Web of Science Group

A Clarivate Analytics company

Case study | Caltech

Using the Web of Science API to meet funding application requirements

"The increasingly interdisciplinary nature of collaborations, collaborators and science in general requires a reliable, comprehensive data source to ensure breadth and depth of coverage across the sciences." - Tom Morrell, Librarian, Caltech

Applying for grants and other funding support, and mustering the necessary data for documentation, can be a complicated and time-consuming process, fraught with exacting requirements and protocols. This phenomenon, familiar to all those involved in the research enterprise, has prompted some researchers and librarians to seek new approaches and efficiencies to make the process more streamlined and effective.

Notable success in easing the workload has been achieved by a group of librarians at the California Institute of Technology (Caltech), in the course of assisting faculty with funding applications. This work involved confronting large data sets, including searches on the *Web of Science*. Fortunately, library resources at Caltech were expanding to meet this demand, increasing the capacity to handle and analyze large files. One step in this evolution was the addition of a programmatic utility designed to speed up search queries and improve the formatting of the resultant data. This added tool was powered by a new API from the *Web of Science Group*.

Collaboration Data

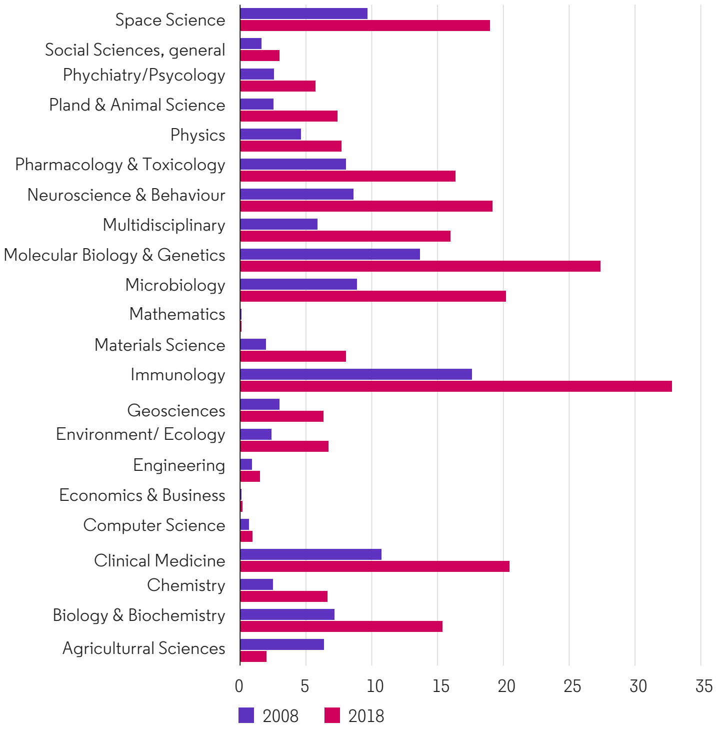
The Caltech librarians were fielding multiple queries from principal investigators (PIs) in physical-sciences disciplines, requesting help with funding applications to the US National Science Foundation (NSF).[1] The NSF stipulations include detailed information on collaborators – or, to use the official term, Collaborators and Other Affiliations (COA) data.

Specifically, in processing funding applications, the NSF uses COA information during the merit-review process, as the agency notes on its website,[2] "to help manage reviewer selection." In practical terms, this means avoiding possible reviewer bias by preemptively excluding anyone with whom the PI may have a business or academic relationship – including coauthorship on a research paper, service on an editorial board, or other official interaction.

In theory, this operation seems straightforward. At Caltech, however, the librarians were assisting PIs in astrophysics, high-energy physics, and other fields in which the extent and extremity of multiauthor papers have increased markedly in recent years.

In 2012, Web of Science data underpinned a ScienceWatch report[4] on the proliferation of multiauthor papers. Four years earlier, as the story noted, papers from the multinational team at the Large Hadron Collider (LHC) at CERN had broken the record for most-authored paper, with more than 3,000 listed collaborators. Subsequently, the hunt for the Higgs boson, nearing its successful conclusion just as the ScienceWatch story appeared, continued the trend of "hyper-authorship," as did other collaborations.

Figure 1: Percentage of Journal Articles with 10+ Authors by ESI Category



Data includes journal articles from Web of Science Science Citation Index Expanded (SCI-E) and Social Science Citation Index (SSCI). Data will not equal 100% due to rounding. Date of extraction: 5 April 2019."

