White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/219963/

Version: Published Version

Book Section:
Whittle, S., O'Sullivan, J. and Pidd, M. (2023) AI and the Editor. In: Hegland, F.A., (ed.) The Future of Text IV. Future Text Publishing, pp. 108-111. ISBN: 9798870688060.

https://doi.org/10.48197/fot2023

© 2023, the Authors. Published by Future Text Publishing. This work is freely available digitally, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

AI and the Editor

Sophie Whittle, James O'Sullivan, Michael Pidd

Digital scholarly editing remains an industrial craft: the materials, medium and methods are technological, but the work itself remains largely manual and bespoke. And because digital editions are labour intensive, they can be limited in scale. Editors—that is, textual scholars and the makers of editions—were among the first in the arts and humanities to recognise the publishing affordances of the digital. And so it is surprising that machine learning and natural language processing have not yet played a greater role in scholarly editing; that newer forms of computation have not advanced editions to the same degree as markup languages did in the final decades of the twentieth century.

It can be easy to get caught up in looking only at the latest, most novel forms of artificial intelligence, for example, using OpenAI's famed ChatGPT to assist with annotations and contextualising materials. But even long-established symbolic and statistical natural language processing techniques could be used to semi-automatically classify and understand the context of words and terms, allowing other kinds of insights into sources. Text collation to identify variants has benefited from computer assistance [1], and certainly, while methods which offer fresh (ie. data-driven) insights are naturally privileged over those which can only offer expediency, there is substantial value in computational tools that can reduce the amount of routine—even tedious—tasks that are required by digital scholarly editors. Natural language processing can assist with time-consuming editorial tasks such as annotating, glossing and connecting texts; such a return not only shortens timeframes to publication (and thus, public consumption), but also allows more time for critical examination (the bit that makes the editing, scholarly). One can only assume that the general lack of adoption of such tools and techniques among scholarly editors is a consequence of ideological opposition, lack of access to the necessary computational expertise, or simply ignorance of their value; certainly, the opportunities are there.

There are many reasons why one would object to the use of AI in editing: the (often valid) concerns over representativeness and accuracy of semi-automated processes [2], and the lack of transparency in the production of (some) tools and processes themselves. The introduction of AI to scholarly editing processes might involve additional work on the part of the editor to ascertain the sources and methods adopted by the model. Editors should also be mindful that any rush to AI might replicate the consequences of the rush to OCR and mass digitisation that occurred in the 2000s and 2010s, which precipitated poor quality outputs of limited use for research, squandering vast amounts of resources and labour.

Yet, there is still value in collaborating with AI, despite having to maintain some of the necessary work of critiquing the output. While there are shortcomings of manual validation in terms of slowing down the technological task in the short term, the method of curating detailed prompts to input into AI models—and subsequently investigating what the AI has come up with—can in fact enrich digital scholarly editing and make the process quicker overall.

At the most basic level, one can certainly see potential in AI as a means of enabling editors and their audiences to intuitively approach their curated materials through distant ways of reading [3]. The majority of text-centred activities in the digital humanities fall into three broad categories: the assemblage of textual resources (including digital scholarly editing), the sharing of such resources (digital publishing), and the use of computer-assisted techniques to quantitatively analyse cultural materials (distant reading or cultural analytics). Katherine Bode criticises the digital humanities for having separate "curatorial and statistical" dimensions, a disciplinary culture which is split between those who gather and edit, and those who analyse with machines [4]. That the aforementioned practices continue to develop in isolation has only served to slow the progress of each: digital scholarly editions and the practice of digital scholarly editing would benefit greatly from integration with data mining techniques and machine-assisted insights, and methods for computer-assisted analysis are only as good as the data on which they operate.

