Biz of Digital – AI Writing Tools and Data Sharing Mandates: Publishing tsunami or FAIR at last?

By **Alvin Hutchinson** (Head, Scholarly Communication, Smithsonian Libraries and Archives, MRC 154, P.O. Box 37012, Washington, DC 20013-7012; Phone: 202-633-1031) <HUTCHINSONA@si.edu>

Column Editor: **Michelle Flinchbaugh** (Digital Scholarship Services Librarian, Albin O. Kuhn Library & Gallery, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250; Phone: 410-455-3544) <flinchba@umbc.edu>

Scientists today are increasingly urged to share research data. This makes sense to just about anyone except perhaps the scientist on whose shoulders falls the task of preparing data for use beyond their immediate project(s). These data sharing mandates, together with the recent development of automated writing tools, present a variety of potential and largely unpredictable consequences to the scientific publishing world. As with many aspects of artificial intelligence, the implications for publishers, scientists, and research libraries may be significant.

Digital tools can now manage and manipulate information far beyond what could have been done just twenty years ago. In newsrooms for instance, computers can create editorial content without human intervention. The simplest example is the 2-3 sentence sports digest which includes the basic facts about a sporting event: which team won or lost; which player had the most points or scored the winning run, goal or basket; the current league standings, etc. All of this data (in the case of baseball, runs, hits, strikeouts, etc.) are structured and documented so that computer software can read it and create a short human-readable synopsis of each game.

"Despite potential drawbacks and limitations to AI-generated texts, the practice could easily compound the already unmanageable volume of scientific research being published."

Scientific research today continually generates more and more datasets from natural phenomena and laboratory experiments, much of it formatted in tables, columns and rows. Today scientists are encouraged if not mandated to share their research data with the aim of making it reusable by future investigators. Currently there are various levels of compliance, but for reuse to become as common as is hoped by the FAIR (Findable, Accessible, Interoperable, Reusable) standard, datasets would need to be carefully documented, normalized and otherwise structured.

Inevitably the software and platforms that are used to collect and manage research data — whether from sensors, survey instruments or manual capture — will become more sophisticated and will smooth the path to reusability, requiring less intervention by scientists. Some tools have come close. For example, WorkBench allows the use of discipline-specific ontologies to standardize and clean up data sets, potentially providing some degree of integration among those data that use the same set of terms or concepts. And like baseball box scores, as tabular scientific data becomes more structured, it will likely become easier to use for automatically generating human-readable summaries.

Even before OpenAI launched ChatGPT in late 2022, there were several software applications used in the research enterprise that included natural language output. For example, SciNote, an electronic lab notebook that has been available for years includes a manuscript writer module that generates versions of a draft of a scientific paper. While the product documentation emphasizes the need for scientists to review the draft and make corrections where needed, it appears to remove at least some of the hurdles of the paper-writing process. Assuming tools like WorkBench, ChatGPT and SciNote proliferate and improve over time, there is reason to believe that they will enable the automated creation of a draft manuscript and eventually, something closer to submission-ready copy.

Initial complaints about ChatGPT center on false and fabricated information created by the software, but that appears to be due to the program relying on unevaluated internet sources. When the software is limited to using structured and curated scientific data as input then presumably much more reliable output can be created. One would hope that any scientist or team using an AI tool to generate a manuscript would review the output carefully to verify accuracy, similar to self-driving vehicles today which despite increasing sophistication still seem to require a human behind the wheel.

Among those impacted by the widespread adoption of text-generating tools by scientists, research librarians will likely not be spared. If current research incentive and reward structures for scientists remain unchanged, labor-saving writing tools may ensure that the ongoing pursuit of increased publication output by individual scholars reaches unprecedented levels.[1] Despite potential drawbacks and limitations to AI-generated texts, the practice could easily compound the already unmanageable volume of scientific research being published. Writing scientific papers — or parts of them — will likely become easier to the point that together with already-strong incentives for scientists to publish and be cited, we will see a steady increase in the volume of papers.

