Operationalizing Bibliometrics as a Service in a Research Library

Susan Makar, M.L.I.S.
Research Librarian
National Institute of Standards and Technology

Amy Trost, M.L.I.S.[1]
Research Librarian
National Institute of Standards and Technology

Abstract

In 2017, librarians in the Information Services Office (ISO) of the National Institute of Standards and Technology (NIST) expanded ISO's bibliometric analysis services to include topical network analysis, bibliographic text mining, and deeper collaborations with stakeholders. The new services help NIST divisions and offices systematically assess research portfolios, document internal and external collaborations, and identify strategic areas for future research. While robust quantitative analysis has been the key to conducting bibliometric analyses and creating visualizations, collaborative relationships with customers are just as critical to understanding impacts and delivering the right answers to the right questions. This paper describes several use cases highlighting the new services, including a network visualization and cluster analysis to identify emerging research topics; an exploration of keywords in a single research area over the course of a decade; and a structured evaluation of collaborations within and outside of our institution.

Introduction

This paper shares the research and methodologies behind ISO's expanded portfolio of bibliometric tools and demonstrates a range of services that can be adapted for use in both academic and special libraries. Bibliometric analysis services can support research objectives from strategic planning to impact assessment, enhancing ongoing efforts to build collaborations with stakeholders.

NIST is a non-regulatory federal agency within the U.S. Department of Commerce (DOC). NIST's mission is to promote U.S. innovation and industrial competitiveness by

[1] Current affiliation is with the Universities at Shady Grove/University of Maryland Libraries

advancing measurement science, standards, and technology in ways that enhance economic security and improve the quality of life. ISO is responsible for creating, maintaining, organizing, and disseminating information to support the research and programmatic needs required to fulfill the scientific and technical mission of NIST.

ISO's path to operationalizing bibliometrics as a service involved responses to a series of customer requests in 2017, often related to strategic planning, that required quantitative analysis beyond traditional citation analysis and research impact assessment. ISO staff already had some experience with co-author network analysis, but had done little with topic or text analysis. The following projects helped ISO expand its bibliometric analysis services to include topic and text analysis, and more customized services.

- ISO's first collaborations involved the identification of emerging research areas relevant to NIST. Program staff became interested in ISO's work searching for "white space" or research opportunities in the strategic areas relevant to the Material Measurement Laboratory (MML).
- The MML collaboration led to a more detailed and focused study of NIST research in the interdisciplinary area of greenhouse gas metrology for a researcher in NIST's Special Programs Office.
- Several researchers have requested analyses of collaborations and co-authorship in their area of research, with a focus on how their specific laboratory contributes to NIST's body of work.
- Staff in NIST's Program Coordination Office wanted to create a network visualization of NIST co-authorship to understand how agency researchers collaborate across organizational lines.
- Materials scientists wanted to understand the landscape of artificial intelligence (AI) and machine learning to identify potential authors for a special issue of a journal, and to later write a perspective paper on AI and machine learning.

Each of these requests presented its own set of challenges. Sometimes the challenge was identifying the right body of literature; other times it was using the right analysis tools to show meaningful results. Close collaborations with researchers were necessary to ensure the project stayed on track and that ISO answered the requests with on-target deliverables. ISO staff met several times with each researcher or research group and emailed or spoke with them on a regular basis.

This paper looks at each type of request listed above and describes the challenges of each while sharing the methods and tools for delivering targeted results.

Background

Bibliometrics is broadly defined by the American Library Association as "the use of statistical methods in the analysis of a body of literature to reveal the historical development of subject fields and patterns of authorship, publication, and use." (Young 1983). Bibliometric activities by library practitioners have focused primarily on citation analysis and research impact assessment until very recently.

The papers written by library practitioners have maintained this narrower focus on the "quantitative measure of research output" (Bladek 2014, 330). For example, the bibliometric services described in Bladek (2014) and Leiss and Gregory (2016) include citation counts, h-index calculations, and impact factor measurements along with customer training and engagement.

This pattern in the literature mirrors actual practice in research libraries, at least in the early part of this decade. Corrall, Kennan, and Afzal (2013, 650) surveyed the bibliometric activities of 140 academic libraries and found that the dominant forms of support offered, after training, were citation reports and calculations of research impact. Cox et al. (2017) develop a competency model for bibliometric practitioners through a survey of 92 practitioners. Many of the "core tasks" emphasized by respondents related to research impact assessment, while tasks such as text mining, network mapping, and keyword analysis were assigned a more peripheral role.

In addition to studying research impacts, academics have been mapping the structure of the sciences and other disciplines since the 1980s; more recent advances in computing allow for a dynamic view of the evolution of research areas (Mingers 2015: 11). Recently, library practitioners have adopted these techniques as well. Eddy and Solomon (2017) document an extensive range of activities, including keyword analyses and journal level co-citations. MacDonald and Dressler (2018) identify research fronts with keyword co-occurrences.

