NASIG
TRANSFORMING THE INFORMATION COMMUNITY

Data Voids and Echo Chambers: The Transformative Journey of Search and Its Consequences

Francesca B. Tripodi and Paul D. Moeller

Abstract

Francesca Tripodi investigates the interplay between user-driven content generation and the manipulation of search in her vision session at the 2024 NASIG Conference. Tripodi shares the outcomes of research into the interconnected nature of data platforms and how the absence of information on Wikipedia may result in information vacuums known as "data voids" which have been manipulated for corporate or political purposes. Tripodi also discusses her research into how our worldviews impact the selection of keywords, which in turn determine the items retrieved in search results, and efforts to boost users' understanding of how search mechanisms work so that they are better prepared to navigate our information-rich society.

Keywords: search, data voids, deep stories, filter bubbles, ideological dialects, literacy

It is fantastic to be here, and I am thrilled that you all care about the outcomes of the research that I have come to talk about. Thanks so much for inviting me. My name is Francesca Tripodi, and I am a professor at the University of North Carolina (UNC) School of Information

https://doi.org/10.3998/nasig.7739

and Library Science. I am a sociologist of the internet, and for the last decade, I've been thinking about how the way people view the world shapes the theories they have and how that ends up impacting their selection of search terms. The talk I have prepared for today brings together aspects of some of the different research projects I've undertaken, and I am looking forward to a good discussion afterwards.

Now, what does it mean to search? I don't have to tell this audience, but back in the day, search was a thing that people would set aside time to do. You would carve a time out to go to the library to do your searching. Obviously, search has changed a lot over the years. Many people within my age bracket have even watched the transformation of search becoming "well, you Google it." To "Google it" has become an accepted norm, but this pattern is now transforming again. My young children, when arguing if something is true or fighting about knowledge gaps will say, "search it up, Mom, search it up." It is interesting because search really is changing, and it is not just something that we're typing, it is increasingly happening on mobile, and more and more becoming voice activated. And now we are seeing a shift to ChatGPT, and I will talk about that a little later today.

A lot of my questions center on the algorithms behind how we find what we want to find. The way that information is ordered is not particularly interesting when what you are searching for can be thought of as static searching. To give an example, I always find that I literally have no idea how many meters are in a mile. I never will, and that is okay as every time I travel internationally, I find myself relying on Google. I use Google because it is very effective for these kinds of everyday search tasks. "What is a good substitute for butter?" is another of the sorts of things that come up constantly, and what's nice about this kind of search is that it is a static search. These are things where the answer is never really going to change. Now of course I have just finished the *Three Body Problem*. I do not know if anybody else here has watched this series, but it makes me wonder if the laws of physics are about to erode, but I am still pretty sure that the number of meters in a mile is not going to change any time soon. Where things get more complicated, and where search becomes more complex is when we search for what sociologist Arlie Hochschild calls "deep stories." Deep stories are these stories that we hear so often in our childhood that we sort of know the end of the story without hearing the full story. These are stories that are told at the dinner table, around a campfire, in a church group, or at camp. These are stories that feel true even if they are not necessarily so. I like to think about how our deep stories have inside of them what I refer to as ideological dialects. Ideological dialects can be thought of as the way the language we use to make sense of our deep stories shifts depending on where we were raised, how we think, and what we want to do with that story. The perfect way to think about the idea of an ideological dialect and deep stories coming together is the color of the sky. I live in the United States, and I learned that the sky is blue. Lots of us grew up with this same story. So, the color of the sky is not just blue, it is also part of our deep stories. "Is the sky blue?" is the answer to questions needing veracity such as "Do you love me?" If you look at a box of crayons, you may find that the color of one crayon is sky. That the sky is blue is a truth that is very much attached to these deep stories. The truths expressed in these sorts of stories, however, may or may not be entirely true. Content analysis of old texts and translations of such works as the Bible show that they didn't say that the sky is blue. Instead, they spoke to the grayness or redness of the sky. It turns out that the idea of blue sky is highly connected to manufacturing because blue is not a color commonly seen in nature. While some plants are a beautiful indigo, for example, blue is relatively rare. Italian culture associated blue with purity, as to manufacture blue paint required the difficult crushing of a particular stone. So blue, the most expensive, pure, and beautiful paint was used for the Virgin Mary. It is also why at that time girl babies wore blue, and boys wore red or baby pink. We did not associate girls with pink until much later. So, the sky is blue is one of those deep stories that has been told so often that it has become an actual truth. And the internet is a great way to check that truth. If you search for why the sky is blue you will get very good results clearly confirming why the sky is blue. You get beautiful pictures from the National Oceanic and Atmospheric Administration (NOAA) and the National Aeronautics and Space Administration (NASA), and you'll see results from Wikipedia discussing shades of blue and why the sky is blue. NOAA, NASA—these are reliable sources. But what is fascinating about search is that it relies very heavily on relevance. This is a wonderful audience as I don't have to explain relevance and why it is significant. But if you search "why the sky is not blue," you get different answers to your question. First you get the song, "Why the Sky Is Not Blue" by Lemon Demon in the knowledge graph of the search results. I must say I've never heard this song. It might be profoundly offensive, and if it is, I'm sorry. I like to think that there is sort of this obscure band that I've somehow bumped up their place in Spotify and that I'm slowly building a whole following for Lemon Demon through use of this example in my talks. I don't know who they are, but they've done a great job at tagging their lyrics. But what you also see in the search results are a lot of studies telling you why the sky is not really blue, and that the sun is not really yellow. What you get in this result is a great example of the inclusion in results of what people also ask about in their searching. We refer to the inclusion of this sort of linking in search results as "semantic media." This is where they are pulling together information based on queries done by other people. The search afterwards in this case is effectively telling you to challenge this notion that the sky is blue and to ask when the sky is not blue. So, with this same deep story in mind, you can really confirm any reality you want with Google. The color of the sky question is one of the games I play with my kids. So, the sky is green apparently right before a tornado and the sky is purple during some beautiful sunsets. "Red sky at night, sailor's delight. Red sky in morning, sailor's warning" is another example of a deep story. The color of the sky is not necessarily uniform. It is not a static search. It is highly dependent on the way we see the world and on the kind of query we ask in Google or Bing or TikTok. What we put into the system is really going to drive what we get out of that system. A lot of the result depends on what Eli Pariser refers to as "filter bubbles." Now this idea of filter bubbles has varying levels of study around it. There have been a lot of studies that show tech companies are not able to keep us in the same types of bubbles as they were in the past, but what we do see are filter bubbles that are created by our own imaginations. So, instead of putting it exclusively on the power of the technology companies, I want to consider the role people play in creating the technology that surrounds them and how we form filter bubbles without even realizing it. I like to look at this through the lens of political thinking.