Discussions about the ethics of using machine-generated scientific manuscripts are too broad to cover here. While several publishers have created policies that deal with artificially-created text (in some cases forbidding AI authorship) it is uncertain whether the generation of draft papers — even if subsequently revised and edited by scientist-authors — is a process that is

universally prohibited. Likewise, given academic misconduct practices today it is not clear to what extent these prohibitions will be followed or could be rigorously enforced.[2] But whether prohibited or not, we can expect both the sophistication and use of these tools by scholars to increase over time which would only apply upward pressure on the amount of publications produced today.

On the other hand, some have expressed optimism that these tools may signal the beginning of the end to the reckless pursuit of publication and citation in researcher evaluation.[3] If ease of generating papers (or parts of them) results in the body of literature being diluted (or viewed as such), it is possible that research evaluation exercises will begin to de-emphasize pure publication and citation counts as a proxy for research excellence and simultaneously emphasize data collection, curation, and sharing. Where it becomes easier to write more papers it becomes easier to generate more citations to one's publications, perhaps diminishing their value. If this came to fruition, instead of a bewildering growth of journals, papers, versions, and repositories, the potential change in research evaluation could lead scientists to spend more time preparing data and less time writing. In that event, sharing one's data sets, assuming they are cited properly and uniformly, could end up being a more desirable goal than simply churning out an increasing number of papers.

One reason for scientists' hesitation to share data is that they fear their work will inform another, perhaps rival scientist's publication(s). These fears may be well-founded, but if the published literature continues to be watered down as described above, and this is recognized by research evaluation committees, one hopes that data sharing would be elevated in importance. As Digital Science CEO, Daniel Hook once optimistically put it, "It is really only a matter of time before having a highly-cited dataset is as important in some fields as a paper in *Nature*, *Science*, or *Cell*."[4]

The magnification of the current overload of scientific publications could play out in one or more ways. A greater stream of literature well beyond current research library capacities to manage it may expose the shortcomings of institutional repositories which will likely not scale well, especially given scientists' known reluctance to deposit reprints. Science librarians will in that case need to further develop cooperative digital archiving solutions. Or perhaps responsibility for collection, discovery and storage of scientific papers will fall on a third party, whether nonprofit or commercial.

Recognition and reward of research data sharing in this scenario might take precedence over publishing more and more papers. Were that the case, it would be reasonable to see research librarians accelerate their shift from traditional acquisition and collection of published literature to supporting the standardization, description, discoverability and access to research data produced at their institution.

AI will undoubtedly affect libraries in many ways outside the possible increase in published content to be managed. Reference, discovery, collection development, information literacy and other areas of librarianship could be offloaded to one degree or another onto machines. But as publishing goes, so go libraries.

NOTE: This piece was written in May 2023. It is possible that AI advances in the few months between draft and publication will have influenced these assumptions.

Endnotes

1. Liebrenz, M., R. Schleifer, A. Buadze, D. Bhugra, and A. Smith. "Generating Scholarly Content with ChatGPT: Ethical Challenges for Medical Publishing." Lancet Digit Health 5, no. 3 (Mar 2023): e105-e06. https://doi.org/10.1016/S2589-7500(23)00019-5.
2. Staiman, Avi. 2023 "Academic Publishers Are Missing the Point on ChatGPT" The Scholarly Kitchen https://scholarlykitchen.sspnet.org/2023/03/31/guest-post-academic-publishers-are-missing-the-point-on-chatgpt/.
3. Lund, Brady D., Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Zhang Wang. "ChatGPT and a New Academic Reality: AI-Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing." Journal of the American Society for Information Science & Technology (March 2023). https://doi.org/https://doi.org/10.1002/asi.24750.
4. Pool, Rebecca, "Dare to Share?," Research Information, 2016, https://www.researchinformation.info/feature/dare-share.