UKSG

Start-up story
Penelope

James Harwood is the founder of Penelope, an automated tool that checks manuscripts and gives immediate feedback to authors before they submit to a journal. It adds comments to a manuscript, just as a real editor might, and tells authors what they must fix to comply with journal requirements and best research practices. He hopes it will make publishing faster and easier, whilst also educating authors about research integrity and best practice.

JAMES HARWOOD

Founder
Penelope Research

How I got started

I have always been devoted to academia, but publishing has always been a pain point. It takes months (or years) to get an article published, and journal instructions can be hideously long. And yet what does get published often is not as good as it could be. My neuroscience lab at Oxford would get frustrated when poor reporting prohibited us from replicating other people's experiments. In fact, this is a problem afflicting the vast majority of published biomedical literature.[1]

I first became interested in text mining through a side project with a friend. Our plan was to mine Twitter to find sport fanatics and build an automated betting machine. Although better than chance, we could never beat the odds. Nonetheless, I saw potential to apply text mining in academia.

Digging around to see what others had already tried, I found numerous projects aspiring to mine published literature for knowledge discovery.[2] I found software packages that could check statistics,[3] or extract specific information from clinical trials.[4] But all of these tools were to be deployed over published articles. None were aimed at helping authors fix mistakes before they get published, and most were academic projects, not mature enough to be scaled or adopted by publishers. None seemed to be widely used.

I started reaching out to publishers and discovered a passionate, enthusiastic community in London that desperately wants to innovate but does not have the expertise. Other than Word macros and plagiarism checking, editors were using few automated methods for assessing submissions.

'it will make publishing faster and easier, whilst also educating authors about research integrity and best practice'

The product

I set out to give publishers a way of using natural language processing to read manuscripts and give immediate feedback to authors.

It had to:

- be easy to use

- be fast

- give feedback that is simple yet educational

- address publisher needs and requirements as well as those of the broader research community.

For maximal impact, I wanted the tool to be available on journal websites and submission portals. This meant it had to be customizable on a journal-by-journal basis, and compatible with existing software infrastructure.

And, of course, it had to have a name. Drawing on the academic's fail-safe method, I worked out that 'publishers' natural language processing' gives the acronym PNLP, hence the name Penelope.

A few months later and Penelope was born. It is a simple widget that can be embedded into any web page, or linked to from a submission system. Authors upload their manuscript, answer a few questions and, within a minute, the file is returned with comments on as if a real editor had checked it.

'within a minute, the file is returned with comments on as if a real editor had checked it'

It checks that files include everything a journal requires and that it is formatted correctly. It checks section headings, declarations and ethical statements. It checks title pages for author and funder information, and ensures abstracts are structured properly with the correct subheadings. It cross-checks citations and ensures every table and figure has a legend.

These were the easy checks to build. The research integrity checks – what I am really interested in – were a little harder. Penelope can already double-check statistics and that the underlying raw data has been published. When relevant, it provides the author with any reporting guidelines they might be expected to follow (e.g. the CONSORT statement for clinical trials), and encourages reporting of randomization, blinding and sample size calculations.

The version available on the Penelope website is a generic one. The checks and advice are all based on general guidelines. But I have also made versions tailored to specific journals. For example, the journal *Addiction* has a version tailored to its author guidelines. It was easy to set up – it just took one phone call – and now it is embedded on their author guideline page and linked to their submission system. It only performs the checks they want, and the wording of the feedback is based on their own author guidelines.

Successes

A thousand authors have now used Penelope and given it an average eight out of ten for satisfaction. Authors like how quick and easy the tool is. They are really grateful when it catches things they have overlooked, and many authors return to use the tool again.

'A thousand authors have now used Penelope'

Increasingly, I see traffic coming from Asia and Africa, and looking at feedback I see users range from experienced professors to junior researchers. I am thrilled about this as I really hope Penelope will help researchers that are less experienced or for whom English is not their first language.

The algorithms have performed well 'in the wild'. I have been monitoring and improving them since the start to get them above 95% accuracy. There will always be instances where Penelope makes a mistake but authors seem tolerant of imperfection. I suppose most people are familiar with other tools that use machine learning – like voice recognition or recommendation systems – and understand their limitations.

Future plans

Currently I am working on opening up an API (application programming interface) so that publishers can build their own interfaces to display the results of Penelope's automated checks to editors and peer reviewers.

