The Journal of Open Source Software

The SAGE Rejected Article Tracker

Andrew Hails[1] **and Adam Day**[1]

1 SAGE Publishing, 1 Oliver's Yard, London, EC1Y 1SP **1**

DOI: 10.21105/joss.03348

Software

- Review
- Repository
- Archive

Editor: Daniel S. Katz

Reviewers:

- @mfenner
- @dhimmel

Submitted: 30 April 2021
Published: 03 August 2021

Summary

Over 3m peer-reviewed research papers are published in academic journals each year (Johnson, 2018). An unknown number of research papers are also rejected by peer-review. There is little understanding of what happens to those rejected papers.

CrossRef is a not-for-profit organisation which maintains a large set of metadata describing the majority of published peer-reviewed research papers. The SAGE rejected article tracker extracts knowledge from that dataset by analysing data from the CrossRef REST API.

Given metadata for a rejected article, the rejected article tracker will:

- search CrossRef's API to retrieve a list of possible matches and
- select the most likely correct result from that list using simple machine learning.

The target audience for the tracker is researchers studying rejected articles. The task performed by the tracker is record-linkage, i.e., finding the correct CrossRef metadata record given incomplete data about a paper. So, while the intended use of the tracker is to track rejected articles, it can also be used by researchers performing record-linkage for other reasons, such as connecting preprints to their published versions, e.g., work by Cabanac (2020). This is a particularly topical application at the current time due to the rapid growth of preprint servers in recent years (Hoy, 2020).

The tracker is available as a Python package with a temporary live demonstration scheduled to run until mid-2021.

Statement of need

As the rate of creation of research manuscripts continues to grow at a rapid pace, the need to understand the peer-review process, improve efficiencies and tackle abuse becomes all the more pressing.

Rejected article tracking has been performed in a number of research settings (Chung et al., 2020; Citerio et al., 2018; Docherty & Klein, 2017; Wijnhoven & Dejong, 2010). Typically, this is done by manually searching for rejected articles over a small dataset. However, commercial rejected article trackers are available (HighWire, 2020; Selmani, 2021). To date, the lack of open source tools has prevented easy acquisition of data on rejected articles for analysis.

Data acquired by rejected article tracking makes various insights into the peer-review and publication processes possible. For example:

- It is possible to measure the rate at which rejected articles are published and cited. This provides evidence for the effectiveness of journal peer-review in identifying (or failing to identify) flaws in research.

Rejected article tracking is also valuable to the study of scientific misconduct (examples: (Bozzo et al., 2017; Ding et al., 2019; Hesselmann et al., 2017)).

Common forms of author-misconduct can be identified and studied.

- Dual submission (where an author has submitted the same article to multiple journals simultaneously) can be detected retrospectively with a high-degree of confidence.
- In a similar way, self-plagiarism can potentially be detected quickly and cheaply by checking new-submissions against CrossRef with the tracker. However, the well-established CrossCheck service based on iThenticate should yield superior results.
- When a rejected article has been later published *and then retracted* due to fraud or other misconduct, this can allow the publisher who rejected the paper to identify that case of misconduct in their own part of the peer-review system.

Finally, the rejected article tracker can also be used to link preprints with their published versions. Due to the rapid recent growth in preprint servers (Fraser, 2021), there is a growing need to improve the data-quality surrounding preprints.

The rejected article tracker is set up, by default, to accept data in the format exported by ScholarOne, a popular system for managing peer-review. However, the input data required is minimal, so data from any peer-review management system should be easily adapted for the tracker. Instructions are given in the `Readme.md` file of the GitHub repository.

How the matching algorithm works

The CrossRef API is often used to perform record-linkage. A typical use-case is adding metadata to incomplete references in the reference-list of a research paper (Tkaczyk, 2018). Under this typical use-case, there are often other data available, such as journal name and publication date as well as issue, volume, or page numbers. However, if we wish to track rejected articles, it is likely that we only have the title and author names for an article and there is a lower chance that it exists in CrossRef's data (since not all rejected articles are published). So, searching the API for just these 2 things often results in incorrect results being retrieved.

We begin with a dataset of ArXiv preprint metadata retrieved from the ArXiv OAI-PMH API. This dataset resembles journal submission data in that it includes the titles and author names of preprints. In many cases, this data also includes the DOI of the same article when it was published. The published version of the article is known as the "version of record" (VOR). This means that we know the correct result of a record-linkage process for this article. We find that the title and author lists are not always identical. Titles often undergo minor (and occasionally major) changes in the time between appearing as a preprint and publication. Author lists can also change in a number of ways (full names might be used instead of initials, or perhaps new authors are added to an author list at some point in the process).

We search the CrossRef API for each preprint's DOI as well as the best incorrect search result. This means that we can fill out 2 rows of a table of data for each preprint.