Furthermore, in a survey designed to better understand the expectations and use of digital editions, participants were asked, "what use would you make of the data published in a digital edition", to which the most frequently cited response was "teaching", and "text analysis" as a very close second [5]. Text analysis, like the edition, is changing, and as text analysis of the quantitative sort becomes increasingly prevalent, so too will the demand for its integration with digital editions as digital systems (which they are). In some instances, users are permitted to download the materials from editions and can subsequently conduct their own analyses if they have the specialist expertise or resources required to do so; but in most cases, it is either impossible or prohibitively cumbersome to use the data contained in digital editions for computer-assisted text analysis.

The future of digital editing should be one in which the publishing platforms are integrated with statistical methods already being used in the digital humanities. Such a future would move scholarly editing, and indeed, the wider digital humanities, beyond the two dimensions identified by Bode. Amy Earhart believes such a future is possible: "data sets and editions can coexist, but only if those from digital and textual editors can find bridges to those approaching digital humanities from other traditions and with other goals" [6]. Seeing the creation of a digital scholarly edition and the application of digital methods for content

analysis as part of a holistic approach to knowledge dissemination would not diminish the intimacy of the editing process, but rather, supplement it. The combination of AI alongside traditional curatorial methods provides editors and audiences with different perspectives, specifically, the type of quantitative evidence that, for better or worse, is valued in today's society as either a form of evidence or a point of entry into complex information. Embedding AI-driven cultural analytics in editions themselves democratises distant reading, as those wishing to apply such methods to the contents of an edition would be able to do so without the need to develop or access specialist expertise or software. And it brings reliability and credibility to datasets. One of the great challenges of distant reading is that methodologies are only as reliable as the data being tested, and in scholarly editions, we find ideal datasets which have been expertly, and more importantly, *transparently* (in that the profile of their curator is visible and human), compiled.

If the ambition of digital scholarly editions is to make digitised text more accessible and searchable, it seems that a PDF of a printed text, archived and well described in a suitable repository, would be sufficient. But if the ambition is to use the digital to transform scholarly editing to a more radical degree, then it would seem that the ways in which critical editions can be read is an obvious opportunity, particularly as scholars across the digital humanities and AI have already developed, adopted, and tested a range of methods for doing just that.

For editors who have a preference for textual editing traditions, their editions can still exist as print or as digitised editions.

Acknowledgements

This research was funded by UKRI-AHRC and the Irish Research Council under the UK-Ireland Collaboration in the Digital Humanities Research Grants (grant numbers AH/W001489/1 and IRC/W001489/1).

References

[1] Haentjens Dekker, Ronald, Dirk van Hulle, Gregor Middell, Vincent Neyt, and Joris van Zundert. "Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project." *Digital Scholarship in the Humanities* 30, no. 3 (2015): 452–70. https://doi.org/10.1093/llc/fqu007.

[2] For an interesting example of how online knowledge resources cannot be taken at face value, even when dealing with "facts", see: Willaert, Tom, and Guido Roumans. "Nitpicking Online Knowledge Representations of Governmental Leadership. The Case of Belgian Prime Ministers in Wikipedia and Wikidata." *LIBER Quarterly: The Journal of the Association of*

European Research Libraries 30, no. 1 (2020): 1–41. https://doi.org/10.18352/lq.10362.

[3] Underwood, Ted. “A Genealogy of Distant Reading.” *Digital Humanities Quarterly* 11, no. 2 (2017).

http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

[4] Bode, Katherine. 2019. “Computational Literary Studies: Participant Forum Responses, Day 2.” *In the Moment* (blog). 2019. https://critinq.wordpress.com/2019/04/02/computational-literary-studies-participant-forum-responses-day-2-3/.

[5] Franzini, Greta, Melissa Terras, and Simon Mahony. “Digital Editions of Text: Surveying User Requirements in the Digital Humanities.” *Journal on Computing and Cultural Heritage* 12, no. 1 (2019): 1:1-1:23. https://doi.org/10.1145/3230671.

[6] Earhart, Amy E. “The Digital Edition and the Digital Humanities.” *Textual Cultures* 7, no. 1 (2012): 18–28.

https://doi.org/10.2979/textcult.7.1.18.