We know that there is a lot of political division in the United States with what we saw in the last election and what we will likely see again. It is as though we are in a permanent case of déjà vu in an increasingly polarized world. What I like to think about in my work is how these ideological world views and ideological dialects shape our version of truth and allow us to create an environment where people can search for alternative facts. What I think about in my research is how ideological dialects shape keywords and the way that ends up impacting the information that is subsequently returned to us.

I like to think about this in terms of immigration. I have created two searches that are not backed by scientific methods to use as examples and then will share a paper where this is done in a scientific setting. Here are two vastly different ways of looking at how citizens participate in the democratic process in the United States. Say I am just an average person in the state of Washington trying to figure out what the noncitizen voting rights are in Washington State. When you search "noncitizen voting rights Washington State," you have search results from the Washington Secretary of State and kingcounty.gov. I have done some fascinating studies on how people put an awful lot of weight in whether something is a .gov, a .org, or a .com. That is a great story, as anyone can buy some of these extensions, and they tend to lend an aura of credibility that is not necessarily down to them. So, we have a lot of information around what rights non-citizens have and what is the dialogue around them. But if I come at this in a different way using "illegal alien voter fraud Spokane WA," dominating the search results is an article almost two decades old from the

Spokesman Review talking about the problem of illegal aliens voting in elections. Those of you doing literacy instruction should emphasize that one should be careful with results that consist of predominantly older materials.

I wanted to test the theory that the way one sees the world impacts their choice of search terms, and a paper I wrote with one of my graduate students, "Abortion Near Me? The Implications of Semantic Media on Accessing Health Information," shares the results of our study.[1] We visited five libraries in different areas of North Carolina with different demographics using a loaner laptop that was provided by the University of North Carolina. The laptop wasn't connected to a profile as we were attempting to control for cookies and any sort of personalization. We effectively sat down with fifty-two people asking them a series of polarizing questions on topics that you really couldn't not have an opinion on. Using neutral prompts, we described a hypothetical and asked them what their position was on it. We then asked them to find information based on their perspective. One topic we tested focused on the issues around abortion. The prompt said a good friend of yours just found out they are unexpectedly pregnant and are considering terminating the pregnancy. Do you have a position on this situation? A little over half of the participants said they would support a friend's decision to terminate a pregnancy. When searching for information to support their friend in this decision, this group used almost the exact same search approach. Almost all of them either searched directly for "Planned Parenthood" or "abortion near me," or abortion plus the name of the state to find resources. The people who did not support their friend's decision had a lot of variation in their queries and approaches. Some sought out abortion alternatives such as adoption. Some searched for counseling services because they were going to direct their friend to them as they would not support a friend's decision to get an abortion. In general, their search efforts did not include a Planned Parenthood option. Effectively we were testing the hypothesis that the way people see the world impacts the kind of queries that they put into their search bar.

