Connecting Digital Scholarly Researcher Identifiers & University Systems: Assessment of Engineering Faculty at a Major Research Institution

Dr. Sarah Over, Virginia Polytechnic Institute & State University

Dr. Sarah Over is the Assistant Director for Research Intelligence at Virginia Tech, serving as their Engineering Librarian and representative for their new Patent and Trademark Resource Center. She is also part of a team focused on research impact and intelligence to support the College of Engineering and Office of Research and Innovation at Virginia Tech. Dr. Over's background is in aerospace and nuclear engineering, with years of experience teaching engineering research methods and introductory coding.

Dr. Mengyu Yin, Virginia Polytechnic Institute and State University

Mengyu Yin is a Research Intelligence Data Scientist at Virginia Tech Libraries, where she supports research analytics and evaluation initiatives using data-driven approaches. She holds a Ph.D. in Agricultural Economics from Oklahoma State University (2021). Prior to joining Virginia Tech, she worked for nearly three years as a data scientist in the consulting industry, developing analytical solutions for clients across multiple sectors. Her expertise includes statistical modeling, machine learning, data visualization, and bibliometric and research analytics. At Virginia Tech, she works with research intelligence tools such as SciVal, Elsevier data products, and Overton to support research evaluation and strategic decision-making.

Emily Mazure, Virginia Polytechnic Institute and State University

Emily Mazure is the Research Impact Librarian at Virginia Tech University Libraries. She specializes in providing support to individuals, teams, departments, and others to increase the visibility of their work and thus their impact. Her expertise includes managing scholarly profiles online, exploring and analyzing scholarly metrics, finding strategic collaborators, and more.

Prof. Connie Stovall, Virginia Polytechnic Institute and State University

As Director for Research Impact & Intelligence, I collaborate with campus stakeholders to translate information to insights. We utilize bibliometric, impact, institutional, funding, and industry data from sources such as Scival, Scopus, Web of Science, Mergent, NSF HERD, IPEDs, Funding Institutional and employ a variety of visualization tools such as Tableau and VosViewer to help identify research competencies, to understand collaboration networks and potential partnerships, and to demonstrate impact.

Shehryar Khan, Virginia Tech

Shehryar Khan is a Graduate Research Assistant in the Department of Research, Impact, and Intelligence (RII) at Virginia Tech, where he is currently pursuing an M.Eng. in Computer Science under the advisement of Dr. Sarah Over. His research focuses on the intersection of machine learning and systems, with specific interests in ML optimizations, post-training evaluations, and security. At RII, his work centers on the development of specialized Large Language Models (LLMs) for patent examination and the engineering of robust systems for the deduplication and organization of large-scale scholarly data. His broader professional interests lie in building high-impact software systems and improving the efficiency of data-driven intelligence.

Emma Jiren Wang, Virginia Polytechnic Institute and State University

Emma Jiren Wang holds a Master of Engineering in Computer Science from Virginia Tech. Her academic interests span Human–Computer Interaction (HCI), Engineering Education, and Library with Information Science. Her research explores interactive agent design, affective computing, computer science education, and social virtual reality.

Brian Craig, Virginia Polytechnic Institute and State University

Brian Craig is an associate, assisting the Research Impact and Intelligence team in multiple ways. With a background in liberal arts, SEO writing, social media, graphic design, printing, customer service, pet care, library collections management, illustration, and fine art, he has a diverse skillset.

Rachel Ann Miles, Virginia Polytechnic Institute and State University

Rachel Miles is the Research Impact Coordinator at Virginia Tech University Libraries. She specializes in research analytics and provides expertise in bibliometrics, altmetrics, research communication, and scholarly publishing. Rachel works closely with faculty, researchers, and administrators to manage the university's Research Information Management (RIM) system, interpret research impact data, and support responsible, ethical research evaluation. She also advocates for awareness of how academic culture and incentive systems affect researcher well-being and behavior.

