AI and NLP for Publishers

How Artificial Intelligence & Natural Language Processing Are Transforming Scholarly Communications

A report from

KGL KnowledgeWorks Global Ltd.
CJK GROUP

ABOUT US

KGL KnowledgeWorks Global Ltd.
CJK GROUP

KnowledgeWorks Global Ltd. (KGL) is the leading provider of editorial, production, online hosting, and transformative solutions for every stage of the content lifecycle. Serving the publishing industry for more than 200 years, KGL offers a full range of technology, content, and delivery solutions that drive revenue growth, streamline operations, and ensure editorial integrity. We are your source for XML-early workflows, intelligent automation, high-speed publishing, accessibility compliance, digital learning solutions, delivery of print and online products, and so much more.

kwglobal.com

CONTENTS

The publishing industry is at a critical tipping point. Artificial intelligence (AI) capabilities have advanced to a degree that scholarly publishers can actually automate significant portions of the publishing workflow. This has massive implications for academic publishers and their authors. From making the editing process more efficient to improving the quality of content published and distributing scientific outputs faster to the research community, AI has the potential to transform the scholarly information chain.

While AI is not entirely new to the publishing industry, it has, up until now, had minimal impact on the content creation and editing process. That is changing with advancements in natural language processing (NLP), a technology that identifies and analyzes the structure of words in order to understand the meaning of content.

With NLP technology, publishers can automate simple editing and formatting tasks and focus their energy on adding greater value to the content. They can also process more journal submissions or speed up tedious peer review without significantly increasing staff or production costs.

Further innovations in AI and NLP are on the horizon. In the not-too-distant future, AI technology will improve every stage of production — not only editing and formatting, but also the content creation itself. The result of these innovations will be near instantaneous publishing.

In this paper, we'll explore how AI and NLP are being used today in scholarly publishing and how it may impact the evolution of research. We'll also explore how the technology works and how publishers like Taylor & Francis, with the help of KnowledgeWorks Global Ltd. (KGL), are realizing the benefits of intelligent automation.

An Introduction to AI & NLP in Scholarly Publishing

What Is AI & NLP?

Before exploring what's possible with AI and NLP in scholarly publishing, it's important to first define these terms. AI is a process by which humans train machines to identify patterns and learn new patterns. It involves developing algorithms that enable machines to quickly process large swaths of data, identify the patterns within that data, and make decisions or recommendations based on that analysis. There are a number of subsets within AI, which span deep learning, natural language processing, cognitive learning, and more.

Often AI is used to analyze numbers. For example, a retailer could use AI technology to identify the best discount to offer individual consumers in order to increase the likelihood of a sale. With the right parameters set by the retailer — not providing a discount steeper than 40%, for instance — an artificially intelligent computer system can process and analyze past product and consumer data quickly and calculate the likelihood that a consumer will purchase when offered a certain discount in the future. The AI-driven system could then make a decision on the discount to offer.

Companies can also implement AI to analyze words, which is known as natural language processing (NLP). NLP incorporates grammar analysis into the machine-learning process. A computer program is trained to recognize the noun, verb, and object in a sentence, as well as the beginning of a sentence versus the end. NLP strives to understand the structure of words in order to discern their meaning.

How Publishers Can Use AI & NLP

Individuals actually interact with AI and NLP regularly in their daily life. One of the most common applications of NLP is when a person searches for a specific term online. The search engine not only looks for pages that feature words matching the search terms, but it employs NLP to understand the context of the keywords by analyzing their structure so that it can provide the most relevant results.

Similar categorization efforts have occurred in the publishing process. "In scholarly publishing, what has been going on for a long time is analyzing author articles to determine what the subject is and define what the metadata keywords are for the article," explains Evan Owens, VP of Technologies at KGL. "In the past, that's been the most common application for artificial intelligence and natural language processing."