Work in ISO has evolved, also, from an initial focus on h-index studies and journal impacts (Bruss 2013). ISO's services expanded to include bibliometric analysis of collaborations across institutions and geographic regions (Makar 2015). Analyses of paper sets in specific subject areas (Makar 2016) used network analysis to show the structure of scientific sub-disciplines. The case studies below build on this work by identifying emerging areas of research, tracking changes in a discipline over time, mapping intra-agency collaborations, and broadening the bibliometric support services available to library customers.

Case Studies

Identifying Emerging Research Areas

In early 2016, researchers in the Material Measurement Laboratory wished to identify emerging research areas where NIST could have the greatest impact on industry and society. We

decided to assist in this effort to identify "white space"—potential new research areas adjacent to current work—through a combination of literature searches, text analysis, and network visualizations. Guided by the Laboratory's strategic goals (https://mmlstrategy.nist.gov/), we performed two analysis projects with two focus areas: microbial metrology in late 2016 and water quality metrology in early 2017.

Search strategies for identifying the two metrology paper sets involved harvesting keywords from strategy documents and identifying important journals. These searches were difficult because we were often looking across disparate disciplines for metrology-related work. NIST's work focuses on measurement science in the areas of physics, chemistry, biology, and engineering. Some analysis considerations included the size of the paper set returned and the level of connection between papers, based on authors, citations, keywords, etc.

We experimented with a variety of bibliometric and visualization tools to both learn about the paper set and see what kinds of analysis would be possible. Our initial analysis focused on keywords: we extracted keyword frequencies in Bibliometrix (a package available for use with the R statistical computing language) and identified emerging or "bursting" keywords with the Science of Science (Sci2) algorithms[2]. While the keyword analysis provided some insights about new research methods and instrumentation, more useful topical analyses were generated using CiteSpace visualizations of clustered disciplines and keywords.

Figure 1 shows the linkages between keywords and disciplines in the microbial metrology paper set. With a cluster analysis of this network, we identified trends that may be of interest to our customers, including: real-time monitoring of aerobic biodegradation, spectrophotometry in food production facilities, and stable isotope/high precision concentration measurements for methane.

[2] Identification of commercial entities is not intended to imply recommendations or endorsement by the National Institute of Standards and Technology.

