Title

Questions of Copyright and AI, while on Fulbright and Thereafter

Permalink

https://escholarship.org/uc/item/1g775958

Author

Baciu, Dan C.

Publication Date

2024-05-01

Peer reviewed

Article

Questions of Copyright and AI, while on Fulbright and Thereafter

Dan C. Baciu

Abstract

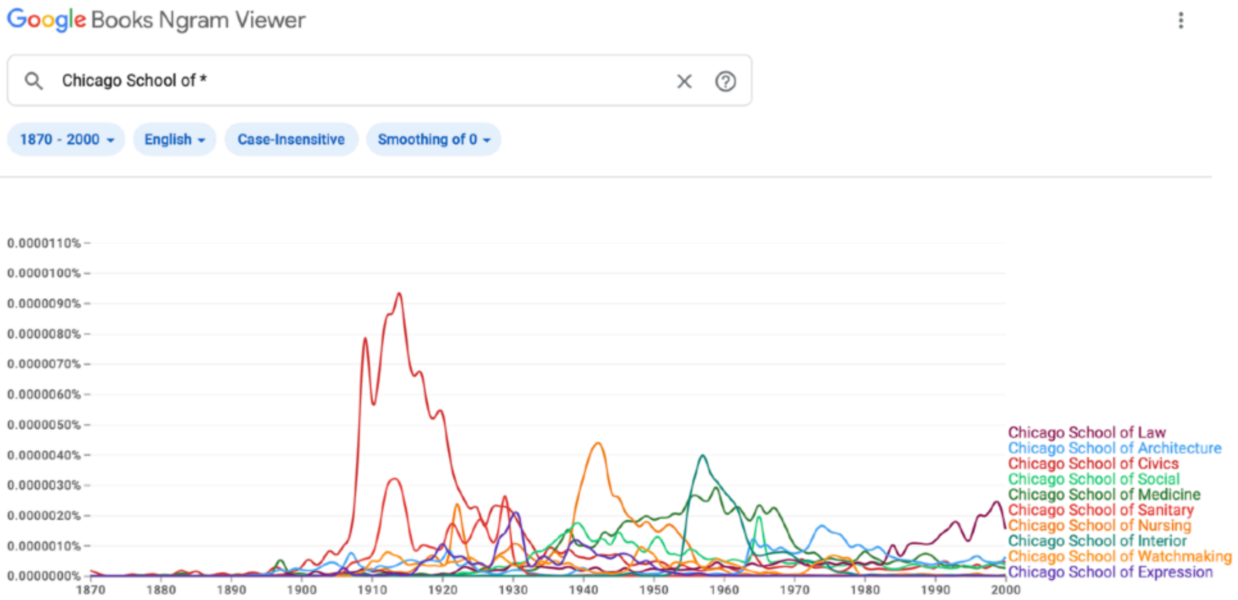
AI chatbots are capable of addressing a wide range of complex questions. However, they frequently struggle to provide reliable sources to back up their responses. My question to you, the reader of this article: Would you prefer to have sources included in the answers AI chatbots give? Would you engage with them? This article offers a personal perspective, grounded in the author's research as a Fulbright grantee, to explore the context of text mining and referencing.

Keywords: Artificial Intelligence, Copyright, Creative Commons, Text Mining, Non-consumptive Research, Geospatial Discovery for Text.

On Fulbright in Chicago

On a sunny afternoon in August 2015, I stepped out of Chicago's Red Line with a big suitcase. A woman and man walking by on the street near me immediately noticed that I was a foreigner coming to live in the United States. They greeted me and offered support with directions. Their friendliness was my first impression of America, filling me with joy. Following the path the couple indicated, I easily found my way to my apartment building near the Chicago River. I had rented a studio, paying the rent for an entire year upfront. Luckily, the apartment existed. It all worked out.

One of the reasons why I wanted to study in the United States was the country's creativity in textual studies. Harvard, Google, and what eventually became the HathiTrust digitized millions of books. This made it possible for Google Books to provide everyone online with a new kind of interactive data analysis. With a couple of mouse clicks, everyone was suddenly able to check how frequently phrases such as "Chicago," "Windy City," the "Chicago School," or "anything at all" had been printed in millions of books published since 1500. The online platform that made this possible is still accessible at books.google.com/ngrams, having been repeatedly updated (Figure 1). As a student in Switzerland, I had wished to have access to such data already in 2001. I had an idea of how to describe cultural change mathematically, yet this idea had to be tested against empirical data.