Connecting Digital Scholarly Researcher Identifiers & University Systems: Assessment of Engineering Faculty at a Major Research Institution

Abstract:

At Virginia Tech, a unit within University Libraries is responsible for supporting institution-wide efforts to improve impact and visibility of campus research. Beyond providing consultations and workshops for growing research impact, the unit evaluated and improved the research data ecosystem to ensure more complete and accurate data capture. Part of this process involved assessment of how researchers use, if at all, researcher scholarly identifiers (ID). Digital scholarly researcher identifiers, such as the ORCID iD and Scopus Author ID, are important for disambiguating researchers' works and helping to capture accurate global impact data. This is especially critical for our institution's College of Engineering, which contributes a large share of campus research outputs.

The researcher IDs and systems involved in this project are Scopus Author IDs, ORCID iDs, Academic Analytics, Symplectic Elements profiles, and our institution's Banner instance (human resources records). This paper will cover the process of finding, assessing, and integrating identifiers across systems for use in research impact reporting and benchmarking, including ways this can be applied to any research institution. The College of Engineering at Virginia Tech's status of researcher IDs will be detailed along with related efforts by the University Libraries' unit. Some surprising results included the number of duplicate IDs and incorrect institutions associated with IDs, leading to inaccurate reporting metrics. In the future, this work will be contained within an internal university data site for assessment by department heads and other leadership to improve Virginia Tech's use of these key researcher IDs and their corresponding metrics.

Key words: research impact; research evaluation; research assessment; research analytics; bibliometrics; scientometrics; institutional research

Introduction

It is not uncommon for large research institutions to tout their research impact, and Virginia Tech is no exception. Virginia Tech is a large, public R1 land-grant university with close to 40,000 students, over 4000 researchers, and $453.4M in sponsored research, a significant portion derived from STEM-based research activities. Its College of Engineering, perhaps the most reputable, ranked in the top 15 nationally for engineering research expenditures by the 2024 National Science Foundation's Higher Education Research and Development (NSF HERD) survey [1]. The college consists of 13 departments, and many of its faculty also work in interdisciplinary centers and institutes throughout the university, and part of the college's role, like the university's at large, is to act as an economic engine for the regional economy.

Much like peers of its size, Virginia Tech and its College of Engineering set goals for growth and benchmarks progress. In addition to routine benchmarking for the College of Engineering, Virginia Tech developed the Global Distinction initiative that centers on its "commitment to empowering impactful research, scholarship, and creative activity," with a particular focus on growing its global reputation, or brand [2]. While Virginia Tech's College of Engineering is ranked in the top 15 nationally, it is ranked in the 251-300 segment globally [3], [4]. As such, many on campus are engaged in benchmarking researchers across diverse metrics, notably many (but not exclusively) included in The Times Higher Education World University Rankings (WUR), and much of the data used for WUR comes from Elsevier's Scopus and SciVal databases.

Additionally, the university benchmarks for many other purposes centered around productivity and impact, and it utilizes a broad array of data sources to do so. However, much of the benchmarking analysis has occurred at the aggregate level, partly because it is difficult to find reliable sources that account for researcher name ambiguity. To that end, our team began a project to identify, validate, and disambiguate researcher names to create more reliable benchmarking – and to supply campus stakeholders with metrics who needed them to nominate our researchers for prestigious awards, an activity for which Virginia Tech lagged behind its aspirational peers. After several months of reconciling university roster names with Scopus profiles and their unique identifiers, both through Elsevier's Application Programming Interfaces (APIs) and through manual means, the team compiled a list of Scopus Author IDs that now allows for calling and conducting publication analysis in a way never done before at Virginia Tech.

Background & Data Sources

Academic librarians are well aware of the important role researcher IDs play in the research life cycle, making discussions of ORCID and Google Scholar frequent discussion points with faculty. However, faculty generally choose to spend their time on the next grant or publication instead of their ID profiles, which is more likely to be rewarded and lead to promotions. A recent study at the University of Manitoba found that for their sample population of researchers, 45.5% had an ORCID profile (75% if you include disputable profiles), 21.2% had a Publons (WoS ResearcherID) profile, and 85.4% had a Scopus Author Profile [5]. Some libraries in consultation with their research offices or other university units have made efforts to raise awareness of researcher IDs to their faculty [6]. Maintenance of these profiles or even uptake can certainly be low for faculty even in STEM, where the processes can be more automated than humanities and arts [5], [7]. Part of maintenance is keeping track of new profiles created by a system as well since this can result in a lowered h-index if split across multiple profiles [8]. Overall, our work seeks to mitigate some of these limitations by taking on curation that can lead to automated results in the future after initial investment.