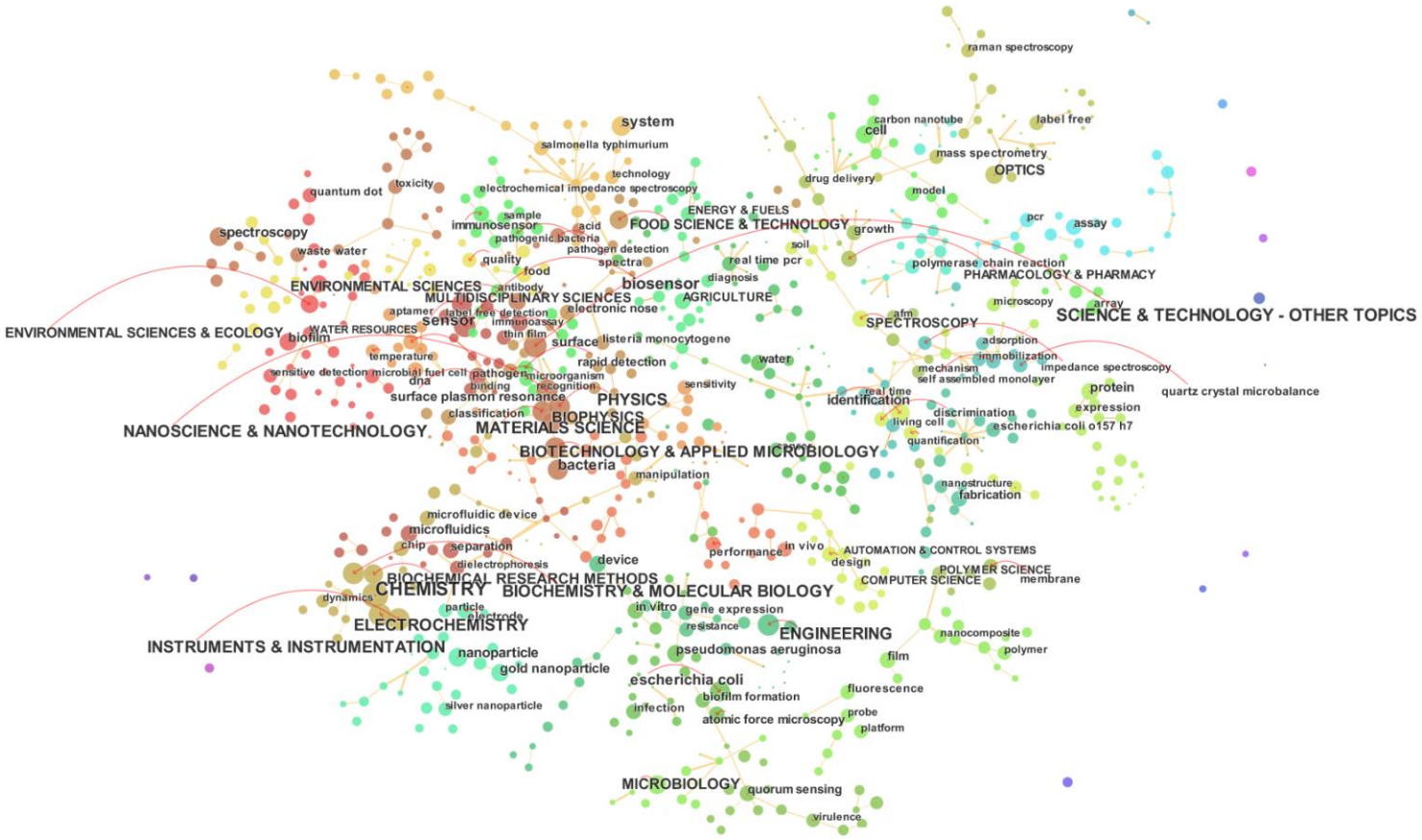

Figure 1. Keywords and Subject Disciplines in Microbial Metrology

Figure 2 shows groups co-occurring keywords related to water quality metrology. A spectral clustering algorithm (Chen et al. 2010) identified 12 groups, or clusters, of keywords. Several recently emerging clusters are described below.