I am also planning to incorporate funder mandates. When it finds a funder, it will automatically check that the manuscript meets funder requirements (e.g. data mandates) and that the target journal has a compatible open access policy.

Long term, my intention is to focus more on research integrity. I have been surprised by how many statistical errors it has already found, so I would like to build this out to cover more statistical techniques. I will add checks for methodological reporting, based on reporting guidelines curated by the EQUATOR Network and BioSharing. These guidelines are a great resource for authors, but they are hard to find and it is too complicated for journals to enforce them. I hope that automating them will increase their impact and make it easy for journals to improve their standards.

> 'Long term, my intention is to focus more on research integrity'

Right now the tool is aimed predominantly at scientific research. Most of its checks do generalize across other disciplines, however, and I would love to collaborate with experts from the social sciences, humanities and arts to make sure the tool addresses their needs too.

How could we get more innovation within academia?

My main challenge has been funding. My initial funding came from two small grants, one from the UK Government and another from Digital Science. This was enough to get the project going and sustain a cheap existence.

A typical start-up route would then be to secure investment. So far, I have managed to survive without seeking investment, and so Penelope is totally independent. This has not been easy, and it is rare for a start-up to be so lean. Most would seek investment from a business angel or venture capital firm, but academia is a hard sell. It is a strange, niche market that few investors currently understand.

> 'a strange, niche market that few investors currently understand'

Looking at others innovating in the academic space, you see that most have received investment from relatively few publishers and are now partly or fully owned by them; Macmillan has the Digital Science companies, Elsevier has the Mendeley suite and SAGE recently invested in Publons. Besides financial support, publishers bring expertise – such as product development, sales and marketing – to help innovators grow into sustainability.

Investing in start-ups is a strategic move for publishers that will reassert their importance in the academic playing field. If other publishers want to disrupt the current power dynamic, then they should consider diversifying the kinds of support and partnerships they offer to innovators.

Funders should consider this too. A common condition of grant funding is to make the code open source which, strategically, does not always make sense, and I have not yet heard of a grant funder that provides assistance beyond finance, as publishers do.

A second challenge I have encountered is that it is not always simple for a business to collaborate with an academic. There are no clear protocols, no standard agreements, and a general wariness amongst academics to trust corporations. Personally, I think that the academic industry would benefit from more interaction between its stakeholders, and so I was delighted to learn of a new research-on-research PhD programme where candidates will spend time working with publishers.[5] I hope to see more collaborative projects like this in the future.

> 'the academic industry would benefit from more interaction between its stakeholders'

Finally, I would advise any aspiring academic innovators to combat their perfectionist predispositions. Perfecting in private is a risky strategy when developing a product, as you can easily spend weeks or months building something the wrong way. I constantly have to quieten my inner perfectionist and force myself to get feedback as soon as possible.

Final thoughts

Running a start-up is hard work and emotional at times, but I love every minute of it. The academic industry is wonderful, frustrating, and has led to so many world-changing developments. It is good. But it is becoming even better. And I am excited to be part of that journey.

Penelope is a great tool that authors are finding useful. As it grows, I hope it will improve research integrity and educate authors whilst reducing publishing times and costs. It would not exist without all the wonderful people that have helped and guided me along the way. I would invite anyone interested in knowing more to contact me; whether you would like to try a tailored version for your journal, or whether you are an academic that would like to help shape its future, just drop me a line (james@peneloperesearch.com).

'I hope it will improve research integrity and educate authors whilst reducing publishing times and costs'

References

1. Glasziou P, Altman D G, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D and Wager E, Reducing waste from incomplete or unusable reports of biomedical research, *Lancet,* 2014, 383(9913), 267–276, DOI: **https://doi.org/10.1016/S0140-6736(13)62228-X** (accessed 3 Jan 2017).

2. Biocreative NLP tools: **http://biocreative.sourceforge.net/bionlp_tools_links.html** (accessed 3 Jan 2017).

3. Statcheck: **http://statcheck.io/** (accessed 3 Jan 2017).

4. Statreviewer: **http://www.statreviewer.com/** (accessed 3 Jan 2017).

5. MIROR PhD programme: **http://miror-ejd.eu/** (accessed 3 Jan 2017).

James Harwood
Founder
Penelope Research, GB
E-mail: james@peneloperesearch.com │ Twitter: @penelope_rsrch

ORCID ID: http://orcid.org/0000-0003-3530-3231