ArXiv title	ArXiv authors	VOR title	VOR authors	correct/ incorrect
title1	author_list1	title2	author_list2	correct
title1	author_list1	title3	author_list3	incorrect

We then:

- Calculate the Levenshtein distance between the titles in each row. This is normalised to a number between 0 and 100 using the `fuzz.ratio` method from the Python `fuzzywuzzy` package.
- Normalise all author names to a single string of `first_initial+last_name` in lower case. Then calculate 2 boolean values: one showing if there is a 100% match in author lists and one showing if there is at least 1 author name matching in the 2 lists.

This gives us a table of numerical data:

levenshtein_similarity	authors_match_one	authors_match_all	label
98	1	1	1
70	1	0	0

Then this data can be used to create a Logistic Regression classifier model using Scikit-Learn with `label` as our target variable. The model essentially learns what the typical difference is between a preprint title and its published version - a bespoke form of fuzzy matching.

This model can then be used to classify search results from the CrossRef API in order to find the correct DOI, and other metadata, for a rejected article.

The complete code required to build and customise the training dataset is included in the SAGE Rejected Article Tracker.

The dataset

The training dataset is also useful for other tasks such as identifying duplicate submissions. For example, if an author submits a paper to two or more journals at once, fuzzy matching on titles and author lists is an effective way to identify this behaviour.

A dataset similar to the one used to train the SAGE Rejected Article Tracker is available to download from Zenodo (Day, 2021).

Acknowledgements

We thank Helen King and Martha Sedgwick for support and advice in the development of this application. We also thank the community at *Journal of Open Source Software* (JOSS) for knowledgeable, patient and highly constructive feedback, particularly: Martin Fenner (Reviewer), Daniel Himmelstein (Reviewer) and Daniel S. Katz (Editor).

References

Bozzo, A., Bali, K., Evaniew, N., & Ghert, M. (2017). Retractions in cancer research: A systematic survey. *Research Integrity and Peer Review*, *2*(1), 5. https://doi.org/10.1186/s41073-017-0031-1

Cabanac, T. B., Guillaume; Oikonomidi. (2020). Day-to-day discovery of preprint–publication links. *Scientometrics*, *126*(6), 5285–5304. https://doi.org/10.1007/s11192-021-03900-7

Chung, S., Lee, J., Yoo, T. H., & Kim, G. H. (2020). The fate of manuscripts rejected from kidney research and clinical practice. *Kidney Research and Clinical Practice*, *39*(2), 230–231. https://doi.org/10.23876/j.krcp.20.392

Citerio, G., Deutsch, E., Sala, E., Lavillonnière, M., Perner, A., Jaber, S., Timsit, J. F., & Azoulay, E. (2018). Fate of manuscripts rejected by Intensive Care Medicine from 2013 to 2016: a follow-up analysis. *Intensive Care Medicine*, *44*(12), 2300–2301. https://doi.org/10.1007/s00134-018-5407-2

Day, A. (2021). *SAGE rejected article tracker training data* (Version 20210722a). Zenodo. https://doi.org/10.5281/zenodo.5122848

Ding, D., Nguyen, B., Gebel, K., Bauman, A., & Bero, L. (2019). Duplicate and salami publication: a prevalence study of journal policies. *International Journal of Epidemiology*, *49*(1), 281–288. https://doi.org/10.1093/ije/dyz187

Docherty, A. B., & Klein, A. A. (2017). The fate of manuscripts rejected from Anaesthesia. *Anaesthesia*, *72*(4), 427–430. https://doi.org/10.1111/anae.13829

Fraser, L. A. D., Nicholas AND Brierley. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*, *19*(4), 1–28. https://doi.org/10.1371/journal.pbio.3000959

Hesselmann, F., Graf, V., Schmidt, M., & Reinhart, M. (2017). The visibility of scientific misconduct: A review of the literature on retracted journal articles. *Current Sociology*, *65*(6), 814–845. https://doi.org/10.1177/0011392116663807

HighWire. (2020). *HighWire analytics*. https://www.highwirepress.com/solutions/highwire-vizor/

Hoy, M. B. (2020). Rise of the Rxivs: How preprint servers are changing the publishing process. *Medical Reference Services Quarterly*, *39*(1), 84–89. https://doi.org/10.1080/02763869.2020.1704597

Johnson, A. M., Rob; Watkinson. (2018). *The STM report*. https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf

Selmani, N. (2021). *Using dimensions to track and report on rejected submissions*. https://www.dimensions.ai/blog/using-dimensions-to-track-and-report-on-rejected-submissions/

Tkaczyk, D. (2018). *Reference matching: For real this time*. https://www.crossref.org/blog/reference-matching-for-real-this-time

Wijnhoven, B. P. L., & Dejong, C. H. C. (2010). Fate of manuscripts declined by the British Journal of Surgery. *British Journal of Surgery*, *97*(3), 450–454. https://doi.org/10.1002/bjs.6880