cover IC_1 and IC_3 through our search string, while we had to check each publication for IC_2.

Conduct and Results. For consistency, the first author of this paper conducted the search and selection process. Initially, we obtained a list of 818 publications, involving many irrelevant ones that were concerned with recommender systems in various domains. After scanning the titles and abstracts, we kept 97 and 23 publications, respectively. To tackle the problems of automated searches [2], [21], [23], we performed additional backwards snowballing on the 23 publications we considered relevant—leading to six new publications. So, our search revealed 29 relevant publications that discuss ARPs and their relation to quality assurance. However, none of the publications performed a comparative analysis as we do.

Web Search. To refine our dataset, we performed a web search for each ARP we identified in the selected publications. Moreover, we used Wikipedia to search for additional ARPs (e.g., through related articles sections). We used the results of this web search (i.e., the official websites, Wikipedia) to improve our understanding of the services and properties of each ARP.

Data Extraction. To address our research objectives, we elicited all ARPs named in the selected publications (RO_1).

TABLE I
Overview of the ARPs we selected.

ID	ARP	Year	# Papers	Url
P1	ArXiv	1991	1,862,711	https://arxiv.org/
P2	FacultyOpinions	2000	81,577	https://facultyopinions.com/
P3	ResearchGate	2008	NA	https://www.researchgate.net/
P4	Publons	2012	7,373	https://publons.com/
P5	PubPeer	2012	NA	https://pubpeer.com/
P6	ScienceOpen	2013	180	https://ScienceOpen.com/
P7	PeerCommunityIn	2016	100–1,000	https://peercommunityin.org/
P8	PreLights	2018	2,095	https://prelights.biologists.com/
P9	Peeriodicals	2018	NA	https://peeriodicals.com/
P10	SciPost	2018	100–1,000	https://scipost.org/
P11	Plaudit	2019	NA	https://plaudit.pub/

NA = Not Available

[2]https://scopus.com/

Moreover, we extracted standard bibliographic data and all correlations between ARPs and other metrics (RO_2). Finally, we extracted all details on the services, properties, pros, and cons of an ARP reported in the publications or in the web (RO_3).

IV. Selected ARPs

While reading through the 29 publications and the results of our web search, we found it challenging to clearly identify ARPs among the different platforms that were mentioned. In the end, we selected 11 ARPs that we could clearly identify to provide post-publication assessments, that seem to be still active, and for which we found detailed descriptions to tackle our research objectives. We display all 11 ARPs in Table I, ordered by the year in which they have been established. All 11 ARPs comprise some form of post-publication assessment, such as recommendations, discussions, comments, or actual reviews. Some ARPs (e.g., ArXiv, ResearchGate) provide quite unique services in addition to commenting on papers (e.g., ResearchGate computes own metrics). Nonetheless, all ARPs share the concept of using expert knowledge for assessing the importance and quality of publications, and thus provide an actual alternative to existing metrics or pre-publication reviews. Next, we briefly introduce each ARP.

ArXiv is a well-known, free distribution service and open-access repository for publications and particularly pre-prints that were not yet peer reviewed. While ArXiv focuses mainly on providing persistent storage, users can still comment and review the available publications. ArXiv is among the oldest and most established platforms, which is highlighted by the large number of publications and comments from various domains.

FacultyOpinions was initially part of F1000,[3] then it was formally F1000Prime, and became FacultyOpinions in early 2020. FacultyOpinions focuses on post-publication assessments of papers that have already been published in journals [31]. Mostly, this ARP is a database of important papers from the biomedical research area with a network of 8,000 scientific experts, who must be Associate Professor (or equivalent) at least. These experts pick publications they consider important and write short summaries of the key insights.

ResearchGate is a popular social network for researchers with over 19 million members. According to a study of van Noorden [47], it is the largest academic social network in terms of active users. ResearchGate provides several services, such as sharing research, collaborating with peers, and commenting. Researchers can solve problems by asking questions and answering or commenting those questions. We consider ResearchGate as an ARP, because it also allows to ask for post-publication reviews and to comment on publications.

Publons is a multi-service platform powered by the Web of Science, and integrates with multiple journals. It was established mainly to strengthen the relationships with peer reviewers and provide recognition for their work. Still, Publons offers pre- and post-publication peer reviewing.

[3]https://f1000research.com/

PubPeer is a non-profit corporation that was established with the goal of improving the quality of research. For this purpose, PubPeer provides innovative services for community interaction and for benefiting readers as well as reviewers. Currently, PubPeer focuses on improving its online post-publication peer-review service.

ScienceOpen is another multi-service discovery platform for scientific publications. This ARP encourages researchers to interact and enhance their research openly, offering post-publication peer review and recommendations to assess publications. It follows strict rules to maintain high quality standards (e.g., some functions require to register at ScienceOpen via ORCID and a certain number of publications).