We know search engines are not librarians. They are produced and managed by corporate entities with particular corporate and political interests. We found when somebody searched for "Planned Parenthood near me," that the search results included the location of the nearest Planned Parenthood. But when the search terms were "abortion near me" or abortion together with the name of the state, the results were more interesting. We found that a lot of the top returns on Google Maps and Google Search were dominated by crisis pregnancy centers (or CPCs) which do not provide abortions and are designed to dissuade people from having an abortion. The fact that the crisis pregnancy centers were so dominant in search results led us to review some research. We found that investigative journalists were able to show that Google benefited by selling millions of dollars of coveted advertisement space to the centers. Although I am using Google in these examples, I do think it is important to recognize that these issues are not exclusive to Google. When I give a version of this talk to my students at UNC some well-meaning undergraduates say, "this is why I use DuckDuckGo." I respond saying protecting your privacy is important, but you still need to consider the corporate and political interests of the company, and you still need to worry about search terms.

So now coming back a little bit more to my other work, what does this mean then in the great scheme of search? A report by two members of Microsoft Research created a concept known as data voids. Their report shows that when little to nothing exists about a subject online it becomes easy to manipulate. One can effectively dominate search results by creating a ton of low-quality content on a subject and tagging that information with keywords. What I consider in my book, *The Propagandists' Playbook: How Conservative Elites Manipulate Search and Threaten Democracy*, is how this process of data void manipulation is used by conspiracy theorists and propagandists.[2] What I talk about in my book is this process of keyword curation and strategic signaling. Individuals or groups may create a lot of low-quality content which can exploit search engines. What's fascinating about this is that for it to work it relies on an extensive network outside of the internet. So, you can't just say something in a ton of searches to effectively manipulate results. These groups often work with podcasts, academic journals, radio and television personalities to build interest in words or phrases which otherwise would be meaningless or a nothing-of-interest subject. An example of this is "stop the steal." "Stop the steal" was a relatively new phrase that many people heard peaking in the period leading up into the 2020 election. What Google Trends shows is that the curation of "stop the steal" started right before the 2016 presidential election. Google Trends provides aggregate data of how people search over time, and it shows a small blip in the use of this search term all the way back in 2016. You see another blip right around the 2018 election cycle, and then you see it peak right before the 2020 election. There is a brief decline in the use of the term after the election, but it spikes again right before the events of January 6, 2021. What I find interesting is the deep story around "stop the steal" because "stop the steal" by itself is not particularly compelling. If you go back to 2016, using Twitter data saved by the WayBack Machine, you can see that the language around "stop the steal" is clearly connected to Make America Great Again (MAGA). In a precursor to the false accusations made against Dominion voting machines, you see issues raised about the security of absentee ballots and voting by mail. You see references to George Soros who is a subject in many antisemitic memes. And then you have references to stopthesteal@gmail.com. I have written to them, and they've never responded. So, the deep story around voter fraud in the United States goes something like this: others, whether that be women, people of color, members of the LGBTQ community, non-native-born persons, non-citizens, and non-Christians are trying to steal our elections. And if we don't push back, what will happen to the America that we love? What's interesting about this deep story is that it is a really tired story that has been recycled as part of misinformation campaigns since the 1800s. W. E. B. Du Bois, in his book *Black Reconstruction in America, 1860–1880*, notes that when reconstruction happened and Black men had the opportunity to vote for the first time in their lives they elected a record number of Black men to represent them in office. What followed were allegations of fraud and suggestions that some sort of dishonesty had led to this outcome. What I have been able to show in my research is that by creating this drum beat around the "stop the steal" voter fraud story, they were effectively able to trick search. In the days leading up to January 6th, "stop the" would autocomplete with "steal rally," and the search results would direct you to the rally closest to you so that you could engage with other like-minded scholars researching the fraud of the 2020 election. This sort of doing your own research I refer to as the IKEA effect of misinformation. Business scholars have shown that when people put together their own furniture, they value it more, even if it is the same quality as something purchased premade. Influential media personalities and politicians tell their audience not to trust them or the media because only you the listener have the power to get answers. I refer to this as participatory disinformation. In a recent paper I wrote with two other scholars, " 'Do your own research': Affordance Activation and Disinformation Spread," we showed how doing your own research can go wrong when the platform afforded has been manipulated.[3]

