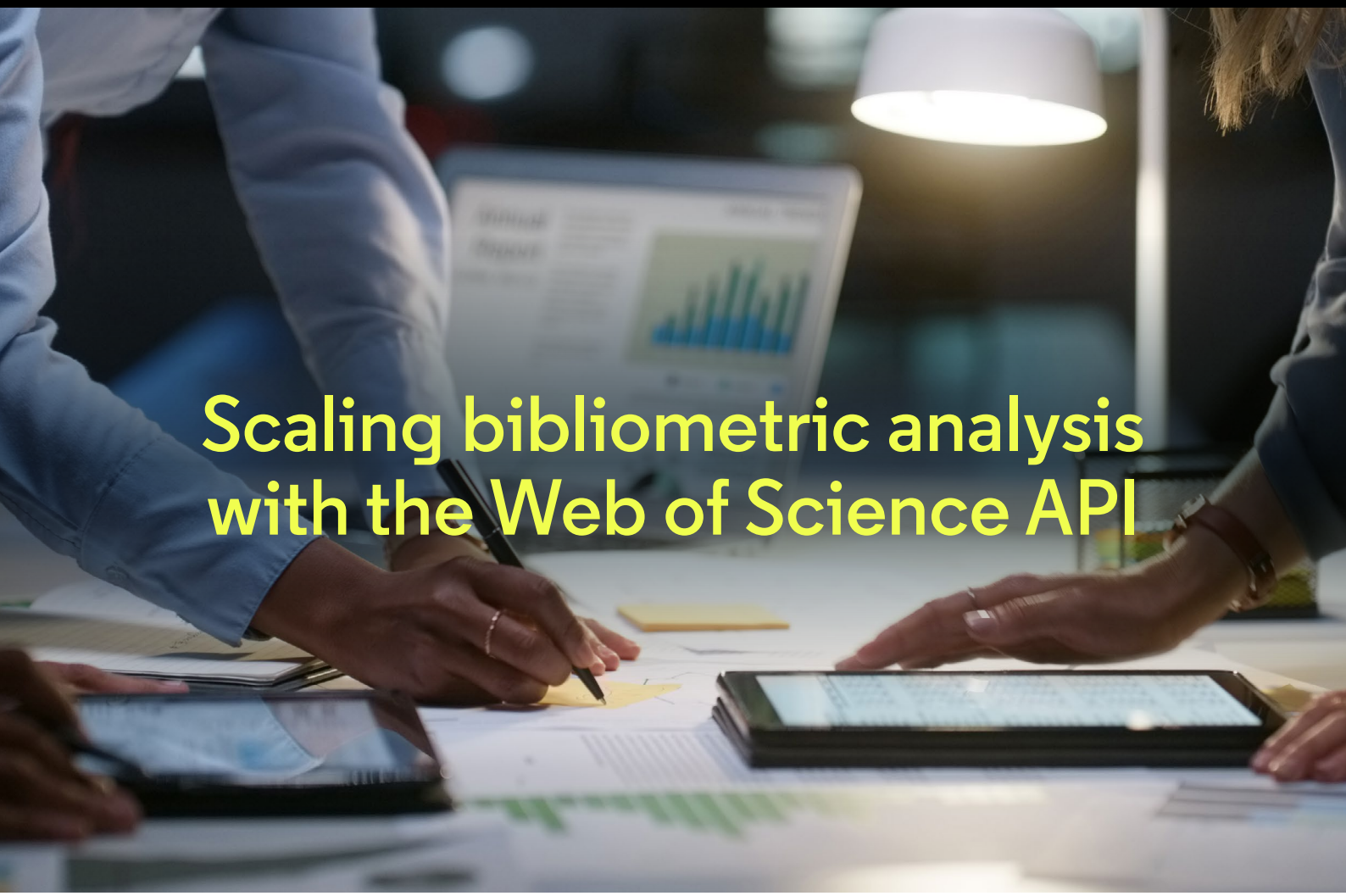

Scaling bibliometric analysis with the Web of Science API

Using the Web of Science API, researchers conducted a large-scale analysis of authorship and citation patterns within a research field.

Customer

Bansal Lab researchers at Georgetown University

Solution

Web of Science API

Challenge

To investigate authorship and citation practices within the interdisciplinary field of infectious disease dynamics (IDD), the research team needed a scalable, high-quality data source that could support complex bibliometric analysis across a diverse and rapidly growing body of literature.

Outcome

Using the Web of Science API, the team conducted a large-scale analysis of authorship and citation patterns within the IDD field. Their work revealed disparities in citation frequency across different demographics, offering evidence to inform future efforts related to representation in the field.