Now, according to Owens, NLP has evolved to the point where it can actually play a role in copy editing. Beyond

"The more we can reduce the touch points and increase the speed through the system, using algorithms, rather than having to wait for humans to review, the faster we can publish content. That helps the immediacy of science."

simply grasping the subject matter of an article, NLP can actually discern the quality of the content in terms of both grammar and format, and assess the piece based on these factors. This is how KGL is implementing NLP technology today.

"We are using it most in the context of copy editing, in terms of reading the manuscript, trying to understand the manuscript, and determine if it's well written. Then we assign a score and even auto-fix some of the simpler grammar and punctuation mistakes," says Atul Goel, President of KGL.

"It's not just using AI to understand what the author is trying to say, it's understanding how to help the author say it in the best way," adds Owens.

KGL has implemented AI to enhance its Smart Edit product, which is part of KGL Publisher Suite. Smart Edit is a cloud-based tool that uses AI and NLP to analyze and edit content and streamline the production process. Smart Edit transforms raw manuscripts into structured XML content that can be converted into the publisher's desired format.

In the last year, KGL has ramped up its use of AI technology through its partnership with Taylor & Francis, a scholarly publisher looking to increase the speed of the editing process and better serve its authors with more timely publication.

Practical Applications for AI & NLP in Scholarly Publishing

A Case Study from Taylor & Francis

Taylor & Francis, an international academic publisher of books and journals, has been mulling over an NLP solution for its production process for several years, says Stewart Gardiner, Global Production Director, Journals. "We've been working on projects to examine how we can improve publication speed through a much better analysis of the content that we publish. We wanted to tailor the level of editing intervention to the language quality of the articles that are accepted into our journals, rather than applying a one-size-fits-all intervention."

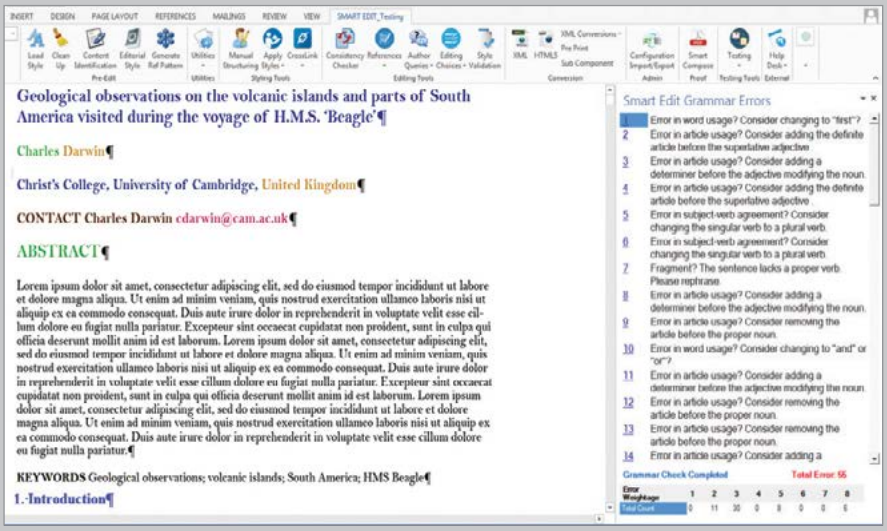
Traditionally, all articles submitted to an academic journal undergo a similar production process. An editor reviews that article and makes corrections. The editor sends the article back to the author for additional changes. This back and forth may occur several times before the manuscript is passed on to the copy editor who makes grammatical corrections or formatting changes to meet the journal's specific style. Then the production team composes the article into the appropriate layout for publication. The editor and author conduct a final proof. And, finally, the article is published. All told, this process could take several weeks.

Aside from taking a great deal of time and energy, this universal approach overlooks the articles that may not need as much editorial intervention, says Gardiner. That is why Taylor & Francis, working with KGL, developed a new system that takes into account the unique needs of each article and the requirements of specific journals.