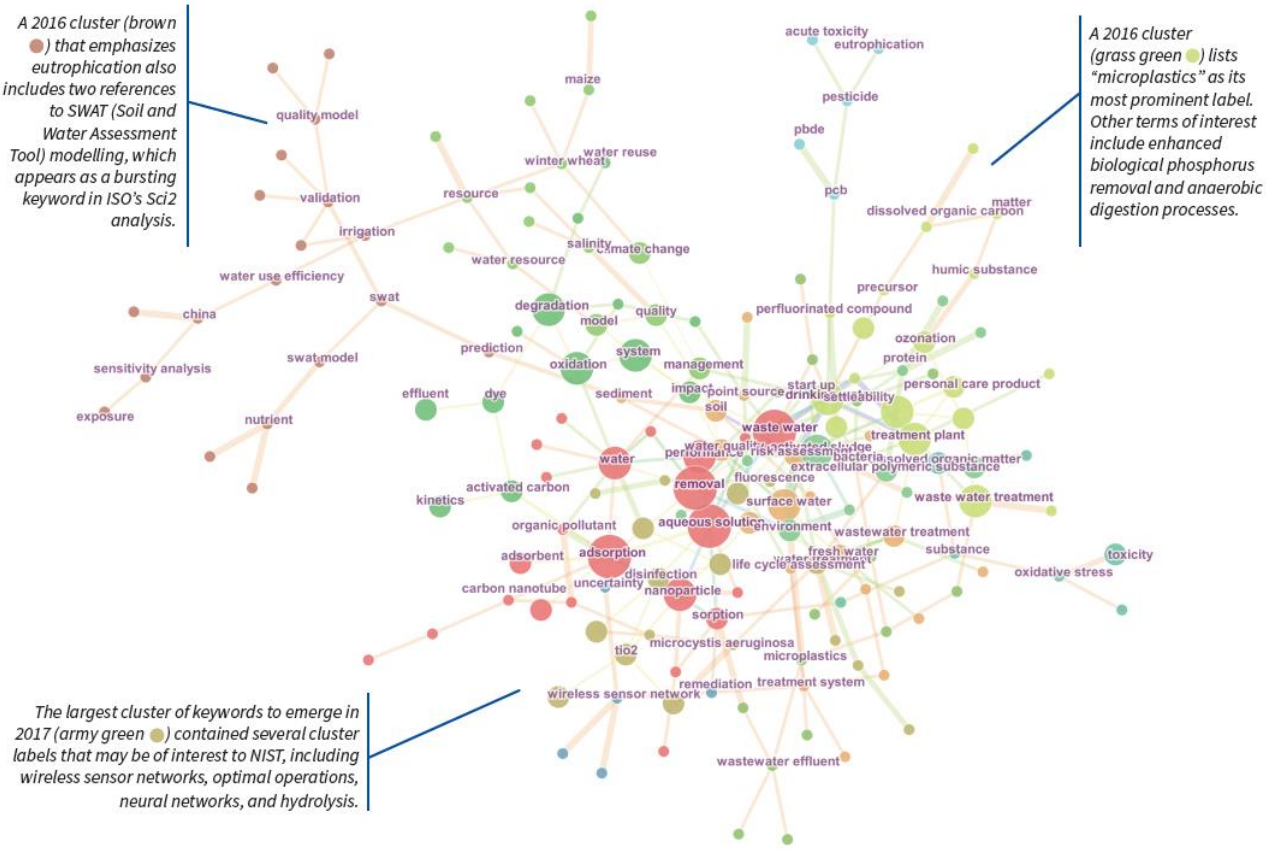

Figure 2. Emerging Clusters of Subject Keywords in Water Quality Metrology

These analyses, which were in many ways more ambitious than later efforts, presented some challenges. Large paper sets from disparate disciplines were not closely interconnected. This low modularity limited the types of networks that we could create. Rather than co-citation or co-author networks, we connected papers via common subject keywords. The resulting visualizations were interesting to generalists but lacked the specificity to inform experts in the field. We remain confident, though, that this type of predictive analysis will continue to grow in importance as text mining algorithms improve.

Keyword Analysis over Time

In the summer of 2017, we were contacted by a customer who coordinates a large portfolio of research projects related to greenhouse gas measurement. She was interested in learning how the field had evolved in the past five years and asked for assistance at the library information desk after viewing some of the network visualizations already completed by the library staff to track research sub-disciplines in forensic science. After a series of initial meetings, we decided the best way forward was to track trending keywords in the publications of the 62 principle investigators associated with GHG measurement.

The initial search was straightforward, but creating the paper set was tricky. There were many discussions with the customer about distinguishing different areas of research by the same investigator in order to limit our search to only papers related to greenhouse gas measurement. After several search iterations, the final paper set was narrowed to include 522 journal articles, proceedings papers, and reviews published between 2013 and 2017.

To provide the customer an overview of the research portfolio, we created a summary report of the full paper set with author and organization names, research areas, and keyword frequencies. This information was gathered from *Web of Science*'s Citation Reports and enhanced by a Bibliometrix summary report in R. Next, with the CiteSpace tool, we generated a keyword co-occurrence network for the paper set. Each keyword, despite occurring over multiple years, was only assigned a single year based on changes in frequency from year to year: what CiteSpace calls "time slicing." The networks generated by CiteSpace were exported into the Gephi network visualization tool. Network visualizations of keywords peaking in 2013 and 2017 are found below in Figure 3 and Figure 4.

