Similarity-Detection Software Use by Scholarly Publishers

Jessica LaPointe

Introduction

Resources like CSE's *White Paper on Promoting Integrity in Scientific Journal Publications*[1] and the Committee on Publication Ethics (COPE) flowcharts[2] provide thorough descriptions of ethical misconduct and guidance for dealing with plagiarism and other ethical violations once they have been discovered. Scholarly publishers like the American Meteorological Society (AMS)[3] and the American Geophysical Union (AGU)[4] provide authors with their own policies regarding text reuse and plagiarism. Others, like the Proceedings of the National Academy of Science (PNAS),[5] direct authors to the COPE guidelines. However, questions remain regarding the most effective ways to use similarity-detection tools before violations occur.

As the use of text similarity-detection tools becomes more widespread across the industry, scholarly publishers are wrestling with a number of common questions: What approaches are other publishers taking with regard to issues of self-similarity and plagiarism? How are they using similarity-detection tools, if at all? What percentage of similarity is acceptable, and how is that percentage determined? What challenges do publishers face, and how are they being addressed? For the CSE Publication Certificate Program, I undertook a project to uncover trends in the scholarly publishing community's approach to similarity detection and identify best practices in the timing of using similarity-detection software (SDS) and workflow management (e.g., who does what when, and how to determine appropriate levels of similarity). Herein, I present my findings, including the survey results, along with some analysis and conclusions.

Results

Of the total of 44 respondents, the overwhelming majority (41) indicated they are currently using SDS. Similarity Check (by iThenticate) was identified as the most commonly used tool, but PlagScan, WCopyfind, Turnitin, Grammarly, and, generically, free online software were also mentioned. According to nearly 67% of respondents, "integration with existing software (e.g., submission/manuscript tracking system)" was the main reason for choosing a particular tool. In addition, 30% claimed "ease of use" and approximately 24% claimed price as the reason they chose one tool over another (for many questions, more than one option could be selected, so the percentages may not equal 100%). Several respondents provided additional comments, many of which indicated the decision was out of their hands or was made before they were hired at their present organization. Others mentioned the lack of choice because only one SDS was offered by their publisher or integrated with their manuscript tracking system. More than 65% have been using SDS for more than 2 years; one replied they have been offering it for over 5 years, but clients have been requesting it only within the last 2 years.

Roughly half the responses stated their organization changed its policies regarding text reuse/recycling and plagiarism as a result of using this SDS, and those changes were principally driven by staff; changes were driven by peer reviewers in only about 7% of the cases. The comments indicate the use of SDS has allowed staff to determine the extent of text recycling and develop more detailed guidelines in response.

While only about half of these organizations have changed their policies, almost 70% have changed their instructions for authors and editors. The comments indicate they provide information to authors and editors regarding their use of the software and clarify expectations for authors. It has also helped them to educate authors about the need to avoid recycling text without proper attribution. It seems one of the key benefits of using SDS is author education. Most respondents have guidelines on text reuse/recycling in their instructions for authors. Some evaluate instances on a case-by-case basis and may offer instructions to authors if necessary. Some direct authors to their Office of Research Integrity or provide authors with standard text from their publisher. Respondents mentioned the need for ongoing author and editor education to improve compliance with their guidelines.

Which papers get the SDS treatment? About 60% replied similarity-detection software is used on all papers. Nearly 17% replied that papers are chosen at the discretion of the editor, and just over 7% check a random selection of papers. Most of the comments revealed that SDS is applied to all the papers that have gone through peer review. In a few other cases, papers are checked only if the reviewer or editor suspects a case of plagiarism or text reuse/recycling. One reply indicated it would be done at the author's request, and another commented only review articles are checked. According to these responses, SDS is used primarily at initial submission (~54%; Figure 1), but the comments indicate papers tend to be checked after acceptance. One reply