The AI and NLP technology analyzes each article it encounters with pre-set grammar and formatting rules. After analyzing the content, the technology scores whether or not the article is of good, publishable quality. The technology will automatically correct minor errors like grammar and punctuation, and flag more complex issues that may need an editor's attention.

Although this may sound similar to spellcheck, Owens says that Smart Edit is far more advanced. A typical grammar and spellcheck tool will flag important parts of a journal article, like the metadata, captions, or running heads as misspelled or grammatically incorrect. While this content is required for most journal publications and incredibly valuable for formatting and categorization purposes, a spellchecker will try to change this content to fit the sentence and paragraph structure of narrative content. KGL's technology, on the other hand, is tailored to comprehend the unique properties of scholarly content and incorporate the requirements of specific publications.

"If the natural language processing tool assesses that the input quality is good, then we put that into a specific production track, which goes straight through to the typesetting stage and skips the copy edit," says Gardiner.



KGL's Smart Edit tool flags content that requires an editor's review.

Articles that require more editing and revision are assigned to a different track and receive greater attention from copy editors.

"This system allows for a fairly intelligent way of looking at content at an article, journal, and publisher level, understanding the requirements of each level, and placing content into the right editing stream," says Goel.

While KGL and Taylor & Francis are still in the early stages of implementing this AI solution, the anticipated benefits are significant. Because editing is often the most time-consuming part of the production process, fast tracking high-quality articles to the composition stage can save a significant amount of time for publishers. "Using this technology, publishers could save between 10 and 40% of production time," says Goel.

Gardiner estimates that 25% of Taylor & Francis's journal submissions are high-quality and can advance straight to the typesetting and composition stage. While the publisher has not yet reached this level of automation, Gardiner expects that within the next few years Taylor & Francis will be able to streamline production for these top tier submissions using the intelligent features in Smart Edit.

The potential time and resource savings are significant, but Gardiner adds that KGL's technology can improve the author experience as well. The speed with which Taylor & Francis can publish an article will likely incentivize more authors to submit articles to its journals.

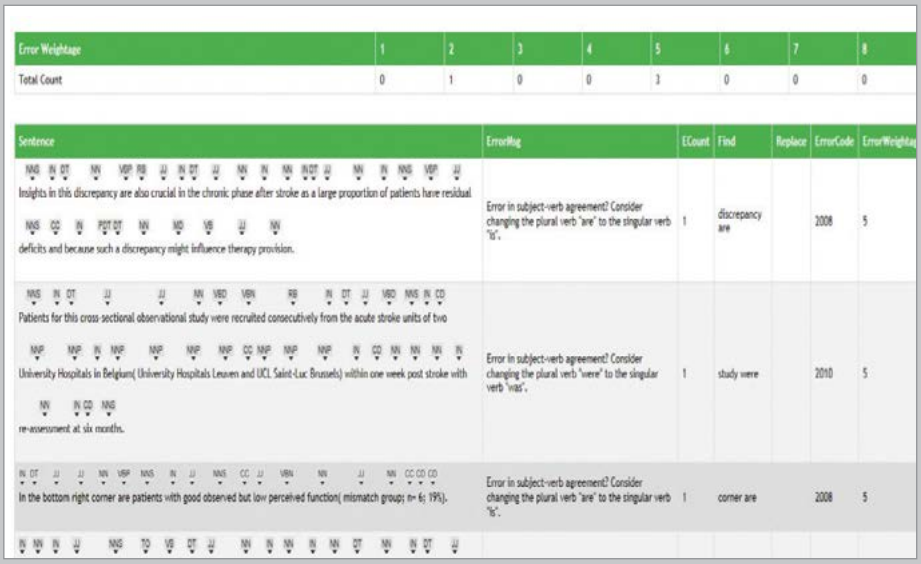
"From an author's perspective, they've had their article accepted on Monday, and then on Tuesday, their proof arrives for them to review. We think that is really the level of customer experience for our authors that we need to be aiming for, and matches the expectation, increasingly, of the younger generation of researchers," says Gardiner.