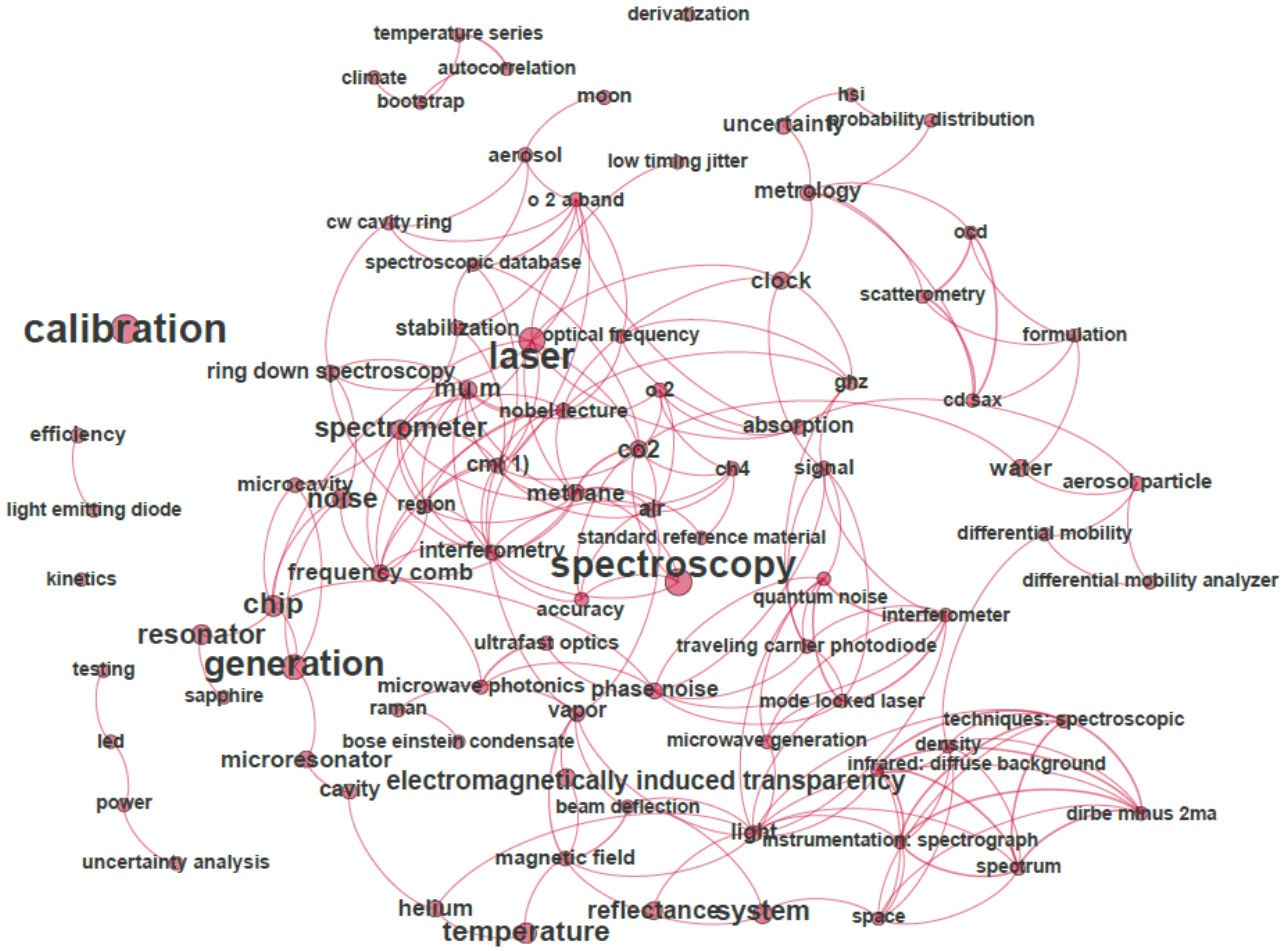

Figure 3. Keyword Frequency and co-occurrence, 2013.

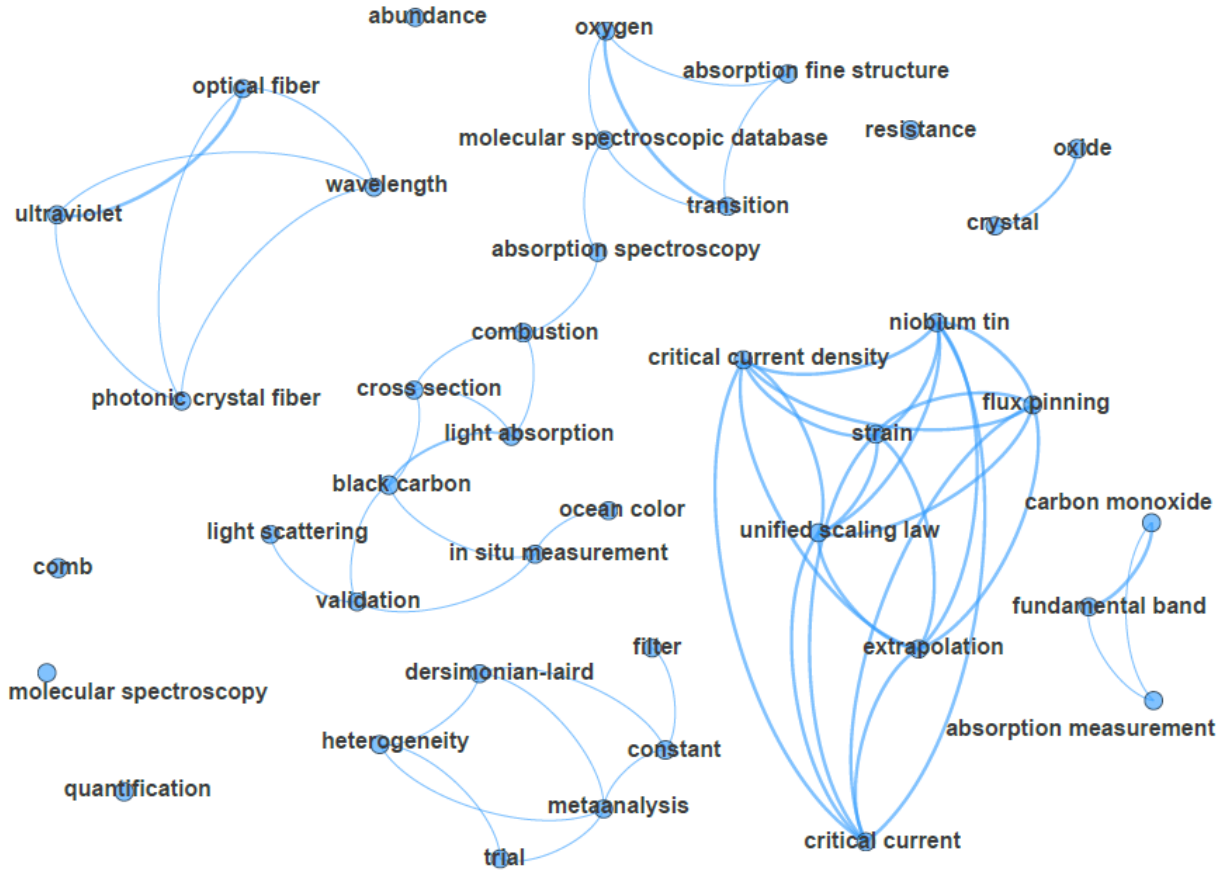

Figure 4. Keyword Frequency and co-occurrence, 2017.