PeerCommunityIn (PCI) is a non-profit organization that aims to create a new scientific publishing system where researchers and reviewers can freely review and recommend scientific publications. PCI has about 1,200 recommenders who review and award recommendations to the pre-prints published on the platform. While the authors are not required to submit their pre-prints to actual journals, PCI argues that the recommendations publicly guarantee research quality.

PreLights is another non-profit publishing platform that provides highlighting and commenting services for pre-prints to the biological research area. Authors are not responsible for submitting their own pre-prints for evaluation. Instead, early-career researchers select interesting publications and write highlight summaries about these.

Peeriodicals is integrated with PubPeer for the purpose of selecting the best publications. It is a virtual journal that allows its users to create a periodical and add the most interesting and useful publications for a specific audience of readers to it. Reviewers can communicate and suggest improvements for any publication.

SciPost is a fully online, non-profit publishing platform managed by researchers who aim to change the current publishing infrastructure. SciPost provides post-publication assessments as quality indicators. For this purpose, it has commentary pages that are associated to publications.

Plaudit is a non-profit publisher-independent plug-in that offers open endorsements. Plaudit aims to provide a simple and accessible alternative indicator for the quality of publications. To this end, this ARP allows trusted researchers to publicly endorse publications, and thus provide credibility for valuable research.

V. Properties of ARPs (RO_1)

For our comparative analysis, we manually inspected each ARP, categorized their services, derived a set of core properties, and defined sets of possible values for each property. To this end, we relied on each ARP's website, available documentation, the publications we selected, and our experiences as researchers in the subject area. In this section, we discuss all properties in a comprehensible order (i.e., a property or its values require an understanding of a previously described property). We display an overview of all properties, their potential values, and the concrete value for each ARP in Table II.

TABLE II
OVERVIEW OF THE PROPERTIES OF THE 11 ARPS WE COMPARED.

		Article Recommendation Platform										
		ArXiv	**FacultyOpinions**	**ResearchGate**	**Publons**	**PubPeer**	**ScienceOpen**	**PeerCommunityIn**	**PreLights**	**Peeriodicals**	**SciPost**	**Plaudit**
Supported documents	Pre-prints	●		●	●	●	●	●	●	●	●	●
	Published		●	●	●	●	●			●	●	●
Coverage	Multiple domains	●		●	●	●	●				●	●
	Single domain		●					●	●	●		
Transparency	Anonymous				●	●		●		●	●	
	Open identities	●	●	●	●	●	●	●	●	●	●	●
	Open review reports					●	●	●	●	●	●	
	Restricted reports		●		●							
Interactivity	Restricted interactions		●	●	●		●	●		●	●	
	Unestricted interactions				●	●	●		●		●	
Accessibility	Free access	●		●	●	●	●	●	●	●	●	●
	Restricted access		●		●							
Functionalities	Commenting	●	●	●	●	●	●	●	●	●	●	
	Endorsements											●
	Post-publication reviews		●		●	●	●			●	●	●
	Pre-publication reviews				●		●	●	●	●	●	
	Recommendations		●		●		●	●	●	●		
	Recommendation classifications		●		●		●			●	●	
Publishing services	External indexing	●		●	●		●	●			●	
	ID for recommendation / review		●		●		●	●		●	●	
	Publishing						●				●	

A. Supported Documents

Definition. This property addresses the different versions of an author's publication that can be managed by an ARP. In the research community, we distinguish several terms for certain versions of a publication, most fundamentally: *Pre-prints* refer to publications before they have been peer reviewed. This version of a research publication can often be shared and posted on ARPs or an author's personal website to make a publication freely available, before and after the actual peer-review. *Published* publications have been accepted at a certain venue and are officially available through a publisher. Usually, these versions of a publication cannot be freely shared.

Pre-prints. It is not surprising that most ARPs support pre-prints, with the sole exception of FacultyOpinions. For example, all papers on ArXiv are pre-prints, even though not all ArXiv pre-prints have been submitted to a peer-reviewed venue. All ARPs that provide pre-publication reviewing (explained shortly), enable researchers to post their unpublished pre-prints in an open-access repository. Interestingly, PreLights allows to add only biorxiv[4] pre-prints.

Published. ARPs that provide post-publication reviewing or services, such as FacultyOpinions, consider officially published publications—but they often allow to post pre-prints to avoid copyright issues (i.e., hybrid ARPs). Some ARPs enforce certain requirements, for instance, PubPeer, Plaudit, or ScienceOpen require some identifier, such as a DOI. In contrast, SciPost allows to add any version of a publication to the commentaries service, allowing other researchers to comment it.