Virginia Tech University Libraries have used a variety of data sources for researcher IDs. As each source has different limitations and data coverage, multiple sources are frequently needed for assessment. Below, we describe each data source we used, including data coverage and any notable limitations. Note that this work can also be done using institutional faculty reporting systems (research information management systems) and ORCID iDs if funds have not allowed for subscription to Elsevier products.

Scopus

Scopus is a multidisciplinary abstract and citation database that includes scholarly literature including journal articles, books, conference proceedings, and more. It has over 100 million records and 2.4 billion cited references [9]. In addition, it generates author profiles automatically whenever two or more articles are linked to the same name. There are currently over 20.5 million author profiles in Scopus, all assigned a Scopus Author ID, a unique number used across Elsevier systems that compile the works associated with that author. Using the metadata from those works, profiles also compile name variations, affiliations, email addresses, subject areas, citations, and co-authors.

ORCID

Open Researcher and Contributor ID (ORCID, orcid.org) is a global, not-for-profit organization with a mission to enable transparent and trustworthy connections between researchers, their contributions, and their affiliations. ORCID provides the ORCID iD, a unique, persistent, 16-digit number identifier for researchers. Researchers can register for free for an ORCID iD and then use the ORCID record (or profile) to connect their research activities – such as scholarly

works, affiliations, professional activities, and funding – to other systems and profiles. Many publishers and journals now require authors to include their ORCID iD when submitting manuscripts. Some federal agencies, such as the National Institutes of Health (NIH) and the National Science Foundation (NSF), also require ORCID iDs for grant applications. ORCID enables integration with other systems, allowing information to be shared between ORCID and those platforms. Researchers are able to configure these connections and choose settings that support automatic updates to their ORCID record.

Academic Analytics

Academic Analytics is a comprehensive database and analysis product focused on faculty at research and medical institutions. The database contains a wealth of data on faculty publications, citations, awards, grants, patents, and clinical trials from over 400,000 faculty across 12,000 departments in 470 universities in the US and abroad [10]. Similar to Scopus and ORCID, identifier numbers are assigned to institutions and faculty members in the database. However, Academic Analytics focuses on tenured/tenure-track and research faculty for most institutions, so other types of faculty are not generally included. Each year, all participating universities supply their rosters to Academic Analytics, including their department and rank. In addition to faculty research activity, Academic Analytics allows subscribers to benchmark against peers such as how their mechanical engineering department compares to others in numbers, funding, publications, and more. One of its most important features is the ability to find prestigious awards won by faculty.

Symplectic Elements

Our institution uses the Research Information Management System (RIMS), Symplectic Elements, to collect faculty activity data [11]. All faculty and graduate students have profiles automatically created in the system, which they can choose to claim and populate with additional data. Profiles are initially populated with core demographic and affiliation information – such as name, prefix, title, college, and department or other unit – ingested from human resources data to ensure that everyone has a profile in the system. A recent analysis found that 95 percent of tenured and tenure-track instructional faculty have claimed their Elements profiles. Those faculty primarily use the system to support annual reporting through their Faculty Activity Reports (FARs). Smaller percentages of other groups, such as graduate students, postdocs, and administrative faculty, claim their profiles. For those groups, Elements doesn't fulfill a similar requirement and those who claim and populate their profile are primarily leveraging it to generate a public profile in Virginia Tech Experts (experts.vt.edu).

The system also ingests data from multiple institutional sources, including Summit (the Office of Sponsored Programs' grant proposal and award system), Virginia Tech Data Repository (data.lib.vt.edu), VTechWorks (vtechworks.lib.vt.edu), and the student information system for teaching and course data. Elements offers an Automatic Claiming feature that helps users collect

data about their scholarly work by enabling connections to external databases and profiling systems. Users can search for, find, and connect identifiers from ORCID, Scopus, Web of Science, Dimensions, and more. As a result, Elements can serve as a centralized source for identifying and collecting researcher identifiers that users have claimed within the system. ORCID iDs are particularly reliable because users are required to log in and explicitly link their ORCID and Elements accounts. Although uncommon, other identifiers, such as Scopus Author IDs, have the potential to be claimed incorrectly by a user; as a result, their accuracy is less reliable. Users are able to review, verify, and supplement automatically imported data after claiming their profiles.