Because the paper set was carefully curated and the researcher was maintaining a broad disciplinary focus, keyword networks were useful here. In other collaborations, we have found keyword terms to be too general to provide insight into the evolution of a research area.

Mapping Collaborations

With more than 2,000 researchers publishing in scientific journals each year, the structure of collaborations, both within NIST and with outsiders, is of keen interest to NIST management at many levels. Below we describe three bibliometric analyses performed to assess both collaborations within NIST and the prevalence of NIST researchers in the body of literature for specific subject areas.

In September 2017, a scientific advisor who was working on strategy and road-mapping approached us for help analyzing a body of environmental literature. She wanted to understand the contributions of her lab, the Material Measurement Laboratory (MML), relative to other environmental science work at NIST. In this case, continuous collaboration with the customer allowed us to generate a set of results very useful to her. We provided her with a list of 2,636 articles, reviews, and proceedings papers. She narrowed this count by hand to 468 papers with

any sort of emphasis on environmental science, and assigned each environmental science paper to one of 13 topic areas. We then used a simple text matching script in R to match the author records to an employee directory, illustrating her lab's involvement in various topic areas (see Figure 5). Our customer's lab (and the divisions within that lab) was involved in nearly half (233 out of 468) of the papers, with a particularly high representation in environmental chemistry (55/81 papers) and exposure (53/73 papers).

A second MML scientific advisor then asked us to assess MML's current involvement in chemical manufacturing and identify potential challenges and opportunities for the lab. Once again, we combined our own internal records about NIST staff with bibliographic records from *Web of Science* to identify MML authors in 258 of 524 articles in this paper set. In this case, the scientific advisor, being the subject matter expert, provided a list of keywords. Our initial search results, based on these keywords, proved to be too broad upon review by the customer. She then tightened and modified the list of keywords, and we reran the search. This time the search results were on target. The effort was successful because we worked very closely with the customer through an iterative process to identify the appropriate paper set.

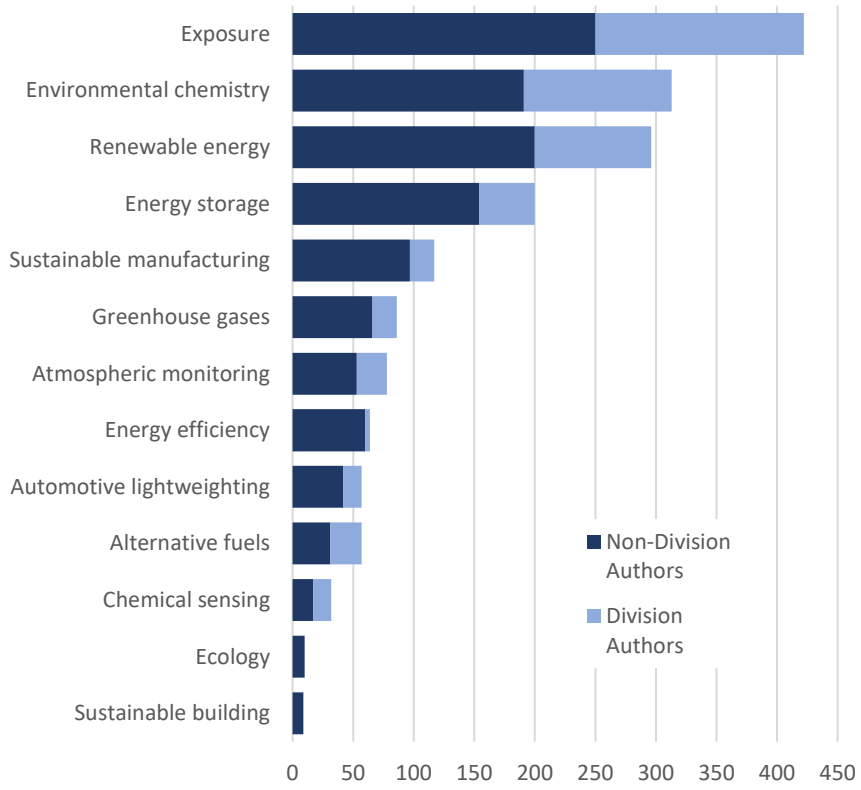

Figure 5. Authorship by Focus Area in Environmental Science, 2016 and 2017

Finally, staff in NIST's Program Coordination Office wanted to understand the full extent of collaborations across divisions within NIST. In this case, the paper set was relatively easy to

identify. The bulk of the work involved assigning each author in *Web of Science* to a NIST division. This was accomplished through a combination of scripting and manual examination. The resulting database of more than 2,000 NIST authors and 3,416 papers could be queried to show collaborations at the individual, group, or institutional level. We built our own node and edge tables using a relational database and created a series of visualizations in Gephi that were very useful to our customer, showing the number of publications (dot size) and number of collaborations (line thickness) (see Figure 6).