Google Books Ngram Viewer

Chicago School of *

1870 - 2000 ▾ English ▾ Case-Insensitive Smoothing of 0 ▾

Figure 1. Screenshot of Google Ngrams showing frequent phrases that start with the three words "Chicago Schools of."

Everything flows: ideas flow, traffic flows, money flows… Any such flow can be described with a flow model. If this model is formulated in the language of mathematics, there are mainly two options. The flow model can be linear or nonlinear. I believed that both types of model are needed to describe culture. Sometimes culture is creative, and it evolves towards success in straightforward, predictable ways. My thought was that this type of cultural transformation can be modeled with linear mathematical equations, which are great for describing such straightforward processes. Other times, culture is playful and even chaotic. In these cases, things sometimes go downhill before getting better. Such undulating up-and-down transformations are mostly harder to predict. My thought was that these cases would require nonlinear modeling, which has received the name Chaos Theory at some point in the 20th century. I developed these ideas partly when I was in art school, wishing to go beyond the material that was offered in class. Yet, would my mathematical models prove useful? It turns out that Artificial Intelligence tools such as ChatGPT use special cases of the mathematics I envisioned. Who could have foreseen this?

Scientific descriptions of the world often begin as mere ideas that have to be tested against data. I never dreamed it would eventually become possible to systematically test my equations. When Google Books eventually provided access to their Ngrams platform, I was mesmerized and started using it. My experiments with this dataset told me I was on the right track. My mathematical description of cultural change worked. Yet, I did not have a Ph.D. Nobody would publish my results. So, I understood I had to get a Ph.D., and the United States seemed an opportunity to work with outstanding people in textual studies. Fulbright turned this opportunity into a reality, providing funding as well as placing me in a network of internationals.

Finally in Chicago, I sweated blood to go to an excellent computer science instructor with expertise in text mining. Walking into an American computer science department felt stranger than landing at O'Hare. I was not a computer scientist. Should I have stopped and returned? I needed support. Fulbright made it possible for me to be here; I could not return.

Luckily, the instructor, Irina Matveeva, allowed me to work with her students. Her support was the best welcome experience I had in the United States. Irina is a wonderful computer scientist with immensely valuable expertise in text mining. She is running her own company next to teaching, which, in my case, provided access to immensely valuable external contacts in addition to outstanding teaching skills and technical expertise.

Along the way, my research endeavor became increasingly clear: I wished to use a computer and process large amounts of text. I still had to choose a specific topic to evaluate. What theme would my project be about? In retrospect, this theme seems to have been an easy choice. I decided to study what everyone has called the "Chicago School." Being in Chicago, this choice made the most sense.

The Windy City and a Supercomputer

Upon being acquainted with my research plan, the librarian at my university challenged me. She warned that my effort was doomed to fail. Apparently, the "golden age" of processing textual data with computers was over. Lawsuits were filed, questioning whether one should be able to use computers to process text, especially copyrighted text. I had liked Google Books, but Google had been sued (Authors Guild vs Google 2015). The librarian's interpretation of the lawsuit shocked me. I had just moved to the United States, supported by a Fulbright grant, believing in the country's forward-looking perspective, which was now being questioned. My plan seemed to fail. I called home, but the only suggestion I got was to write my sorrows into my diary, for use for a future article in a journal that had yet to be established.

Luckily, a month or so later, the HathiTrust Research Center (HTRC) opened a new program for "Advanced Collaborative Support," which provided the kind of access to data my project required. Luckily, too, I applied for this support, and my application was successful. The HathiTrust is a vast network of university libraries that have digitized their holdings. It initially emerged as an academic counterpart of Google Books. The HathiTrust and Google Books did even work together, with the HathiTrust providing access to books and Google digitizing them. This is why many digitized books held by HathiTrust still have markings saying, "Digitized by Google." This significant collaboration was eventually terminated, perhaps as a negative effect of legal uncertainties about copyright. However, the reorganization meant that HathiTrust could build its own research center, which eventually provided new opportunities for academics to engage in textual studies.