[4]https://biorxiv.org/

B. Coverage

Definition. This property refers to the domains each ARP covers. The ARPs we studied have mainly two types of coverage: Most cover multiple research domains, but with a limited number of publications. Others cover only a single domain, and potentially only some of its subject areas.

Multiple domains. ResearchGate and PubPeer support all research domains, without providing a classification. ScienceOpen includes all disciplines and has more than 60 million publication records available for post-publication assessments. Other ARPs support multiple defined research areas, for instance, ArXiv defines eight different domains and 155 subject areas. Similarly, Publons supports multiple disciplines, but computer science remains at the end of the list in terms of post-review publications, with only 312 publications compared to, for example, 1,775 in medicine and 2,016 in psychology. SciPost covers only the domains multidisciplinary, formal sciences, natural sciences, and social sciences.

Single domain. A minority of the ARPs cover only a single research domain. For instance, FacultyOpinions focuses on the biomedical domain with 47 subject areas. Similarly, PeerCommunityIn covers the biomedical domain, while PreLights and Peeriodicals focus on the biological domain. Interestingly, all single-domain ARPs stem from the biomedical or biological domains.

C. Transparency

Definition. With this property, we capture to what extent an ARP reveals or hides information. To this end, we consider two aspects: First, whether it is possible to hide a reviewer's or commentator's identity. Second, whether peer reviews are openly available or restricted.

Anonymity. Anonymous interaction allows users to comment or review while hiding their identity. This feature is optionally supported in Publons, PubPeer, Peeriodicals, SciPost, and PeerCommunityIn. Particularly, PubPeer hides the reviewers' real identity, asking them to obtain a safeguarded account and authenticating their identity before approving a review.

Open identities. Essentially all ARPs allow their users to reveal their identities. However, ScienceOpen and FacultyOpinions actually forbid anonymity, which is why the identity of the reviewers and their comments are visible at all times. Other ARPs, such as PeerCommunityIn, allow reviewers to freely choose to reveal their identity. Most strictly, Plaudit requires an ORCID account for the endorser.

Open review reports. For six ARPs, reviews or recommendations are publicly available. Such transparency has the pro of allowing others to judge a publication based on the review. However, the review also reflects on the reviewer, which is why such transparency may not be desired by everyone.

Restricted reports. Only two ARPs restrict the visibility of reviews or endorsements. Namely, FacultyOpinions makes reviewing and recommendation reports only available after paying the registration fee. In contrast, for Publons, the availability of reports depends on the connected journals' editorial policies. Finally, three ARPs have no reports (only comments).

D. Interactivity

Definition. With this property, we capture to what extent researchers can interact on an ARP. Essentially, we distinguish between restricted and unrestricted interactions. The former means that authors are not able to directly interact with their reviewers, while the latter means that even multiple rounds of open discussions are possible.

Restricted interactions. FacultyOpinions, PeerCommunityIn, Peeriodicals, ArXiv, ResearchGate, and Plaudit restrict interactions. For instance, at Peeriodicals, the reviewers are allowed to openly interact with each other, but this excludes authors. Almost all ARPs define certain restrictions and constrains regarding interaction between researchers and with the ARP. For example, ScienceOpen aims to maintain a high quality standard by defining that their expert reviewers must have at least five official publications in their ORCID.

Unrestricted interactions. Five of the 11 ARPs allow authors to freely interact. This means that the authors can respond to reviewer questions and amend the work upon their notes. Similar to a rebuttal or typical journal revisions, this procedure allows to explain unclear points through multiple rounds of discussion and to resubmit a modified version.

E. Accessibility

Definition. This property covers the different options for accessing an ARP. Precisely, we captured whether users have free access to the ARP, or must pay to participate. Note that this covers mostly access to the post-publication assessments, but also the publications themselves in one case.

Free access. All ARPs provide a free access option, except for FacultyOpinions. Still, to fulfill their purpose, all ARPs require that the user creates an account.

Restricted access. FacultyOpinions offers only a 30 days free trial for interested users. Afterwards, a subscription is needed to access the recommendations posted by the experts. Interestingly, Publons allows authors to choose whether their publications are freely available or not.

F. Functionalities

Definition. This characteristic describes the quality-related services provided by an ARP. Pre-publication reviews are mainly those commissioned by a venue during the path to publication. Post-publication reviews are those about published publications, and are mostly not considered for a journal. Recommendation, commenting, and endorsement functionalities are essential to discover important publications, as proposed by experts.

Commenting. ResearchGate and ArXiv provide no peer-reviewing services (ResearchGate allows to ask for feedback, but this is not visible to others), since they focus on making publications accessible and visible. Still, commenting is directly available in ResearchGate. In contrast, commenting in ArXiv is restricted to Scirate[5] and Arxivsanity,[6] which are collaborative tools allowing users to vote and comment on publications. Other ARPs provide organized recommendation services, and support commenting at the same time. *SciPost* offers their Commentaries service, which aims at adding papers to collections to make them open for comments. As the only exception, Peeriodicals does not support commenting.