[1] https://www.nsf.gov/pubs/2019/nsf19003/nsf19003.pdf

[2] https://www.nsf.gov/bfa/dias/policy/coa.jsp

[3] https://clarivate.com/?p=35127&preview=1&_ppp=0f3e27ff4c

[4] ScienceWatch, July 2012: http://archive.sciencewatch.com/newsletter/2012/201207/multiauthor_papers/

Nature magazine reported in 2018[5] (quoting data from Web of Science) on the continuing rise of papers with more than 1,000 authors. By then, the 3,000-author mark had been handily surpassed by a 2015 Physical Review Letters[6] report listing more than 5,100 authors. (The paper, not terribly surprisingly, described follow-up work at the LHC on the Higgs boson.) Overall, as Nature reported, the fields of nuclear and particle physics accounted for the majority of 1,000-plus-author papers.

Meanwhile, as Web of Science figures attest, the percentage of papers with 10 or more authors has steadily increased in all main fields over the last decade. Overall, in journal articles indexed in the subject fields covered in Essential Science Indicators, the percent of share papers with more than 10 authors has more than doubled in 14 of 22 subject areas (see Figure 1).

454

publications and nearly 11,000 unique coauthors

Complex Reports

At Caltech, the surging trend of multiauthor papers was clearly evident as library staff addressed the COA component of faculty NSF applications.

The problem:
Sometimes faculty requests involve a dozen or more PIs, each of whom is responsible for considerable research output, routinely entailing hundreds of publications and thousands of coauthors (in one instance, 454 publications and nearly 11,000 unique coauthors). The multistep data collection consumed a huge amount of time, as did the processing and vetting of the results – all while the librarians had their hands full with other requests and activities.

Another complication:
Library staff were starting to field additional faculty requests that went beyond the astrophysics specialization of their databases. Although Caltech's own repository[7] and the Harvard Astrophysics Data System (ADS) in astrophysics allowed for programmatic access to the required data,faculty from other departments were consulting librarians for assistance with their grants. The problem called for a multidisciplinary resource.

[5] www.natureindex.com/news-blog/paper-authorship-goes-hyper

[6] https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.114.191803

[7] https://authors.library.caltech.edu/

"A lot of libraries are starting to build out teams of data specialists – people who can interact with data programmatically for the purposes of reporting or systems integration..."

Joy Painter, Librarian, Caltech

Web of Science and a REST API

Caltech Library was able to utilize the Web of Science as a broader source of data, as it has coverage of more than 250 subject fields. Given the scale of the data, however, even using the web application on its own to support data collection and curation proved problematic. This capacity received a boost in mid-2018, when the Caltech staff began to use a new version of the Web of Science API as part of a broader collaboration between the two organizations. This new API version evolves the existing framework for programmatic Web of Science data interrogation, making it easier to leverage Web of Science data for use in internal projects and systems. By reducing the technical complexity of the API (REST vs. SOAP) and expanding the supported data formats (XML/JSON), the utility's usability and flexibility surmounts the enormous scale of the information involved, automating much of the multistep process for pulling the data, and giving the library staff more time to concentrate on other aspects of assisting with grant applications and the faculty they serve.

The experience of Caltech and the Web of Science API can be instructive for any library that wants to build its data-processing capabilities and handle large-scale, cross-disciplinary data requests in order to serve the faculty and otherwise fulfill its mission.

Joy Painter, the Physics, Math and Astronomy Librarian at Caltech and a key staff member dealing with NSF applications, had this to say about the addition of the Web of Science and new API to the library's arsenal:

"A lot of libraries are starting to build out teams of data specialists – people who can interact with data programmatically for the purposes of reporting or systems integration. So the question is, as the library moves deeper into the digital-world, how can libraries position themselves to leverage both domain expertise and expertise with working with data to better serve their faculties and the universities as a whole? With this work, we think it has shown that it is not only possible but paramount to bring together roles that had operated independently of each other to effectively address real problems that exist in the research environment."

Caltech has made their 'collaborator_reports' project code open source, available on Github: https://github.com/caltechlibrary/collaborator_reports

Contact our experts today:

+1 215 386 0100 (U.S.)
+44 (0) 20 7433 4000 (Europe)

webofsciencegroup.com