Methodology for Data Automation

As we highlight throughout the following sections, faculty usage of profiles produces data that needs extensive manual curation. Global Distinction efforts at Virginia Tech have only accelerated the need for accurate, usable data as well. Faculty are also encouraged to publish more; meaning they are less likely to take the time to manage all their profiles (including our engineering faculty who drive many of our ranking metrics). As manual curation work takes time, our team decided to pick one source to start and gradually add more as the project gains momentum with our campus stakeholders and faculty.

We developed a structured workflow to identify and validate Scopus Author IDs (AUID as they are abbreviated within Elsevier's systems and in this paper) for engineering faculty at Virginia Tech to support institution-scale research assessment. Scopus was selected in consultation with the Office of Research and Innovation due to its curated coverage and use in external rankings, including the US News and World Report. The workflow combines automated querying via the Scopus API, with high-precision matching strategies, in which uncertain matches are intentionally left unresolved, alongside manual verification to prioritize accuracy while explicitly managing ambiguity.

Finding and Processing Scopus Author IDs

To identify AUIDs for engineering faculty at an institutional scale, we implemented a Python-based workflow using the Elsevier Scopus Author Search API (Figure 1). The workflow accepts a roster of faculty names and issues structured API queries to retrieve candidate author profiles. Because name-based querying frequently yields ambiguous results (e.g., homonyms or fragmented author profiles), the workflow was deliberately designed to accept only high-confidence automated matches while explicitly identifying uncertain cases for human review.