So, I am going to take a small turn and then we will come back around as I am sure you are all wondering about ChatGPT. What I have been studying a lot about with ChatGPT is how data voids manifest themselves in these types of systems. While I have been researching search for over a decade, I have recently been investigating artificial intelligence (AI) enabled search systems. So, what about ChatGPT? Research shows that there is an intimate connection between Wikipedia and search results. Many of the knowledge graphs that you see in traditional search engines are pulled from Wikipedia, and Wikidata was one of the largest sources of training materials for many of the large language model systems that we are using now. There is a lot of good that has come from Wikipedia, but there are some issues as well. In my article, "Ms. Categorized: Gender Notability and Inequality on Wikipedia," I looked at how women editors and pages about women are disproportionately challenged in Wikipedia after an article's creation, so we have a gender gap in representation that manifests itself in both editorial activities as well as content.[4] Upwards of 80 percent of editors are men, and likewise most pages are filled with men's interests, and biographies about significant women are severely underrepresented. So, what I talk about in this paper is how this problem of the gender gap is even more layered. I was interested in what happens in Wikipedia edit-a-thons, which are special meetups designed to improve the encyclopedia. What I found at these events was that pages about women were being declined as non-notable or tagged for deletion during the edit-a-thon. I thought this was weird, and all these women editors I interacted with at the edit-a-thons said that this happens. So, I partnered with an amazing computer scientist, and we scraped all articles marked for deletion over a four-year period, and I analyzed them trying to determine whether the same proportion of men's and women's articles had been nominated for deletion but should have been included in Wikipedia. I have scientifically referred to this act as an oopsie. My null hypothesis is that the proportion of oopsies would be the same if no inherent bias exists around gender and that the portion of oopsies for men's and women's content would remain constant. What I found was that the number of women's content that was marked for deletion was disproportionately high except for during one short period. This period took place after Donna Strickland won the Nobel Prize in Physics and everyone went to Wikipedia for information about her and there was nothing there. There had been a Wikipedia page for Donna Strickland, but she had been nominated for being a non-notable subject and the page had been deleted. After this event there was a period of more balance in articles nominated for deletion. This has since passed, and the disproportionate removal of women's articles has resumed. So, what does this have to do with ChatGPT? If you search ChatGPT for notable scientists or notable artists, a disproportionate number of them will be men. So, thinking about how this search base is interconnected with AI-enabled search tools is really important. I'm also thinking about how data voids in particular impact ChatGPT. With that, I will open it up to questions.

Discussion

Several questions were raised by members of the audience. One attendee asked if there were issues around the representation of transgender people in Wikipedia. Tripodi noted that one of her graduate students had done a study and found issues with vandalism in which transgender scientists were dead named or had their preferred pronouns repeatedly changed. For individuals who obtained notability before transitioning, Wikipedia maintains a practice of keeping their dead names, and her graduate student was also looking at problems associated with this policy.

When asked why articles about women were getting deleted from Wikipedia, Tripodi responded that there is room for deletions in some respects. She supports deletions for vandalism. Vandalism happens on Wikipedia, so it is important to make sure that you do not have harmful content posted because it is so powerful in terms of the access to that information. Wikipedia is often used by people trying to bump up their ranking. Someone opening a donut shop could create an article touting the shop's amazing donuts. To try to combat this sort of manipulation, the Wikipedia Community has policies in place that say something is not considered notable unless it has been covered by an independent third-party source. Why are articles by and about women more likely to be challenged? Tripodi thinks it is sexism. Tripodi received hate mail after National Public Radio did a piece on her "Ms. Categorized" paper that showed gender inequalities are real. Current studies replicate ones done all the way back in the 1950s by sociologists that show resumes getting passed over when a woman's name or anything other than a white dominant presenting name is featured. To be fair, Tripodi noted that she cannot use her data to test the motivations behind the nominating of articles for deletions, but she thinks a lot of it is due to implicit bias.

When asked if she thought that Wikipedia and other tools were just dumb interfaces that reflect and reinforce what is happening in our society and are not in themselves the problem, Tripodi was in full agreement.