CONTINUED

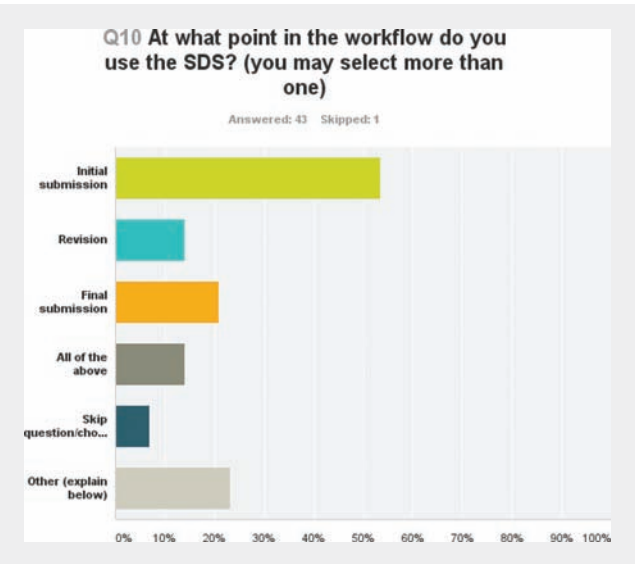

Figure 1. Graph of responses to Question 10: At what point in the workflow do you use the SDS?

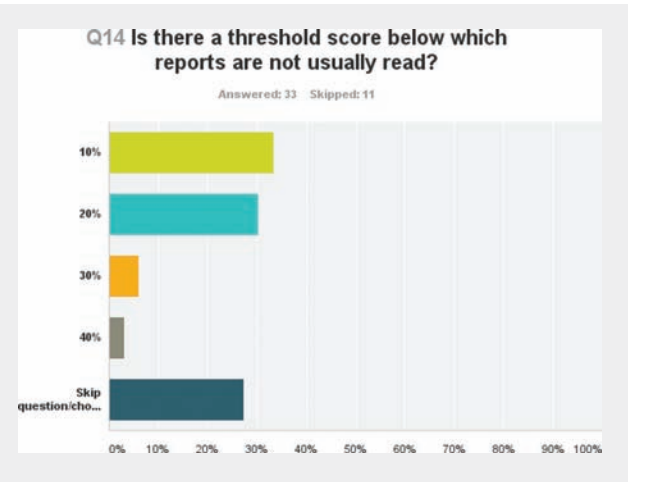

Figure 2. Graph of responses to Question 14: Is there a threshold score below which reports are not usually read?

explained that checking papers after acceptance, rather than upon initial submission, reduces the workload associated with processing the papers. Nearly 14% of respondents indicated SDS is used when an author submits a revision, and about 21% use it after the final submission.

Nearly 60% of respondents indicate they use SDS at the beginning of the workflow to weed out potential problem papers as early as possible; over 44% replied they do it to keep papers with serious problems out of peer review to save the peer reviewers time. Others identified cost and staff/volunteer time as reasons they use SDS at a particular point in the workflow. One detailed comment explained that checking all papers would be more trouble than it is worth, considering text reuse/recycling is not that much of a problem for that organization. Because the percentage match can be misleading, reports are read carefully, which takes time but results in a more accurate assessment of the paper.

The responsibility for reading the reports that are generated by the SDS falls both to "editorial board members (chief editors, editors, etc.)" and "staff (e.g., production staff, copy editors, technical editors)". As expected, the reports are analyzed by the same people who are responsible for making the final decision to accept or reject the paper, which in some cases is a paid staff member (like the managing editor) and in other cases is the editor-in-chief.

The percentage similarity match score may be a "red herring," but having a threshold for not needing to read the report would surely save peer reviewers and staff some precious time. Over 60% of the responses indicate a similarity score of 20% or less means the report will likely not be thoroughly read (Figure 2). Two respondents indicated they would likely not read the report with a score of 30% or less, and one respondent gave a threshold of 40%. Nearly a quarter of respondents chose to skip this question, but there were several detailed comments. Some provided an alternative score, such as 5%, 15%, or 25% (and in one case, 50%), at which they would read the SDS reports. Most of the comments indicated that the complete report would be read, regardless of the similarity score. As far as why a particular threshold was chosen, responses ranged from "it's important to read every report" to "the threshold seems to let through almost all good papers while catching the majority of plagiarized ones." Many comments pointed to the value of experience when reviewing the reports, since a low similarity score might disguise the fact that an entire sentence was taken verbatim from another uncited source, while a high score might be the result of many short phrases or terms that appear in other papers, which does not constitute plagiarism or a violation of the publisher's ethical guidelines.