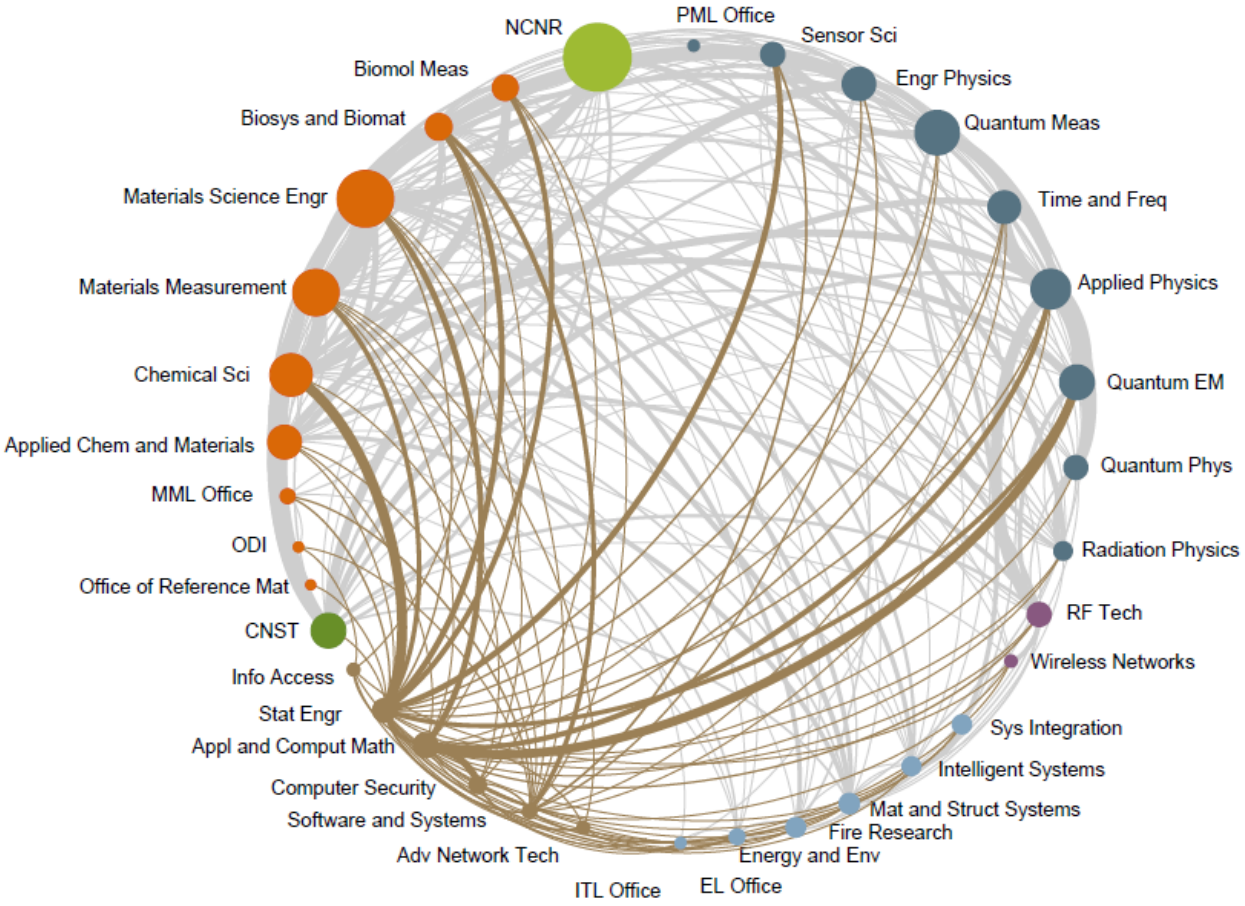

Figure 6. Collaborations within NIST, 2016 and 2017

With continuous communication and meetings, we were able to discern the real intent behind customer questions. This meant that the customer also needed to consult with her colleagues to discuss options as a follow-up to our meetings with her. Communication became the key to meeting and even exceeding the customer's needs and expectations.

Landscape Analysis and Assessment

In January 2018, a group of materials scientists approached us for help in understanding research related to artificial intelligence (AI) and machine learning within their discipline. More

specifically, the scientists were interested in identifying potential authors for a special issue of a journal and illustrating this process though a perspective paper about the use of AI and machine learning to identify authors and emerging topic areas.