Endorsements. Plaudit is the only ARP that provides recommendations in the form of endorsements. We distinguish between these two values, because recommendations in all other ARPs are textual descriptions or summaries of a publication. In contrast, an endorsement in Plaudit essentially means liking a publication as on social-media platforms.

Post-publication reviews. A total of seven ARP provide services for actual post-publication reviews. Such post-publication reviews are a helpful means for readers to get an overview of a publication's content, quality, and importance. Only three of

[5]https://www.scirate.com/

[6]http://www.arxiv-sanity.com/

those ARPs focus on post-publication only, while the remaining four also allow pre-publication reviews.

Pre-publication reviews. Six different ARPs explicitly offer pre-publication reviews. Such reviews are intended to improve a publication before it is actually submitted or published at a peer-reviewed venue. Only two of the six ARPs do not provide post-publication reviews.

Recommendations. FacultyOpinions, Publons, ScienceOpen, PeerCommunityIn, and Peeriodicals support textual recommendations. Highlights in PreLights summarize selected pre-prints, explain their pros, cons, and relevant comments or answers from the pre-prints' authors themselves. Thus, they are highly similar to recommendations in other ARPs.

Recommendation classifications. Classifications help to increase the visibility of publications by guiding interested researchers. Publications in FacultyOpinions can be either interesting, novel, challenging, negative/Null result, confirmation, or other reasons. Then, a semantic rating (exceptional, very good, good, and dissent) is provided. ScienceOpen uses five-star ratings with four factors (importance, validity, comprehensibility, and completeness) followed by a formally written review. For Publons, publication scores are assigned based on two properties (quality and significance), which allow to gain a fast understanding of how a publication is rated. Also, users can filter based on these properties to see only publications of a certain quality. At PeerCommunityIn, recommendations are short texts (about half a page), whereas at Scipost, the reports can have several ranks for different properties (validity, significance, originality, clarity, formatting, and grammar). ScienceOpen adds a special property called "collection," which allows researchers to select important publications on a specific topic and incorporate these into a single document. Peeriodicals employs the same concept.

G. Publishing services

Definition. This property describes whether and how publishing activities are carried out or supported by an ARP. Interestingly, some of the ARPs are not only recommendation or storing services, but actually publish publications. Note that three ARPs have no support for any publishing activities (e.g., not even linking to the publication).

External indexing. PreLights and Publons incorporate publishing indirectly through journal partners. PeerCommunityIn succeeded to get the support of 45 journals that consider submissions of publications recommended by it. SciPost is listed in a part of Web of Science's Core Collection, in Google Scholar, and in INSPIRE. ResearchGate and ArXiv are repositories that allow to submit the posted pre-prints to any publication venue, and ArXiv is listed at least in dblp and Google Scholar.

ID for recommendation / review. Reviews are considered valuable sources of information for authors and researchers in general. FacultyOpinions, Publons, ScienceOpen, SciPost, and PeerCommunityIn publish their reviews or recommendation reports, and can provide a separated ID for them—which makes the reports fully citeable, transparent, and more visible. In Peeriodicals, the editors decide whether the reviews for a publication are published. On the contrary, other ARPs do not support publishing reviews or recommendations, mainly because they do not offer such services (i.e., ResearchGate, ArXiv, Plaudit) or because they behave similarly to open commenting platforms (i.e., PubPeer, PreLights).

Publishing. SciPost and ScienceOpen actually enforce an actual peer-review process. Publications that pass this process, and only those, are then stored on the corresponding ARPs. The other ARPs do not provide direct publishing options, since they mostly act as reviewers, recommender systems, or repositories for pre-prints.

VI. Discussion

In the following, we discuss our results in the context of each of our research objectives.

A. RO_1: ARPs Comparison and Properties

Importance of post-publication review. Discussions about already published publications can be highly valuable, if performed by experienced peers. Such reviews can identify flaws that were missed in the actual peer review, highlight the quality and importance of the research, and suggest opportunities for future work. Interestingly, while some ARPs focus on changing quality and impact assessments based on such reviews, others aim to reform various steps of current publishing practices. Still, the post-publication reviews share the common goal of improving even on published results. In this regard, some studies describe FacultyOpinions as an aid for researchers to get pointers to relevant publications, as an important tool for assessing research, and a valuable complement to existing metrics [10], [52]. Reflecting on these insights, we consider particularly platforms that focus on recommendations and rating services (e.g., FacultyOpinions, Prelights, Peeriodicals) as helpful means to navigate through the ever growing number of publications.