There was no clear consensus regarding a similarity score at which the report would always be read (Figure 3). Many replied that all reports are read, regardless of score. This highlighted the fact that SDS users seem aware of the limitations of similarity scores: they can only indicate the percentage of text that matches an existing publication, but they give no indication of the nature of the text recycling and whether it constitutes a problem that needs to be addressed. Experience was cited as key in deciding how to use the SDS scores. Once editors gain familiarity with the tool and the similarity reports, they are better able to gauge what similarity score might indicate an actionable problem with the text. Respondents clearly identified "methods" sections and references as likely areas of text matching,

CONTINUED

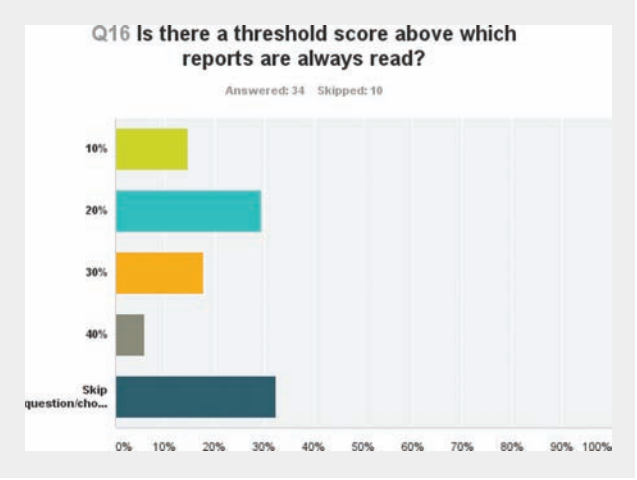

Figure 3. Graph of responses to Question 16: Is there a threshold score above which reports are always read?

and so reports would have to be analyzed to rule out these obvious sources.

Nearly 75% of respondents indicated a paper with an unacceptable (by their standards) level of text similarity would be rejected by their journal, but a large number would also recommend a "revise" decision, would ask the author to include appropriate citations during copyediting, or would put the paper on hold while the matter was investigated (Figure 4). Many would contact the author and explain the problem, even in the event of a rejection. In potentially serious cases of plagiarism, a few replied they would contact the author's institution. In cases where a problem has been identified, 40% of respondents allow authors to access the similarity reports themselves (i.e., by sending authors a PDF

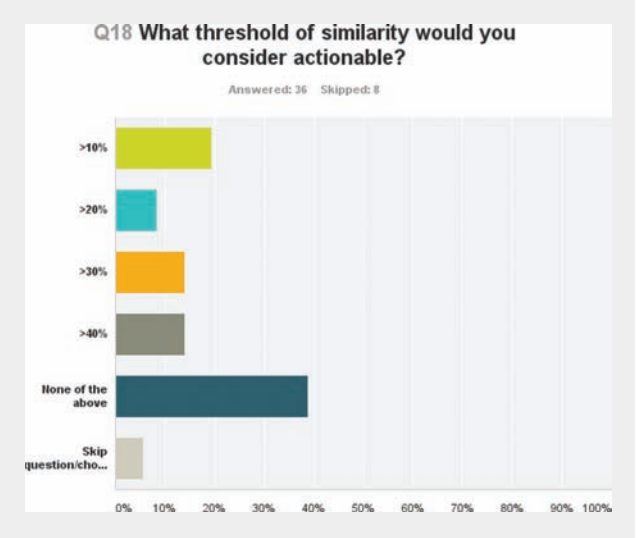

Figure 4. Graph of responses to Question 18: What threshold of similarity would you consider actionable?

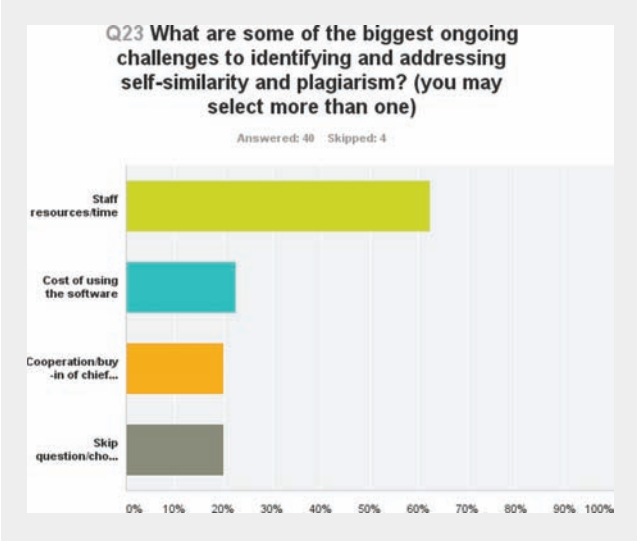

Figure 5. What are some of the biggest ongoing challenges to identifying and addressing self-similarity and plagiarism?

of the report). This allows authors to see exactly where the problems are and can help them better understand what they need to do to make their paper ready to pass peer review.