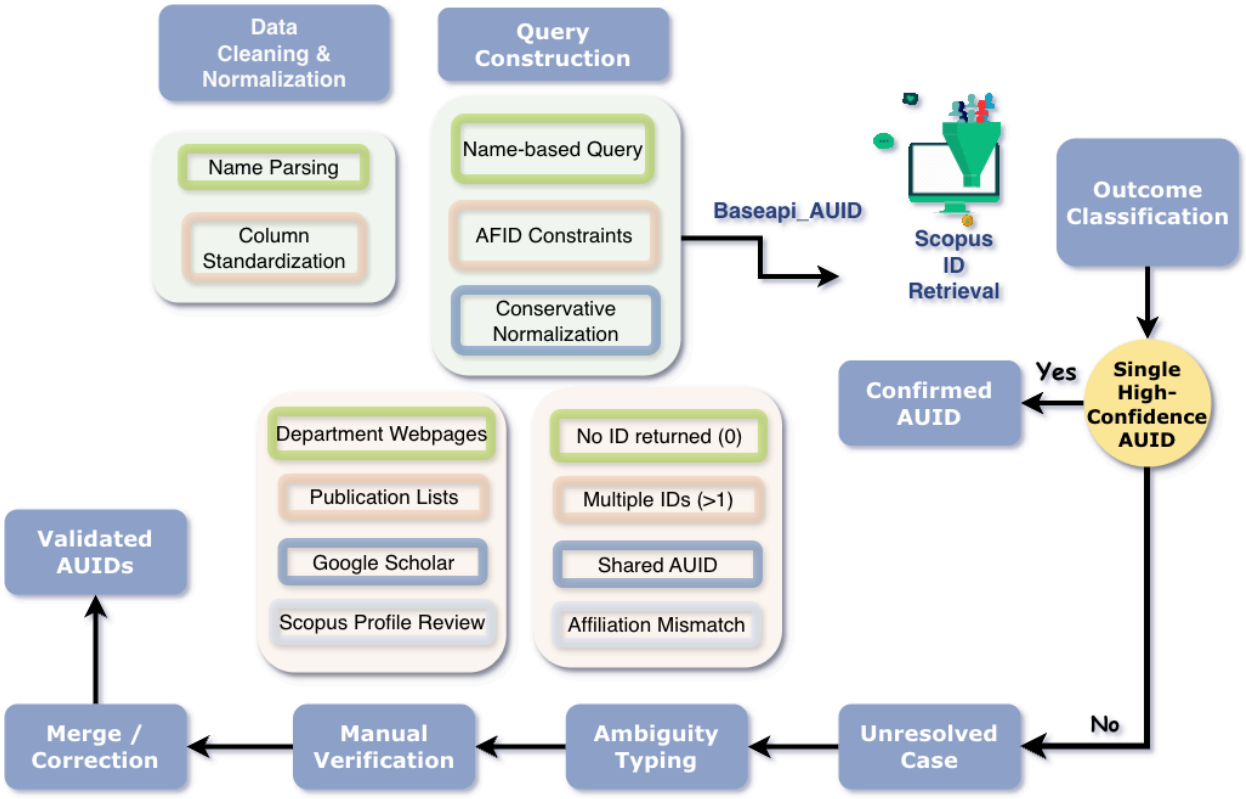

Figure 1: Workflow for AUID Identification and Validation. The diagram illustrates the structured process developed to identify and validate AUIDs for engineering faculty. The workflow integrates roster cleaning and normalization, API-based author querying, conservative outcome classification, and targeted manual verification to ensure high-confidence identifier assignment while explicitly managing ambiguous cases. (Note: AFID is Affiliation ID, an identifier for institution affiliation.)

Step 1: Preparation and Standardization of Faculty Roster Data

As the initial procedural step, a roster of engineering faculty and other colleges on campus was obtained from university human resources and prepared for automated processing. The data were cleaned and organized to ensure consistency in author name representation, primarily by adjusting and aligning relevant columns, including separating and mapping name fields and standardizing the column structure across records. This standardized roster served as the input for all subsequent automated identification steps.

Step 2: API-Based Author Identification Under System Constraints

The standardized faculty roster was used to query the Elsevier Scopus Author Search API to identify candidate AUIDs. Queries were constructed using author names and institutional affiliation identifiers (AFIDs) to improve precision. During execution of the API-based author identification step, several practical challenges became evident that limited the effectiveness of naïve automated querying at institutional scale.

First, name ambiguity was common: author names were parsed using restrained normalization rules that favored precision over recall, thereby reducing the risk of conflating distinct individuals with similar names. In practice, a number of faculty shared identical or highly similar first and last names, which occasionally caused the API to return the same AUID for different individuals. An initial mitigation strategy involved incorporating middle names or initials into query construction to further differentiate authors. However, this approach introduced additional inaccuracies, as middle names are not consistently represented in Scopus records; some authors do not include middle names in their profiles, while others may have changed name representations over time. As a result, reliance on middle-name matching led to an increase in false “not found” outcomes, reducing overall data coverage. To balance accuracy with completeness, the workflow ultimately avoided strict middle-name enforcement and instead preserved cases in which multiple faculty members were associated with the same AUID as unresolved. These cases were explicitly filtered from automated matches and deferred for manual investigation. Although this approach increased the need for human review, it ensured that valid author records were not excluded from the dataset and prevented incorrect automated assignments. This design choice reflects a deliberate trade-off favoring data completeness and integrity over aggressive automated disambiguation.

Second, institutional affiliation information within Scopus is not consistently represented across author profiles. Faculty members may be associated with alternate institutional names, interdisciplinary units, or legacy affiliations, which can result in incomplete retrieval when affiliation constraints are applied too narrowly. For example, College of Engineering, Virginia Tech was treated differently than Virginia Tech. To address this issue, the workflow implemented affiliation-based query scoping using a curated list of institutional affiliation identifiers rather than relying on a single institutional record. By incorporating multiple known affiliation identifiers into each query, the approach increased coverage of institution-affiliated authors while maintaining sufficient precision to exclude profiles from unrelated institutions.

Third, author name data also exhibit substantial variability across systems, including differences in the use of initials, middle names, diacritics, hyphenation, and name changes over time. These inconsistencies can significantly affect search results, increasing the likelihood of ambiguous matches or missing profiles when strict matching criteria are applied. To mitigate this challenge, the workflow adopted conservative name parsing and avoided aggressive disambiguation strategies that could exclude valid records. In parallel, fault-tolerant API execution was incorporated through enforced delays between requests, limited retry attempts for failed queries, and systematic logging of errors and response codes. Together, these measures supported scalable and repeatable querying across large faculty rosters while minimizing false-positive matches, preserving transparency around unresolved cases, and preventing silent data loss due to unhandled API failures.