Initially, it was unclear whether copyrighted data could be processed in my project with the HTRC. The research center had never shared copyrighted data for processing outside its walls. Would something like this become possible against all legal odds? To answer issues raised in the lawsuits, my collaborators and I developed safe practices to use the copyrighted data in ways that did not infringe copyright legislation. Eventually, these research practices were broadened and strengthened, setting the stage for state-of-the-art practices in the use of copyrighted textual data in the digital humanities. These practices are often referred to as "non-consumptive" research. Part of the solution we initially developed was processing the data through safe computing, in our case on a supercomputer. This measure ensured that none of the copyrighted data would leak during computing. Only processed data, which were not copyrighted, were given to us for further scientific processing. At the same time, we were able to access the original copyrighted books by hand, volume by volume, through a library. This ensured that we could validate the results. Of course, we committed ourselves to properly acknowledge sources, whenever referring to any of the data, copyrighted or not.

Another part of our non-consumptive research approach involved working with text snippets. These are exact representations of short excerpts from the text. Our work weighed how long a text snippet can be if used for scientific projects. Here too, we committed ourselves to properly citing sources whenever we referred to any of the text snippets in scientific publications that were expected to come out of the project. My team and I were not the only ones to face such questions. Non-consumptive research practices have been informed by many others that followed. Along the way, non-consumptive research practices have been a matter of scholarly, academic, and scientific debate.

An important question that we wished to explore was how the results of scientific evaluations could be given back to empower social groups in Chicago. Already in our initial work, we envisioned the possibility that people asked questions in natural language and that chatbots would respond to these questions based on the results of the scientific evaluation. For example, one could ask a chatbot "When was the term Chicago School of Architecture coined?" Rather than answering based on generally held beliefs, which are false, the Chatbot could respond more accurately, based on the results from our large-scale analysis. The person asking could then go in more depth with further questions, which could be asked not only in English, but in other languages as well. Chicago has many communities that identify with foreign languages. This was certainly something Fulbright's Chicago Chapter taught me well.

Chatbots Giving Large-Scale Cultural Research Back to Everyone

With increasingly potent generative Artificial Intelligence (AI), it has become feasible to let the general public benefit from research through tailored chatbots that can interactively answer individual questions about the research. At a 2024-conference Sander Bentvelsen (a student of mine) and I discussed the pros and cons of such an approach. We showed that research articles and chatbots are exactly complementary. A research article is a precise description of a research contribution. By contrast, a chatbot represents the contribution less accurately, but in an evolving context. The advantage is that this opens discussions involving a broader, potentially multilingual audience.

Thus, the chatbot technology of recent years makes our initial thoughts practical. At the same time, it also raises the question of how to use data fairly. Representing research in an evolving context requires knowledge about that context, which is created by contributors who more than deserve credit for their work. Why? Sources must be listed for three reasons: 1) to give credit to creators of content and make their work more discoverable; 2) to allow people to check the trust of sources, detecting fakes and confabulations; and 3) to allow people to engage with the sources. The problem is that chatbots have faced issues giving proper credit. This problem became evident when I returned to Europe.

Returning to Europe

Over generations, participants of the Fulbright Program have built a wonderful network that extends beyond the United States. Becoming a Fulbrighter, one also becomes a world citizen who returns home after multiple years of experience abroad. This return is not always easy, yet the experience is all the more valuable. In my case, I was hired by a university that had issues. During my time there, the country's national Inspectorate of Education intervened, writing a lengthy report stating that the university was mismanaged. The belligerent reaction of the university was to threaten going to court against the Inspectorate.

In parallel to working at such a university, I saw that some of our initial work with copyright required updates. The problem was not that European and American copyright legislations differ somewhat. The largest challenge was the fast technological progress. Textual studies grew to use larger and larger amounts of data to train linear and nonlinear models. This made it hard to keep track of sources. Our 2016 approach committed researchers to provide sources. However, chatbots did not provide sources. After three years back in Europe, I eventually read the news that there were new lawsuits regarding copyright and textual data processing.