There are also benefits to the scholarly community at large when articles are published faster. "The more we can reduce the touch points and increase the speed through the system, using algorithms, rather than having to wait for humans to review, the faster we can publish content," says Gardner. "That helps the immediacy of science. It allows us to deliver important research and ideas to the scholarly community faster."

The AI behind KGL Publisher Suite

The **Smart Edit** AI and NLP technology is powered by a rules-based system. Editors and programmers inform grammar rules and style rules that may be specific to a certain publisher or journal. For example, if a scientific journal needs specific funding information cited in an article, publishers can develop rules that instruct the algorithm to look for and identify this data. Essentially, the rules tell the AI technology what word patterns to look for and how to identify patterns that are missing or incorrect.

"There are some edits the AI technology can identify and make the necessary changes. Those are primarily the instances where the rules are fairly black and white," says Goel. "But wherever there is an issue of interpretation, ambiguity, and context, the AI highlights



Error Weightage	1	2	3	4	5	6	7	8
Total Count	0	1	0	0	3	0	0	0

Sentence	ErrorMsg	ECount	Find	Replace	ErrorCode	ErrorWeightage
Insights in this discrepancy are also crucial in the chronic phase after stroke as a large proportion of patients have residual deficits and because such a discrepancy might influence therapy provision.	Error in subject-verb agreement? Consider changing the plural verb "are" to the singular verb "is".	1	discrepancy are		2008	5
Patients for this cross-sectional observational study were recruited consecutively from the acute stroke units of two University Hospitals in Belgium(University Hospitals Leuven and UCL Saint-Luc Brussels) within one week post stroke with re-assessment at six months.	Error in subject-verb agreement? Consider changing the plural verb "were" to the singular verb "was".	1	study were		2010	5
In the bottom right corner are patients with good observed but low perceived function(mismatch group; n= 6; 19%).	Error in subject-verb agreement? Consider changing the plural verb "are" to the singular verb "is".	1	corner are		2008	5

Smart Edit assigns scoring metrics for each grammatical error category.

that section and presents options so the editor can review and make a decision about what needs to be done."

While many rules are created manually for each publisher that works with KGL, Owens envisions a future where we will use AI and NLP to understand the style and formatting requirements of different publications simply by analyzing them. In a few years, he predicts, the rules that guide AI decision making will be somewhat automated. "Artificial intelligence can look at existing publications to determine standards in terms of formatting and editorial style. That will reduce the need for a lot of manual programming," says Owens.

KGL plans to roll out additional AI enhancements to the rest of its Publisher Suite, which includes Smart Compose, Smart Proof, and Smart Track.

Smart Compose ingests structured XML created by Smart Edit and formats that content based on the formatting rules and style of a specific publication. The goal is to further automate composition and improve the overall composition quality using AI and NLP.

Smart Proof is a tool that allows KGL to share an online proof of the composed article with editors and authors where they can input in-line edits and queries directly into the XML using a user-friendly interface. These edits are tracked and approved directly in the platform. AI and NLP may eventually suggest corrections or formatting changes to authors and editors and reduce the time needed to proof.

Finally, **Smart Track** logs every transaction so publishers know exactly what edits were made to an article as well as the stage in the production cycle. AI technology can power automatic alerts to publishers and authors when human intervention is needed and further streamline the production process.

Ultimately, KGL's goal is to use this deeper understanding of publishers' content to automate further areas of the production process. "AI technology will evolve to all parts of production process and for certain publication types, it potentially could come close to an instantaneous publishing model," says Goel.

Looking Ahead

How to Get Started with AI & NLP

For publishers looking to follow in Taylor & Francis's footsteps, Goel recommends taking a close look at the goals of different stakeholders in the business. He acknowledges that all scholarly publishers are trying to meet goals around quality, cost, and schedule. "It's not one or the other. It's what is the relative importance of those aspects, and how do publishers want to prioritize them? Editing is the step which takes the longest in

"AI technology will evolve to all parts of production process and for certain publication types, it potentially could come close to an instantaneous publishing model."

the current production environment, so if schedule is important, you might put greater emphasis on editing and automating this stage."

