Mobilising our skills and values for the data centric world of artificial intelligence

Andrew Cox
Information School, University of Sheffield, Sheffield, UK

Abstract
Because current conceptualisations of how to achieve Artificial Intelligence are data driven, so information professional skills applied to data become highly relevant. Translating our well established information skills to the context of data management and stewardship could be invaluable in such areas as data search, understanding data provenance, copyright issues, promoting data sharing and standards based description of data, data disposition or preservation, data ethics, and in promoting data literacy. As a profession we have a valuable and unique contribution to make through information skills applied to data, but we need to include data more in our vocabulary and thinking.

Key words: *artificial intelligence; data management; data curation; data stewardship; data governance.*

I suggest that librarians have a potential role in artificial intelligence (AI) because of the relevance of our existing skills and values to data and because data is the foundation of AI.

The current conceptualisation of Artificial Intelligence is based on training algorithms with data, in the case of large language models (LLMs) like GPT, very large amounts of data. It is not by chance that our era of AI follows the decade of big data. It is partly Google's and Microsoft's access to massive amounts of data that enable them to do clever things with AI.

Equally, when we understand the importance of data we better understand some of the problems of AI. For example, we do not know exactly what data was used to train ChatGPT which makes it harder to fully understand its operations. What we do know about the data sources used, we have an explanation why there is so much bias in its outputs (1). Because ChatGPT was trained on material from the Internet and sites like reddit, it reflects many of the biases and stereotypes propagated in those spaces. Another data problem is that the harvesting process for data to train AI was under a claim of fair use, but the legality of this is in question. A number of copyright holders are suing OpenAI over their alleged use of copyright material without permission. The new EU legislation also requires a clear statement of what training data was used to train AI services.

The hunt for data

Such concerns around data ownership and data quality are an increasingly important aspect of how AI is developing. CBInsights report that one of the key trends for generative AI in 2024 is that "We are running out of high-quality data to train LLMs" (2). A recent AvePoint report suggested that many organisations are keen to exploit AI, but are finding data quality a barrier (3). The search for high quality data also explains the news that OpenAI recently signed a content deal with the Financial Times (4). They have similar deals with Associated Press, Axel Springer, and Le Monde. Google has signed a deal with the Wall Street Journal. The data centric nature of AI opens up one avenue by which libraries might be involved in AI development: through the use of the high quality material in library collections as training data. Data in the context of AI can be structured, quantitative data, but could also be text (including publications), sounds, images, anything

Address for correspondence: Andrew Cox, Information School, University of Sheffield, The Wave, 2 Whitham Road, The University of Sheffield, Sheffield S10 2AH UK. E-mail: a.m.cox@sheffield.ac.uk

that has been made machine readable. The term collections as data has been coined to think about library collections as machine readable data, rather than primarily for humans to read. The Vancouver statement on collections as data, which lays out principles for ethical use of collections as data is highly relevant to AI (5). Meanwhile the availability of open data may be key to blocking the tendency of big Tech companies to control the direction of AI development. Advocacy for open science including open data is highly relevant as a result (6).

Data skills

And that leads us to another dimension of AI's data-centrism that opens up possibilities for librarians, including health librarians: the importance of data skills. I would argue that many of the key skills and values of the information professional are highly relevant to management and use of data in such contexts as machine learning and AI. These competencies could be very helpful in supporting data scientists within organisations like the NHS, as well as the growing number of researchers in all disciplines who want to use AI based research techniques.

Here are some of the relevant skills and values:

- *Data search expertise.* Many funders expect researchers to undertake a data search prior to commencing research. Indeed, a data search should be part of any literature review. Yet, searching for data sources remains hard in a fragmented data landscape. Librarians are good at search. We can support data scientists to uncover valuable data sources.
- *Understanding data provenance.* Using data for any purpose, including AI, is fraught with problems if there is not a good understanding of how and why it was produced and how far it is a valid form of data for a proposed analysis or other use. Again, this is where, as information professionals, we can play a role in informing the use of data.
- *Copyright knowledge.* Expertise in IPR is highly relevant to understanding how data can be used. Researchers often turn to the library to understand copyright better.
- *Belief in data sharing and standards-based description.* It is second nature for information professionals to promote the sharing of information and the use standards in describing information to ensure that it can be found. The FAIR principles encapsulate this perspective. But not every researcher thinks like this. Librarians offer a distinctive contribution to the data ecosystem in promoting open data and data sharing more generally.
- *Expertise on preservation/ disposition.* Retention or destruction of data is an important, e.g. to comply with GDPR. Again, these are areas where our profession has long had expertise.
- *Strong stance on data ethics.* Our core professional values and ethical principles are relevant to data and AI. Our values include emphasis on equal access to information, avoidance of bias and misinformation, protection of confidentiality and support to intellectual property rights. Such guiding principles are highly relevant to data stewardship and AI.
- *Desire to promote data and AI literacy.* Our commitments access to information imply promoting literacy. In the AI context this includes data literacy, as an essential component of AI literacy (7).

In short, all these aspects of data stewardship are a strong match to librarians' skills and values. As a profession we talk a lot about "knowledge and information", "the evidence" and "the literature". We perhaps do not use the language of "data" enough. We may need to translate some of our skills to operate in a data world. But it is clear that our information skills are highly relevant to data stewardship and so to AI.

There are many other ways AI will touch information work (8), such as: through using generative AI for professional tasks, e.g. summarisation; through chatbot services to users; and in roles supporting users to pick from the plethora of generative AI-based tools for tasks such as search (9). So, there is a lot to do in terms of adjusting our professional skills, thinking and language for an AI world. But this paper argues strongly that translating our information management skills and values to the data centric world of AI would be a really positive path for the profession.

Submitted on invitation.
Accepted on 7 June 2024.

REFERENCES

1. Webb M. Exploring the potential bias in ChatGPT. 2023. In: JISC involve blog [Internet]. Available from: https://nationalcentreforai.jiscinvolve.org/wp/2023/01/26/exploring-the-potential-for-bias-in-chatgpt/
2. CBInsights. Generative AI predictions for 2024 [Internet]. 2024. Available from: https://www.cbinsights.com/research/report/generative-ai-predictions-2024/
3. AvePoint. AI and information management report: the data problem that's stalling AI success [Internet]. 2024. Available from: https://www.avepoint.com/shifthappens/reports/artificial-intelligence-and-information-management-report-2024
4. Murgia M. The Financial Times and OpenAI strike content licensing deal. Financial Times. 2024 Apr 30.
5. Padilla T, Scates Kettler H, Varmer S and Shorish Y. Vancouver Statement on Collections as Data [Internet]. 2023. Available from: https://zenodo.org/records/8342171
6. Ziesche S. Open data for AI: What now? [Internet]. Paris: UNESCO; 2023. Available from: https://www.unesco.org/en/articles/open-data-ai-what-now
7. Long D, Magerko B. What is AI literacy? Competencies and design considerations. Proceedings of the 2020 CHI conference on human factors in computing systems. 2020 Apr:1-16. DOI: 10.1145/3313831.3376727
8. Cox A M, Mazumdar S. Defining artificial intelligence for librarians. J Librariansh Inf Sci [Internet]. 2022 Dec;56(2). DOI 10.1177/09610006221142029
9. Baytas C Ruediger D. Generative AI in Higher Education: the product landscape [Internet]. 2024 Mar. Available from: https://sr.ithaka.org/publications/generative-ai-in-higher-education/

This paper is published under a CC BY license