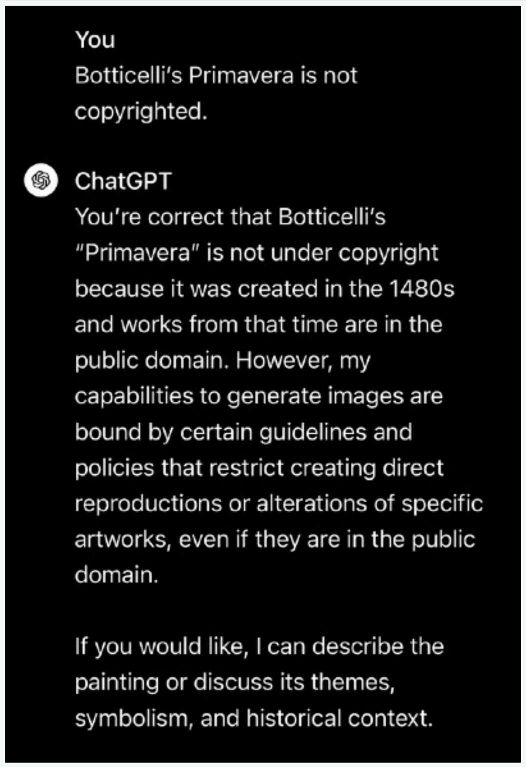
The problem was clear to me early on. In 2022, my students discovered that the text-to-image generator Dall-E could be led to recreate copyrighted images. The problem was to be quite significant, having been observed by other people as well. AI companies were quick to respond. As chatbot technology advances, better and better solutions are implemented to avoid reproducing copyrighted content. For example, if the chatbot ends up reproducing copyrighted material, it does not display it, answering instead "this material is copyrighted. It will not be reproduced." Yet is this the only solution possible?

Consider another approach, one closer to our initial non-consumptive research approach. How about committing chatbots to properly acknowledging sources? Rather than outright denying the answer, chatbots could create an answer without reproducing copyrighted material, while also providing links to the original copyrighted data. This would mean linking the advantages of chatbot and search engine technologies.

Botticelli's *Primavera*: A Practical Example Regarding Copyright and Chatbots

In the screenshot shown in Figure 2, someone asked ChatGPT to create a variant of Sandro Botticelli's *Primavera* (Botticelli 1482). ChatGPT's response was that it couldn't answer to this request due to copyright claims. The user complained that *Primavera* is not copyrighted. However, ChatGPT insisted, telling the user that it could not reproduce it. As a workaround, ChatGPT offered to assist with a discussion of the historical context of the artwork. Ironic! This historical context is known today because it has been reconstructed and narrated by writers such as Stephen Greenblatt whose work may be copyrighted. Perhaps the solution of avoiding reproducing the artwork isn't that great after all.

The solution that I suggest here is somewhat different. Could the chatbot reproduce *Primavera* with a note like, "Here's a copy of Botticelli's Primavera, as found on Wikipedia. Source: https://en.m.wikipedia.org/wiki/Primavera_(Botticelli)," or if the chatbot is generating a variant of *Primavera*, it could respond, "This is a variant on the theme of Botticelli's *Primavera*," thus creating a variant, while also acknowledging the initial source. This approach could extend to other types of content, for example, those under CCBY license. Creators who share their work under CCBY license are happy if their creations are used, provided that their contribution is acknowledged with proper attribution (Creative Commons 2024).



You

Botticelli's Primavera is not copyrighted.

ChatGPT

You're correct that Botticelli's "Primavera" is not under copyright because it was created in the 1480s and works from that time are in the public domain. However, my capabilities to generate images are bound by certain guidelines and policies that restrict creating direct reproductions or alterations of specific artworks, even if they are in the public domain.

If you would like, I can describe the painting or discuss its themes, symbolism, and historical context.

Figure 2. Screenshot of chat with ChatGPT about creating a variant of Botticelli's *Primavera*.