We already had experienced difficulties searching the AI and machine learning literature in the *Web of Science* database by keyword; *Web of Science* does not offer subject indexing or a thesaurus. Since machine learning can be described in many ways, important articles might have been missed. Even the search for AI articles proved problematic since not all AI was relevant to the researchers.

We then decided it was best to perform author searches for individuals who were known to be working in NIST's fields of interest. However, this posed a different set of challenges, since there were over 100 researchers, who over the course of 10 years may have worked at multiple institutions and published research results outside of AI and machine learning. We retrieved 7,177 papers of varying degrees of relevancy, but our customers were comfortable with this limited precision so long as all important papers were included in the set (100% recall).

We conducted some initial analysis on this paper set in Sci2 and Gephi, and linked bibliographic data with employee directory records. We also educated the researchers about the tools we had used in the past, noting the strengths and weaknesses of each. However, in a departure from previous collaborations, these researchers conducted much of the bibliometric analysis themselves. Using our work as a starting point, they applied machine learning algorithms and topic models using Python libraries to isolate the most relevant papers in the initial set.

We intend to replicate our approach to this project with other NIST researchers. This hybrid model of both creating and enabling analysis would be effective at any organization with a highly technical scientific workforce. Similar work is already in the offing with other labs, as we are beginning a partnership in Spring 2018 with the nanofabrication research group to uncover emerging areas of research in their field.

Conclusions

ISO's expansion of bibliometric services has required a change in our standard practices. Search strategies have evolved to include larger paper sets in some cases and more explicitly linked papers in others. Staff have developed new expertise in a broad range of bibliometric tools and program libraries. The creation of new support services, which tie closely to a customer's existing research portfolio, requires a degree of interaction beyond that of a simple reference interview. ISO has found that a more iterative process with regular contact ensures the most useful end product.

ISO's current approach is limited by the amount of text available via a standard bibliographic search in *Web of Science*. We hope to conduct text analysis of all sections of a research paper in future collaborations, perhaps through the use of arXiv records or the public records of NIST's own research in PubMed Central. Future results will also be more useful to our customers as clustering and text mining algorithms continue to improve.

References

Bladek, Marta. "Bibilometrics services and the academic library: meeting the emerging needs of the campus community." *College & Undergraduate Libraries* 21, no. 3-4 (2014): 330-344.

Bruss, Stacy, and Susan Makar. “Bibliometrics and Impact Analyses by Librarians at the National Institute of Standards and Technology.” Presented at Bibliometrics by Librarians and Information Professionals, January, 2013. Online at https://www.nist.gov/publications/bibliometric-and-impact-analyses-librarians-national-institute-standards-and-technology.

Chen, Chaomei, Fidelia Ibekwe-SanJuan, and Jianhua Hou. “The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis.” *Journal of the American Society for Information Science and Technology* 61, no. 7 (2010): 1386-1409.

Corrall, Sheila, Mary Anne Kennan, and Waseem Afzal. "Bibliometrics and research data management services: Emerging trends in library support for research." *Library Trends* 61, no. 3 (2013): 636-674.

Cox, Andrew, Elizabeth Gadd, Sabrina Petersohn, and Laura Sbaffi. "Competencies for bibliometrics." *Journal of Librarianship and Information Science* (2017): 0961000617728111.

Eddy, Mark A., and Daniela Solomon. "Leveraging Librarian Liaison Expertise in a New Consultancy Role." *The Journal of Academic Librarianship* 43, no. 2 (2017): 121-127.

Leiss, Caroline, and Kathleen Gregory. "Visibility and Impact of Research: Bibliometric Services for University Management and Academic Staff." (2017).

MacDonald, Karen I., and Virginia Dressler. "Using Citation Analysis to Identify Research Fronts: A Case Study with the Internet of Things." *Science & Technology Libraries* (2018): 1-16.

Makar, Susan, and Amanda Malanowski. “Taking Impact Metrics to the Next Level at the National Institute of Standards and Technology.” Presented at the Special Libraries Association Annual Conference, June 2015. Online at https://www.nist.gov/publications/taking-impact-metrics-next-level-national-institute-standards-and-technology.

Makar, Susan, Amanda Malanowski, and Katie Rapp. "Visualizing Forensic Publication Impacts and Collaborations: Presenting at a Scientific Venue Leads to Increased Collaborations between Researchers and Information Professionals." *Science & Technology Libraries* 35, no. 2 (2016): 109-118.

Mingers, John, and Loet Leydesdorff. "A review of theory and practice in scientometrics." *European Journal of Operational Research* 246, no. 1 (2015): 1-19.

Young, Heartsill, and Terry Belanger. *The ALA Glossary of Library and Information Science*. Chicago: American Library Association, 1983.