According to Goel, understanding the level and type of editorial intervention that is acceptable is key. For example, a publisher may want to make editorial interventions around the tone of the article to keep it consistent with the voice of its journal. The level of AI intervention the publisher might find acceptable may be lower than a publisher motivated primarily by speed.

Once publishers understand their goals for automation, Gardiner recommends partnering with a vendor to implement these solutions. And in particular, he says publishers should look for a vendor that already has advanced editing and structuring tools. For example, can the vendor easily translate raw manuscripts into structured XML?

"It's no use developing an AI tool unless the structuring and editing tool is already very advanced. That needs to be a core capability," says Gardiner.

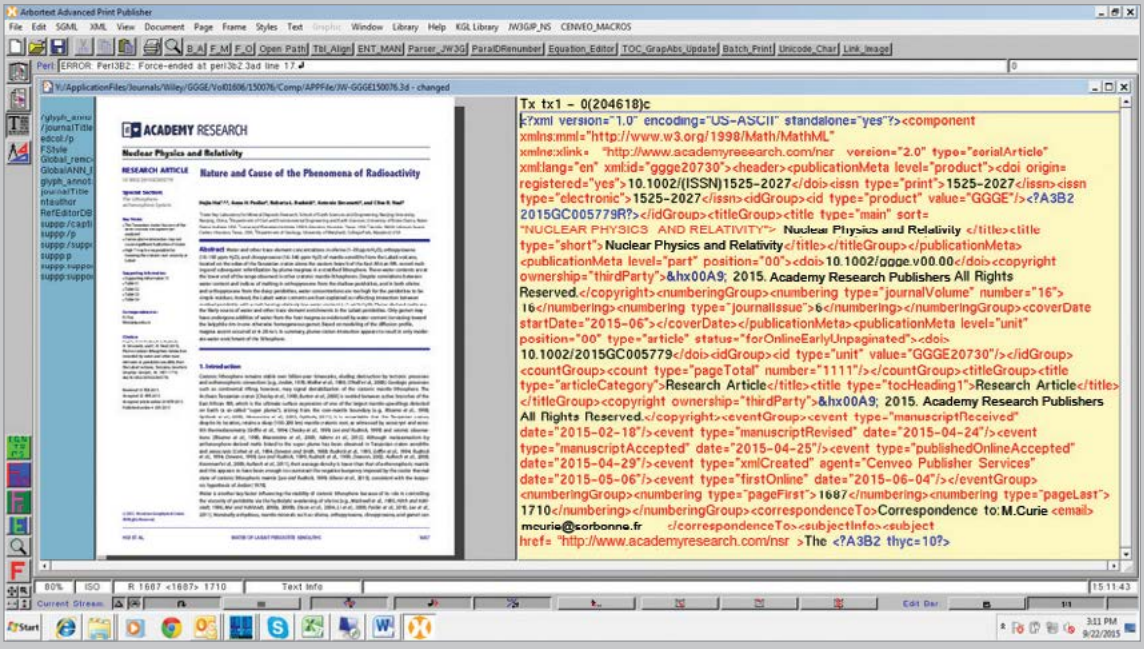
Examining a vendor's research and development investments is also important. Does the vendor have the structure in place to continue to develop and advance their AI solution? Gardiner says publishers should research the size of a vendor's development team, the amount of energy they place on studying next generation technology like AI, and how the vendor is implementing that technology into publishing workflows.

"There's a great deal of variation from one vendor to another in terms of research and development and the amount of revenue they're reinvesting into that area of the business. I think that needs to be something that publishers look at when they assess suppliers," says Gardiner.