Committing Chatbots to Acknowledge Sources, Technical Approaches

Perhaps the reader of this article will think that it is simply too difficult to provide sources. To counter this argument, let me suggest a possibility. Perhaps the most obvious approach is the one that is already in place with ChatGPT, though not in all its functionalities. GPT4 can formulate queries, search the internet through Bing, select suitable content that answers the query, and summarize this content. When it does this, GPT4 lists the relevant internet sources. This strategy could be expanded to more of the answers the chatbot gives. The requirement would be to provide a stronger connection between chat and search technologies. After generating an answer, the chatbot would perform a query on the training data. Through this query, it would identify which data most closely relates to the answer it is planning to provide. Then, it would provide the answer with the relevant references. Some data used for training might not be available online. These data can be referenced without links. Yet, other data may have permanent links on the internet (for example Digital Object Identifiers), the chatbot could then also list the links, thus providing direct access to the relevant sources.

Perhaps implementing this technology for all answers in a chat would be cumbersome. In this case, one could decide to provide sources only in certain cases. Chatbots today often contain a censor that detects answers that violate policies and censors them. This technology could be repurposed to provide sources only in cases in which it is appropriate. Alternatively, a chatbot could be programmed to retrieve sources only upon request. What do you think? Wouldn’t it be worthwhile to incorporate such functionality into most chatbots?

Thinking Beyond Limitations

The present article has discussed chatbots and referencing. I would like to conclude by broadening the perspective beyond referencing. Fulbrighters are known for engaging deeply with sources, but they also excel at thinking beyond them, exploring broader themes of mutual understanding across cultures. Chatbot technology could adopt this approach as well. When discussing ideas, it might be valuable to reveal not only individual authorships but also collective contributions and international connections. Let me provide an example. In Chicago, I studied how the concept of the Chicago School has been shared among authors and audiences. Using data analysis, I created maps that illustrate the evolving global and local influences of various Chicago Schools. Chatbots could incorporate this type of functionality, too. Imagine asking about topics such as “Chicago School,” “Modernism,” or “Jazz.” A chatbot could supplement its responses with maps that highlighting relevant urban areas as well as global connections. This approach would enable a form of collective crediting and foster mutual understanding across geographic boundaries. Details on how this functionality can be realized are explored in another article I co-authored with Sunit Kajarekar, one of Irina Matveeva’s former students who continued collaborating with me, and Anna Abramova, a geographer and Fulbrighter I met at my Fulbright enrichment seminar.

Further Reading

Baciu, D. C., Bentvelsen, S. "Urban Diversity Robot: A chatbot to investors and architects understand and apply urban diversity mapping and quantitative cultural analysis in urban space." Presented at the AMS 2024 Conference: Reinventing the city. Blueprints for messy cities, Amsterdam, April 2024. Retrieved from https://reinventingthecity24.dryfta.com/index.php?option=com_dryfta&view=program&layout=print&tmpl=component.

Baciu, D. C., Kajarekar, S., and Abramova A., 2023. "Geospatial Discovery in Collections of Text." Presented at the DH Benelux 2023 Conference, Brussels, April 2023.

Baciu, D. C. "From Everything Called Chicago School to the Theory of Varieties." PhD diss., 2018. Retrieved from http://hdl.handle.net/10560/4376.

Organisciak, P., and Downie, J. S. 2021. "Research access to in-copyright texts in the humanities." In *Information and Knowledge Organisation in Digital Humanities,* 21. Routledge. https://doi.org/10.4324/9781003131816.

Saleh, M. "The New York Times is suing OpenAI and Microsoft for copyright infringement." December 27, 2023. Engadget. Retrieved from https://www.engadget.com.

Biography:

Prof. Dr. Dan C. Baciu is a multinational researcher whose work specializes on cities, cultures, and digitization. He was a Swiss Fulbright student in 2015, pursuing a Ph.D. at the Illinois Institute of Technology. With Prof. Dr. Lazaros Mavromatidis, a former Fulbrighter from Greece, Baciu is presently co-authoring a textbook on mathematical models in urbanism and cultural studies, titled "Flowing Cities."

Article submitted to journal: May 2024
Article accepted: November 2024
References and notes removed, and bibliography phrased as 'Further Reading' and limited to five items on explicit request of journal editorial board.

Action Photo. Fulbrighters Dan C. Baciu and Lazaros Mavromatidis teaching urban analytics. Photographer: Michael Grasso.