Centralized platform. Research publications are sometimes debated on different social-media platforms or developer forums. Consequently, most discussions are scattered across various platforms and may never reach the authors. Moreover, valuable information to assess the quality and impact of a publication may get lost. ARPs could provide a centralized platform, maintained by experienced researchers, to accumulate and share discussions on publications—which would arguably lead to great benefits for research.

ARPs vs social media platforms. Social-media platforms are open for everyone and often limited in their expressiveness (e.g., character limits). This can lead to non-constructive criticisms, misinterpretations, or simply a waste of time aiming to inspect and understand all arguments. However, there are also pros, for instance, a higher visibility of the research or insights from communities outside of the own domain. Still, establishing ARPs would reinforce post-publication assessments and mitigate the problems of social-media platforms. Potentially, feedback on different platforms could be automatically collected and synthesized to support researchers who contribute to an ARP.

Users anonymity. Enabling anonymity may encourage junior researchers to participate more actively in reviewing and discussing research, since they would not have to fear negative consequences for criticizing more senior researchers. However, anonymity may not only empower researchers, it also introduces biases regarding the fairness of the reviews. ARPs that do not employ anonymity argue that this increases the quality of reviews, and yields more comprehensive, constructive, as well as well-written reviews. This point is essentially the same as for traditional peer-review strategies. Still, high quality reviews could be directly recognized in an ARP.

Coverage barrier. A study by Xuemei and Thelwall [26] pointed out that recommended publications receive 1.30 recommendations on average, and more than 90 % of these are given within half a year after a publication has appeared. Moreover, most ARPs focus on the biomedical domain and comprise only a small number of publications. Such problems may be caused by reviewers selecting which publications to review and recommend, which is why most are not post-peer reviewed [22], [52]. Namely, only 2 % of the publications in FacultyOpinions received at least one recommendation [48]. Consequently, there is a coverage barrier with respect to the considered domains as well as the time period in which publications are visible.

B. RO_2: Correlations Between ARPs and Metrics

ARPs that include experts' post-publication assessments are a new opportunity to complement bibliometrics and Altmetrics for assessing the quality of publications [34]. In Table III, we summarize correlational studies between ARPs and other metrics we identified during our systematic literature review (cf. Section III-B). Most of the publications analyzed F1000Prime or F1000, which is now called FacultyOpinions—and which provides the F1000 Article Factor score (FFa score) that is based on recommendations. We can see that the studies span a longer period and have been performed on various datasets from different research areas. Most results suggest that recommendations in ARPs (e.g., FFa score) correlate with citation counts. This may indicate that positive post-publication assessments lead to more citations, which supports the assumption that ARPs can indicate important and high-quality publications before traditional metrics can. However, previous studies could not demonstrate a causal relationship for this correlation. Consequently, we do not know if positive comments in ARPs lead to more citations or vice versa—or whether both are simply caused by high-quality publications.

C. RO_3: Challenges of Establishing ARPs

Challenges. ARPs are relatively new and have rarely been adopted across different domains, with our results showing a prevalence of biomedical research. Surprisingly, the computer-science community, as a primary candidate for developing new software platforms, seems to barely notice ARPs, yet. There are several challenges connected to establishing and maintaining an ARP, as highlighted by the prior medical ARP PubMed Commons that was discontinued in 2018. From our analysis, the related work, and such negative examples, we argue that ARPs face the following problems:

- Contributing to ARPs (e.g., reviewing, maintaining, publishing) represents additional workload for researchers on top of their usual obligations.
- Discussing and sharing ideas or research opportunities during post-publication assessments may be problematic, due to the competitive nature of research.
- The number of different ARPs and their various options may confuse and demotivate researchers to contribute.
- The post-publication assessments must be comprehensible for other users who have not read the publication; so, ensuring the quality of the report is key.

While this list is incomplete, we argue that these are major problems we need to solve to actually establish ARPs.

Steps to encourage the adoption of ARPs. ARPs must be promoted by researchers and the scientific community by highlighting their potential benefits. Particularly the computer-science community could help develop ARPs that can help analyze larger numbers of publications and contribute to a common knowledge base. We propose the following steps to advance the use of ARPs:

- Extensive collaborative research that investigates the pros and cons of ARPs for providing high-quality post-publication assessments of publications.
- Improving the perception of post-publication assessments as a beneficial means for the overall research community.
- Further unifying and studying the properties (cf. Section V) to understand their pros and cons in more detail.
- Consolidating ARPs by advancing towards a centralized platform that semi-automatically collects information on publications (e.g., bibliographic data).
- Properly recognizing the effort researchers spend on performing post-publication assessments for the community (e.g., citeable reports, awards).
- Defining criteria for selecting experts and for assessing the quality of the post-publication assessments.

Obviously, ARPs will not immediately succeed in the near future. However, with the increasing use of some ARPs (e.g., ArXiv) across all research communities and future advances towards open research, we argue that the insights and open problems we derived in this paper are key directions for future work.