Clarivate

Georgetown University, located in Washington, D.C., is a leading research institution with $376.6 million in research and development expenditures in 2024. Over 1,000 researchers and scholars in the natural and biomedical sciences, social sciences and humanities collaborate across the university's campuses and around the world to conduct research for the common good.

Dr. Shweta Bansal is a professor of biology and principal investigator for Bansal Lab, a disease ecology and network epidemiology research group at Georgetown University. She supervised and guided the research conducted by Alexes Merritt, research associate at Bansal Lab, and Juliana Taube, Ph.D. candidate in biology.

Their research focuses on the social dynamics of science, and they were inspired by a neuroscience study that explored differences in representation and impact across demographic groups. The IDD field gained visibility during the pandemic, and Taube and Merritt were curious whether patterns found in other quantitative fields manifest similarly in the IDD community. They aimed to establish a data-driven baseline to assess the current state of the field and inform future recommendations.

Mapping a diverse and evolving field

The IDD field is highly interdisciplinary, spanning biology, epidemiology, network science, economics and more. Unlike fields with centralized publication venues, IDD research is dispersed across journals and disciplines, making it difficult to define the field using traditional journal-based methods.

To address this, the team adopted a novel approach: identifying the field through citation relationships. By analyzing who cites foundational IDD papers, they could map the community of researchers contributing to the field. This required access to a large, high-quality bibliometric dataset.

Leveraging the Web of Science API

Locating a reliable dataset is critical for 'science of science' studies such as this. Taube and Merritt knew that they would need to work with massive amounts of data for the project and found that Web of Science offered well-documented API access to a high volume of articles spanning disciplines and countries. When discussing why Web of Science was selected as a source, Taube explained, "It's helpful to use a database that is well respected and recognized." Relying on a reputable, validated data source lends credibility to a study.

The team started by using the Web of Science Core Collection to identify influential articles in infectious disease dynamics. They then constructed a dataset of citing articles from 2000–2019, which became the authorship dataset. From there, they created additional citation-based datasets. They applied machine learning tools to infer author gender and race/ethnicity, analyzed citation patterns and assessed disparities across geographic and demographic dimensions.

To conduct a demographic analysis, the team needed robust metadata on author names and affiliations. Web of Science Core Collection indexers capture all author names and their affiliations and unify affiliation variants to standardized organizations. "I don't think there was another way without Web of Science. I heavily relied on Web of Science for the country data. You can't easily find the same level of data elsewhere," said Merritt.

Merritt also shared that the thoroughly indexed, well-structured data from Clarivate ultimately saved her countless hours. "Web of Science data played very well with other tools because the names are very clean, and that's the input that most of the programs use. The process of cleaning data to feed it into some of the more finicky processing tools was much easier for us."

The team used the web interface for some parts of the data collection and the API for others. The platform's efficient batch download options were helpful for getting started quickly. However, as the project evolved in scope over time, the API offered flexibility and simplified the process of pulling additional data points for defined publication sets down the line.

Clarivate experts helped Taube and Merritt customize API queries, extract citation networks and adapt their data pipeline as the project progressed. "Our customer success consultant was phenomenal," Taube shared. "I worked with him on tailoring code to our needs, and we were able to get multiple generations of citations looping back on each other."

> “I don’t think there was another way without Web of Science. I heavily relied on Web of Science for the country data. You can’t easily find the same level of data elsewhere.”

Alexes Merritt,
Research associate at Bansal Lab

Evidence of disparities in access and recognition

The team’s analysis revealed several key insights into the dynamics of IDD research. IDD is a rapidly growing field, with significant expansion even before the COVID-19 pandemic. Authorship is dominated by men and white scholars, particularly in senior author positions. Women and scholars of color are undercited, and the undercitation of women authors intensifies when they occupy senior roles, as indicated by the last author position.

Impact and future directions

In their paper, the team introduced a set of tools and suggestions aimed at supporting authors, journals and institutions interested in addressing this dynamic. They developed a tool that allows authors to assess the diversity of their bibliographies relative to the broader field and offered optional strategies, such as removing citation limits, that could help expand representation if desired.

Looking ahead, Taube and Merritt hope their work inspires similar analyses in other scientific fields. They also see value in revisiting the field to assess progress and potentially exploring how the influx of COVID-era publications has reshaped the IDD landscape.

BXD1748534816 / 04

About Clarivate

Clarivate is a leading global provider of transformative intelligence. We offer enriched data, insights & analytics, workflow solutions and expert services in the areas of Academia & Government, Intellectual Property and Life Sciences & Healthcare. For more information, please visit **clarivate.com**.

Discover solutions at:

clarivate.com