Twenty-five of forty respondents (62%; not all respondents answered all questions) identified staff resources and time as the biggest ongoing challenge to using SDS, followed by cost of using the software and editor cooperation (Figure 5). A number of comments pointed out that getting authors to understand what they might have done wrong—for example, it is not acceptable to reuse large sections of text from one's own articles—is one of the biggest challenges they face. Another challenge is the fact that these tools are deployed mainly at the discretion of the editors, each of whom may apply different standards and requirements to the papers submitted to their journals. A few noted the interface of the SDS was insufficiently user friendly, but this might improve as more publishers adopt the software, the interfaces become more intuitive, and users become more familiar with them.

Discussion and Conclusions

I originally hypothesized SDS users might have developed common approaches and even established some general standards for evaluating SDS reports and similarity scores. According to the survey results, SDS users have taken largely similar approaches to using these tools, but not in the way I imagined. Unexpectedly, there appears to be little consensus regarding the percent of text similarity that requires a particular action. The majority of the responses indicate that users are aware of the limitations of the similarity scores—that is, they reveal little about the extent of text recycling, much

CONTINUED

less plagiarism—and rather than check the reports only if they hit a certain score, *all* reports are carefully read.

Curiously, several comments stated it would matter whether the text was being reused from another work by the same author(s), rather than whether it was from so-called gray literature (e.g., dissertations/theses, conference materials) versus copyrighted published work (e.g., journal articles, book chapters). According to these respondents, self-plagiarizing is a less serious offense than reusing text from a source not authored by the same person.

For many of the survey questions, the answers boiled down to "it depends," which may indicate that not only are cases of text reuse and recycling being taken very seriously by publishers, but they are also being approached with great care and deliberation.

Clearly, more research is needed on this fertile topic. But a few general themes arise out of these data. There is quite a bit of consensus among respondents in how SDS tools are used. The challenges and limitations also seem to be fairly common across the board. Editors and others who use SDS are circumspect in its use and application to address potential problems in manuscripts. Many enlist the authors themselves to correct problems and make sure all references are appropriately cited prior to publication. The use of SDS is likely to increase in the future, and it will continue to be a powerful tool for educating authors about appropriate reuse of material as well as how best to avoid problems that can lead to ethical lapses and retractions.[6,7]

Sincerest thanks go out to all the survey respondents, as well as to Anna Jester for "boosting the signal" by mentioning the survey on her LinkedIn page. Special thanks also go to Gwendolyn Whittaker, AMS Peer Review Support Coordinator, for help in developing the survey questions. To view the full survey results, go to https://www.surveymonkey.com/results/SM-8RZXD2XB/.

References

1. Scott-Lichter D and the Editorial Policy Committee, Council of Science Editors. CSE's white paper on promoting integrity in scientific journal publications, 2012 update. 3rd edition. Wheat Ridge, CO: 2012. http://www.councilscienceeditors.org/wp-content/uploads/entire_whitepaper.pdf.
2. Committee on Publication Ethics. Flowcharts. http://publicationethics.org/resources/flowcharts.
3. American Meteorological Society. Author disclosure and obligations. Boston, MA: AMS; 2010. https://www.ametsoc.org/ams/index.cfm/publications/authors/journal-and-bams-authors/author-resources/author-disclosure-and-obligations/.
4. American Geophysical Union. Scientific ethics for authors and reviewers. Washington, DC: AGU. http://publications.agu.org/author-resource-center/scientific-ethics-authors/.
5. Proceedings of the National Academy of Sciences of the United States of America. PNAS information for authors. Washington, DC: PNAS; 2018. http://blog.pnas.org/iforc.pdf.
6. Stern V. Prominent physicist loses four more papers for duplication. Retraction Watch. April 18, 2017. http://retractionwatch.com/2017/04/18/prominent-physicist-loses-four-papers-duplication/.
7. Oransky I. "Highly unethical practices" force four retractions for nanotech researcher. Retraction Watch. February 27, 2013. http://retractionwatch.com/2013/02/27/highly-unethical-practices-force-four-retractions-for-nanotech-researcher/.