VII. Threats to validity

A threat to the external validity of our study is that we covered only a set of all existing ARPs. Consequently, other platforms may have interesting properties we could not consider. One reason for this issue is the limited number of publications describing and surveying ARPs. We argue that our methodology led to a representative sample, and we included the most prominent ARPs. While we cannot fully avoid this threat, we think that such a comparative analysis is needed as a starting point for future research.

Another threat to the internal validity is that we may have wrongly classified specific properties. We aimed to

TABLE III
OVERVIEW OF CORRELATION STUDIES BETWEEN ARPS AND QUALITY METRICS.

Ref	Year	Purpose	Dataset	results
[3]	2009	Compare FFa with bibliometrics to assess quality.	687 Welcome-Trust publications.	A strong, positive association between FFa and citations with the exception that highly rated publications were not highly cited during the first three years after publication.
[50]	2010	Compare F1000Prime ratings with citations.	1530 publications from seven major ecological journals in 2005.	Approximately one-third of the recommended publications were cited less frequently. F1000Prime publications performed poorly in predicting citation.
[26]	2012	Investigate correlations between F1000, Mendeley, and traditional bibliometrics.	1397 genomics and genetics publications selected by F1000 members from 172 journals.	FFa, Mendeley counts, and citations correlated significantly at the 1 % level with an FFa score of at least seven. These sources are useful for post-publication assessments.
[31]	2013	Examine the relationships between publication types (assigned labels), citation counts, and FFa.	Random sample of F1000 medical publications from 2007 and 2008.	Highlighting key properties could help to reveal the hidden value of some medical publications; citation counts and FFa scores were significantly different for two classifications: new finding and changes clinical practice.
[8]	2013	Investigate the correlation between peers' ratings and bibliometrics.	125 publications published in 2008 in cell biology or immunology subjects.	Correlation between the ratings and <times cited, 2nd generation citations, category actual expected citations> reached at least a medium effect size.
[48]	2014	Compare between F1000 recommendations and citations.	Complete database of F1000 recommendations.	2 % of the biomedical publications received at least one F1000 recommendation. There was a relatively weak correlation between F1000 recommendations and citations.
[13]	2015	Investigate the effect of research level and publication type on the consistency of assessments based on citations and FFa.	28,254 cited publications in F1000.	Research level had little impact, but publication type affected correlations (evidence-based research is more often cited but not highly recommended, while transformative is the opposite).
[9]	2015	Examine the influences of F1000Prime scores on citation counts over 10 years.	9,898 F1000Prime recommended publications from 2000 to 2004 and cited until 2013.	F1000Prime rating scores as quality proxies of the publications played less of a role in the later citation counts than journal impact factor.
[34]	2018	Analyze Publons metrics and their relationships with bibliometric and Altmetric indicators.	45,819 publications extracted from Publons.	Correlations between bibliometric, Altmetric counts, and Publons metrics were very weak ($r < 0.2$) and not significant.
[7]	2018	Investigate the relationship between altmetrics and F1000Prime scores.	178,855 recommendations for 140,240 publications.	Citation-based metrics and readership counts were significantly more related to quality (FFa) than tweets.
[49]	2019	Investigate the correlation between F1000Prime rating scores and WoS citations.	F1000Prime recommended publications from four medical journals in 2010.	Recommended publications in F1000 were cited significantly more in three journals, the correlations were significant.

FFa = F1000 article factor; F1000Prime and F1000 are identical to FacultyOpinions

mitigate such threats by starting with individual properties and synthesizing them into common ones. Still, our classification may be incomplete or not ideal, but we are not aware of another survey on ARPs to which we could compare our classification. For this reason, we argue that we were able to derive meaningful results that are reliable based on the available data

VIII. CONCLUSION

In this paper, we presented a comparative analysis of 11 ARPs and their properties. We derived seven properties and defined their potential values based on our analysis. Moreover, we derived open problems and future challenges for establishing ARPs in a research community. Overall, our core findings are:

- We compared 11 ARPs from which we elicited seven core properties that we explained in detail: supported documents, coverage, transparency, interactivity, accessibility, functionalities, and publishing services.
- We discussed and compared the trade-offs between the possible values of different properties to reason on their pros and cons.
- We reviewed related work to identify correlational studies, which suggest that ARPs are a helpful means to assess the quality and importance of publications, particularly before they gain citations.
- We discussed four major challenges and six directions for future work that we have to tackle particularly as computer scientists to improve our understanding of ARPs and to achieve their full potential.

So, this paper shares detailed insights with other researchers, hoping to inspire future research on an important, but not well explored research direction. More research is needed to overcome the previously discussed challenges. Therefore, for future research, we are especially interested in detailed investigations of the validity of ARPs and their correlation with other quality indicators. Moreover, we plan to conduct a survey in which the participants are scientific researchers from different domains and levels to understand their opinions of ARPs, potential barriers of adopting them, and perceived benefits for research.