She noted that lot of what we see in search is society reflected back to us. When research shows that a search query for "Black girls" returns pornography, this is indicative of the problematic way in which we as a society see young Black girls. She also thinks it is more than that. Part of what she is trying to show in her research is when you are searching "illegal aliens and voter fraud," that is reflecting your own world view. She thinks the issues we are seeing around search are both a reflection of society at large and a reflection of one's own world views.

When asked what she would like from the librarians in the room, Tripodi noted that better literacy tools could give people a sense of how search engines work or how technology works so that they can be empowered to make sense of it all. She referred to a recent incident with AI-enhanced search in which people jokingly asked how many rocks a person should eat in a day, and the chat box responded something like according to geologists you should consume one small rock a day. This is effectively a silly but great example of a data void. She would like people to understand that the starting point of search is really important, especially during an election cycle. She recognizes that funding and job security for librarians can be precarious depending on where they're situated so they do not want to be seen as inserting politics into an information literacy session. But if they say "just look at your starting point and understand how that will impact search results," that should be apolitical. This is where the sky is blue example can be powerful in helping people understand what relevance is in a meaningful way. When she does a talk for kids, she likes to search "dogs are the best" and then search "cats are the best" because Google will return valid points for both searches. She wants people to validate information and utilize critical thinking. It is paramount for patrons to understand search, and librarians offering search literacy workshops can be very helpful.

An audience member noted that in the early days of search there was a lot of competition, and the information search and retrieval landscape had not been monetized yet. This has changed over the last twenty-five years, and he thought we were seeing more misinformation as a result. Tripodi agreed that the monetization of search is problematic. She raised the issue of the recent accidental release of Google's search algorithms. Money is still one of the primary drivers of search regardless of what they want us to think. She has had a lot of back and forth with Google where they've said what she has written is not accurate, and now, she is seeing the algorithms and thinking what she wrote was pretty accurate. Google has been dominant, but we are also seeing many people turning to Siri, Alexa, YouTube, and TikTok for their questions. We are seeing a war being waged on libraries, and Tripodi thinks people recognize that the library is the last information frontier. The reason libraries are being targeted is because in a saturated search environment fueled by corporate interests, we still have the library; and that poses a thorn for some. Tripodi believes the monetization of the dominant model for search really stresses the importance of the library.

In a follow-up question to what libraries can do in the current scenario, Tripodi was asked how much of the problem is learning how to search and how much of it is the data void. Tripodi thinks part of it is learning how to search. She has done a lot of work with Wikipedia on gender inequality but has also done a lot of amazing work with Wikimedia. Through a collaborative lab (Search Prompt Integrity and Learning Lab) she and other scholars are investigating how tags for vandalism and contested phrases could provide a lens into data voids as they are forming. It is not the data voids that we care about. It is the garbage they are filled with that is the issue. There will always be some data voids, but we would like to develop a way to identify keywords as they are forming and to identify the curation process as it happens instead of afterwards. Her goal is to create a dashboard for librarians, journalists, and technologists. Data voids are always going to be around. They are cyclical, and journalists, once they are aware that a data void is being manipulated, are getting good at having them filled with high quality content. It is those pivotal points when people are searching for a topic newly of interest that are critical. Developing a way to circumvent the manipulation and curation of keywords is an area of focus for her.

Contributor Notes

Dr. Francesca B. Tripodi is an Associate Professor at the School of Information and Library Science and a Principal Investigator at the Center for Information Technology and Public Life at the University of North Carolina at Chapel Hill.

Paul D. Moeller is an Associate Professor and Director of the Digitization, Description, and Discovery Services Team at the University of Colorado Boulder Libraries.

Notes

[1] Francesca Bolla Tripodi and Aashka Dave, "Abortion Near Me? The Implications of Semantic Media on Accessing Health Information," *Social Media + Society* 9, no. 3 (2023), https://doi.org/10.1177/20563051231195548.

[2] Francesca Bolla Tripodi, *Propagandists' Playbook: How Conservative Elites Manipulate Search and Threaten Democracy* (New Haven: Yale University Press, 2022), https://doi.org/10.2307/j.ctv2pwtnwn.

[3] Francesca Bolla Tripodi, Lauren C. Garcia, and Alice E. Marwick, " 'Do Your Own Research': Affordance Activation and Disinformation Spread," *Information, Communication & Society* 27, no. 6 (2023): 1212–1228, https://doi.org/10.1080/1369118X.2023.2245869.

[4] Francesca Tripodi, "Ms. Categorized: Gender, Notability, and Inequality on Wikipedia," *New Media & Society* 25, no. 7 (2023): 1687–1707, https://doi.org/10.1177/14614448211023772.