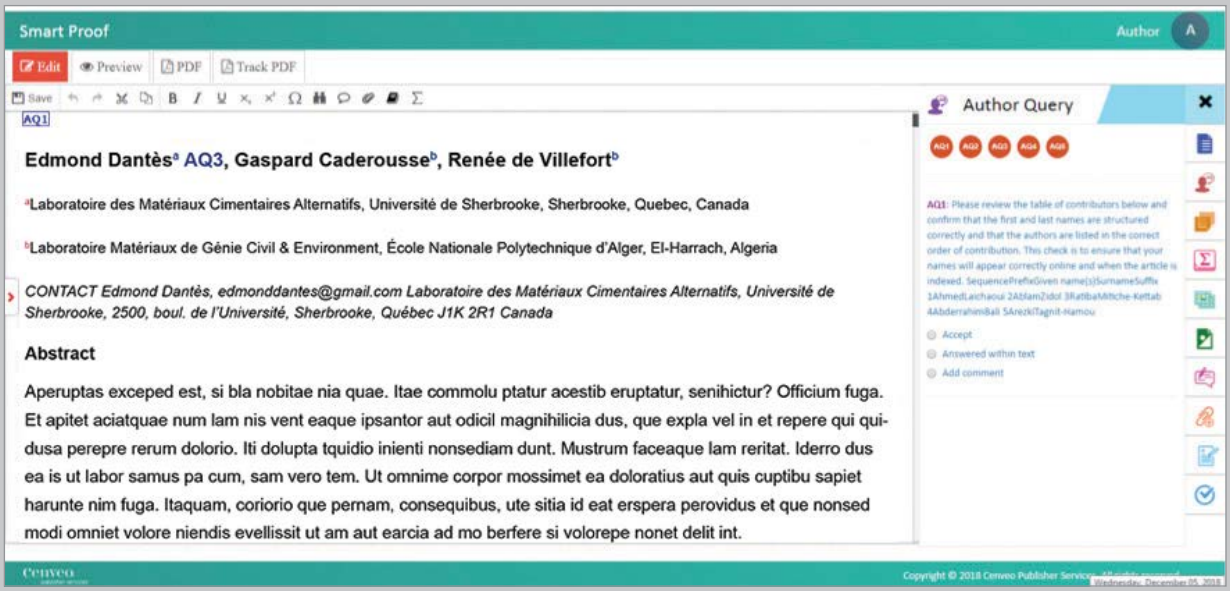
The Future of AI & NLP in the Scholarly Ecosystem

Currently AI and NLP largely impact the production process at Taylor & Francis, but Gardiner sees opportunities for this technology to streamline everything from content creation to peer review.

"I can imagine a future in which some companies offer authors the ability to analyze their article and see what



Smart Compose formats structured XML into the appropriate layout style for a specific journal.



Smart Proof delivers a web-based proof to editors and authors for close review.

journal they should submit it to — a kind of journal matching service," says Gardiner.

Likewise, the technology could analyze a database of scholars who have expertise in different areas of scholarship and match the best individuals with a scholarly article to peer review. Not only is this an enhancement that will speed production, but Gardiner anticipates it will vastly improve the quality of many scholarly articles.

AI technology can even resolve issues of plagiarism, says Gardiner. Once AI technology has analyzed a variety of journals and understands the content that exists on certain subject matters, it can identify when sections of articles have been plagiarized or where citations are missing or incorrect.