ACKNOWLEDGMENT

This research has been supported by the German Research Foundation (LE 3382/2-3, SA 465/49-3).

REFERENCES

[1] D. Aksnes, L. Langfeldt, and P. Wouters, "Citations, citation indicators, and research quality: An overview of basic concepts and theories," *SAGE Open*, vol. 9, no. 1, 2019.

[2] M. Ali Babar and H. Zhang, "Systematic literature reviews in software engineering: Preliminary results from interviews with researchers," in *International Symposium on Empirical Software Engineering and Measurement*, ser. ESEM. IEEE, 2009.

[3] L. Allen, C. Jones, K. Dolby, D. Lynn, and M. Walport, "Looking for landmarks: The role of expert review and bibliometric analysis in evaluating scientific publication outputs," *PloS One*, vol. 4, no. 6, 2009.

[4] V. Batagelj, A. Ferligoj, and F. Squazzoni, "The emergence of a field: A network analysis of research on peer review," *Scientometrics*, vol. 113, 2017.

[5] J. Beel and B. Gipp, "Google Scholar's ranking algorithm: The impact of citation counts (an empirical study)," in *International Conference on Research Challenges in Information Systems*, ser. RCIS. IEEE, 2009.

[6] L. Bornmann, "Scientific peer review," *Annual Review of Information Science and Technology*, vol. 45, no. 1, 2011.

[7] L. Bornmann and R. Haunschild, "Do Altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data," *PloS One*, vol. 13, no. 5, 2018.

[8] L. Bornmann and L. Leydesdorff, "The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000," *Journal of Informetrics*, vol. 7, no. 2, 2013.

[9] ——, "Does quality and content matter for citedness? A comparison with para-textual factors and over time," *Journal of Informetrics*, vol. 9, no. 3, 2015.

[10] L. Bornmann and W. Marx, "Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts?" *Journal of Informetrics*, vol. 9, no. 2, 2014.

[11] H. Carlsson, "Allocation of research funds using bibliometric indicators – asset and challenge to Swedish higher education sector," *Infotrend*, vol. 64, no. 4, 2009.

[12] D. Crotty, "Altmetrics: Finding meaningful needles in the data haystack," *Serials Review*, vol. 40, no. 3, 2014.

[13] J. Du, X. Tang, and Y. Wu, "The effects of research level and article type on the differences between citation metrics and F1000 recommendations," *Journal of the Association for Information Science and Technology*, vol. 67, no. 12, 2015.

[14] G. Eysenbach, "Can Tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact," *Journal of Medical Internet Research*, vol. 13, no. 4, 2011.

[15] F. Galligan and S. Dyas-Correia, "Altmetrics: Rethinking the way we measure," *Serials Review*, vol. 39, 03 2013.

[16] L. M. Galloway, J. L. Pease, and A. E. Rauh, "Introduction to Altmetrics for science, technology, engineering, and mathematics (STEM) librarians," *Science & Technology Libraries*, vol. 32, no. 4, 2013.

[17] E. Garfield and R. K. Merton, *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Wiley, 1979.

[18] A.-W. Harzing and S. Alakangas, "Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison," *Scientometrics*, vol. 106, no. 2, 2016.

[19] G. Holden, G. Rosenberg, and K. Barker, "Tracing thought through time and space: A selective review of bibliometrics in social work," *Social Work in Health Care*, vol. 41, no. 3–4, 2005.

[20] Y. Kiduk and M. Lokman I, "Citation analysis: A comparison of Google Scholar, Scopus, and Web of Science," *Proceedings of the American Society for Information Science and Technology*, vol. 43, no. 1, 2007.

[21] B. A. Kitchenham, D. Budgen, and P. Brereton, *Evidence-Based Software Engineering and Systematic Reviews*. CRC Press, 2015.

[22] G. Kreiman and J. H. R. Maunsell, "Nine criteria for a measure of scientific output," *Frontiers in Computational Neuroscience*, vol. 5, 2011.

[23] J. Krüger, C. Lausberger, I. von Nostitz-Wallwitz, G. Saake, and T. Leich, "Search. Review. Repeat? An empirical study of threats to replicating SLR searches," *Empirical Software Engineering*, vol. 25, no. 1, 2020.

[24] C. J. Lee and D. Moher, "Promote scientific integrity via journal peer review data," *Science*, vol. 357, no. 6348, 2017.

[25] G. Lewison, "Evaluation of books as research outputs in history of medicine," *Research Evaluation*, vol. 10, no. 2, 2001.

[26] X. Li and M. Thelwall, "F1000, Mendeley and traditional bibliometric indicators," in *International Conference on Science and Technology Indicators*, ser. STI, 2012.

[27] D. Lindsey, "Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid," *Scientometrics*, vol. 15, no. 3-4, 1989.

[28] Y. Liu, W. Zuo, Y. Gao, and Y. Qiao, "Comprehensive geometrical interpretation of h-type indices," *Scientometrics*, vol. 96, no. 2, 2013.

[29] B. Lutz, "Do Altmetrics point to the broader impact of research? An overview of benefits and disadvantages of Altmetrics," *Journal of Informetrics*, vol. 8, no. 4, 2014.

[30] H. Moed, *Citation Analysis in Research Evaluation*. Springer, 2005.

[31] E. Mohammadi and M. Thelwall, "Assessing non-standard article impact using F1000 labels," *Scientometrics*, no. 97, 2013.

[32] J. Nicolaisen, "Citation analysis," *Annual Review of Information Science and Technology*, vol. 41, no. 1, 2008.

[33] A. G. Nuzzolese, P. Ciancarini, A. Gangemi, S. Peroni, F. Poggi, and V. Presutti, "Do Altmetrics work for assessing research quality?" *Scientometrics*, vol. 118, no. 2, 2019.

[34] J. Ortega, "Exploratory analysis of Publons metrics and their relationship with bibliometric and Altmetric impact," *Aslib Journal of Information Management*, vol. 71, no. 1, 2018.

[35] B. Patro and A. Aggarwal, "How honest is the h-index in measuring individual research output?" *Journal of Postgraduate Medicine*, vol. 57, no. 3, 2011.

[36] J. Priem, P. Groth, and D. Taraborelli, "The Altmetrics collection," *PloS One*, vol. 7, no. 11, 2012.

[37] P. Regibeau and K. E. Rockett. (2016) Research assessment design and the role of bibliometrics. [Online]. Available: https://voxeu.org/article/using-bibliometrics-gauge-research-quality

[38] K. A. Robinson and S. N. Goodman, "A systematic examination of the citation of prior research in reports of randomized, controlled trials," *Annals of Internal Medicine*, vol. 154, no. 1, 2011.

[39] T. Ross-Hellauer and E. Görögh, "Guidelines for open peer review implementation," *Research Integrity and Peer Review*, vol. 4, no. 1, 2019.

[40] Y. Shakeel, J. Krüger, G. Saake, and T. Leich, "Indicating studies' quality based on open data in digital libraries," in *International Conference on Business Information Systems*, ser. BIS. Springer, 2018.

[41] Y. Shakeel, J. Krüger, I. von Nostitz-Wallwitz, G. Saake, and T. Leich, "Automated selection and quality assessment of primary studies," *Journal of Data and Information Quality*, vol. 12, no. 1, 2020.

[42] H. Shema, J. Bar-Ilan, and M. Thelwall, "Scholarly blogs are a promising Altmetric source," *Research Trends*, vol. 37, 2014.

[43] K. Siler, K. Lee, and L. Bero, "Measuring the effectiveness of scientific gatekeeping," *Proceedings of the National Academy of Sciences*, vol. 112, no. 2, 2015.

[44] V. Spezi, S. Wakeling, S. Pinfield, J. Fry, C. Creaser, and P. Willett, ""Let the community decide"? The vision and reality of soundness-only peer review in open-access mega-journals," *Journal of Documentation*, vol. 74, no. 1, 2018.

[45] J. P. Tennant and T. Ross-Hellauer, "The limitations to our understanding of peer review," *Research Integrity and Peer Review*, vol. 5, no. 1, 2020.

[46] M. Thelwall, "The pros and cons of the use of Altmetrics in research assessment," *Scholarly Assessment Reports*, vol. 2, no. 1, 2020.

[47] R. van Noorden, "Online collaboration: Scientists and the social network," *Nature*, vol. 512, 2014.

[48] L. Waltman and R. Costas, "F1000 recommendations as a potential new data source for research evaluation: A comparison with citations," *Journal of the Association for Information Science and Technology*, vol. 65, no. 3, 2014.

[49] P. Wang, J. Williams, N. Zhang, and Q. Wu, "F1000prime recommended articles and their citations: An exploratory study of four journals," *Scientometrics*, vol. 122, 2019.

[50] D. Wardle, "Do 'Faculty of 1000' (F1000) ratings of ecological publications serve as reasonable predictors of their future impact?" *Ideas in Ecology and Evolution*, vol. 3, 2010.

[51] J. Wilsdon and H. Services, "The metric tide: Correlation analysis of REF2014 scores and metrics. (Supplementary Report II to the Independent Review of the Role of Metrics in Research Assessment and Management)," 2015.

[52] P. Wouters and R. Costas, "Users, narcissism and control—tracking the impact of scholarly publications in the 21st century," 2012.