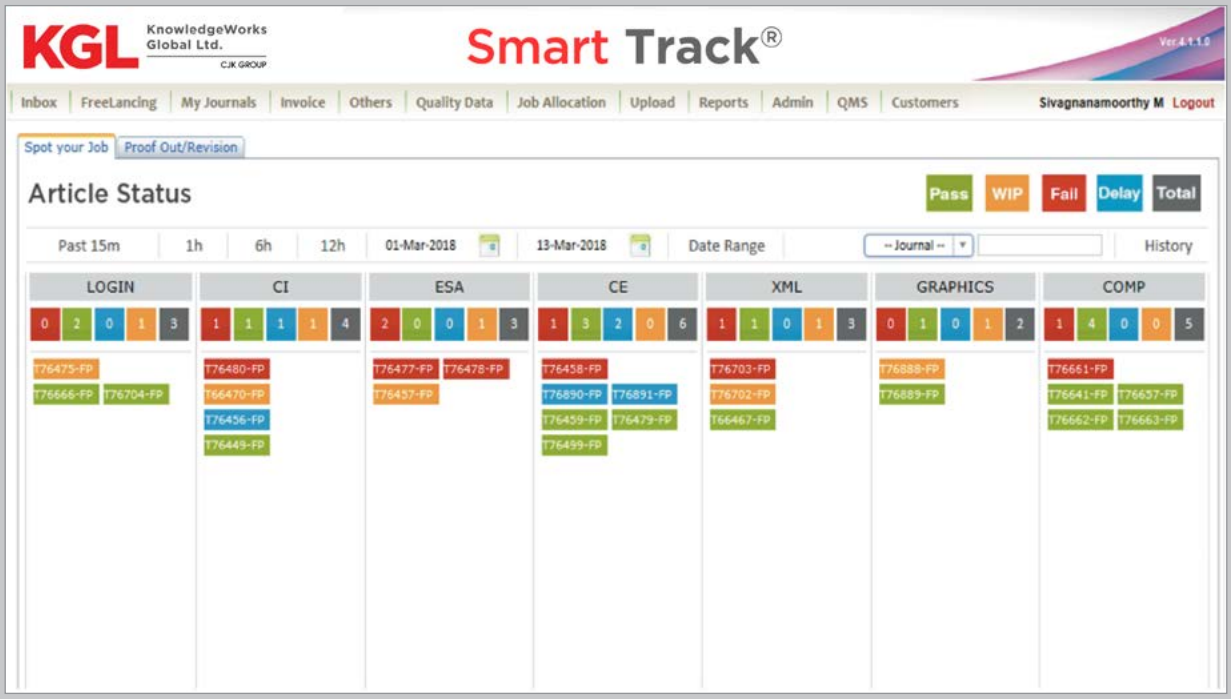
With further automation of content creation and production processes, some publishers may assume that the role of the editor will become obsolete. In Owens's view, AI will not replace editors, but rather it will empower editors to provide more value to their authors.

"I don't think editing will go away. It will be focused on the very high-value aspects," says Owens. "I think it will empower editors to focus on the most important value-adds, rather than having to worry about punctuation or just the ordinary logistics. Tidying articles up will be automated so that they can focus on more important issues like querying authors about the meaning of their articles."

Gardiner adds that the growth of AI adoption at Taylor & Francis is not motivated by the desire to replace existing staff, but to manage journal submissions that its current staff cannot handle while improving journal quality.

"Submissions to Taylor & Francis journals are growing year over year at 10%. We cannot increase headcount 10% each year, and we cannot spend 10% more on our publication and production budgets each year. One of the reasons why we've developed AI solutions is to try and cope with that growth without sacrificing quality and without increasing costs," says Gardiner.

"AI will not replace editors, but rather it will empower editors to provide more value to their authors."



Smart Track documents each change made to an article and tracks its progress through the production cycle.

Looking ahead, Gardiner is particularly excited by the impact AI can have on authors' work. He envisions a self-service model that limits the number of touch points for authors and helps them publish their content more quickly.

"It's about that sense of delighting our customers, by having their article accepted on the Monday, and ready for review on a Tuesday. I think the potential to get to that kind of author experience really is a game changer. For a company like Taylor & Francis, which aspires to be the place where authors want to publish, I think it would be an immense achievement if we got to that speed of publication with a good proportion of our articles."

Goel says that KGL's work with AI and NLP is just the beginning. "We're seeing the first wave of implementation, but this technology is here to stay. There is huge opportunity for publishers to adopt this technology and really improve every stage of the publishing process," says Goel.

kwglobal.com

North America | Europe | Asia

info@kwglobal.com

@KwGlobalLtd