Step 3: Classification of Identification Outcomes and Output Management

Retrieved Scopus Author Search results were classified based on identifier uniqueness. Records returning exactly one AUID were considered high-confidence matches, and the identifier was directly recorded and displayed. Records returning no profiles or more than one AUID were classified as unresolved; in these cases, the workflow recorded and displayed the observed number of returned IDs (0 or >1) to explicitly capture ambiguity. Unresolved outcomes typically resulted from author name homonymy and/or multiple AUIDs associated with a single researcher. To avoid erroneous automated assignment, unresolved records were retained for subsequent manual verification and targeted searching.

To support downstream processing, results were exported into two separate datasets: (1) a confirmed file containing faculty with a single AUID for direct integration into institutional systems and bibliometric analysis, and (2) an unresolved file containing faculty with zero or multiple AUIDs returned, requiring manual review and profile reconciliation. This separation minimized downstream errors and ensured that automated metrics were computed only from high-confidence identifiers.

Step 4: Manual Verification and Faculty Validation of Scopus Author IDs

Following automated classification and output segregation, unresolved cases required targeted manual investigation to confirm or correct AUIDs. These cases primarily included authors for whom no Scopus profile was returned under automated constraints, as well as authors associated with multiple candidate profiles. Manual research drew on supplementary information sources such as departmental webpages, publication lists, and other scholarly identity systems to verify author identity and determine appropriate Scopus profiles or merge requirements.

During this process, it became evident that a subset of faculty had not actively managed their Scopus Author profiles, including outdated affiliations, unmerged duplicate profiles, or incomplete metadata. The lack of proactive profile management substantially increased the labor required for manual reconciliation, particularly for authors with long publication histories or multiple institutional affiliations. As a result, manual verification remains a necessary but resource-intensive component of the workflow. Further details are presented in the following section on manual searching since it requires significant effort and attention to detail.

These steps are ongoing due to regular faculty turnover at Virginia Tech and informs future work aimed at reducing manual burden through faculty engagement and guidance. Based on observed patterns, future efforts will focus on developing clear instructions and outreach materials to encourage faculty to review, maintain, and validate their AUIDs. By promoting consistent identifier management practices among faculty, institutions may reduce downstream ambiguity, improve data quality across systems, and increase the effectiveness of automated research assessment workflows.

The Manual Search for Scopus Author IDs

Although many scholarly IDs can be identified through the API process above, many unfortunately remained unidentified. In the case of the College of Engineering alone, this amounted to 283 faculty of 731 total. Of the 283, 210 researchers had no AUID returned through the API, while 73 returned more than one AUID, requiring further investigation. Unless someone manually searched for them, the researchers' work would not be credited, certainly an unacceptable proposition.

To help ensure all our researchers were properly credited, our team spent months manually searching (for all colleges on campus) to confirm if an AUID existed for anyone the API did not find matches. This involved searching for clues from multiple sources such as college and professional websites, Google Scholar, OpenAlex, ResearchGate, and LinkedIn. It takes determination, but if scholars have made efforts to showcase their work online, it greatly helps.

With time, we discovered complex reasons that cause AUIDs to remain missing or incorrect. For example, it can happen when just one letter is off in a name, or when their name is the same as another person's. Many authors even use nicknames or names different from their scholarly profiles. In these cases, this presents considerable challenges and requires thorough investigation to locate the correct person.

Sometimes, AUIDs remain unidentified when the author's college affiliation is not shown as primarily Virginia Tech, but rather a related entity. As a result, our team has worked on unifying the hierarchy of all our researchers, as well as making it easier to find them if their affiliation in Scopus has a spelling or other variation.

In the project's initial search via API, we determined when a scholar likely has multiple AUIDs. However, these each needed to be manually located, verified, and merged into one AUID. While most authors had no more than several, for some outstandingly prolific authors, Scopus has generated a dozen or more over the years. In those cases, a source like Google Scholar is needed to correctly identify them. Once we had all the AUIDs, we can select each profile and send a merge request to Elsevier directly on the Scopus website. We selected the authors by checking the box next to their name and then click, "Request to merge authors" at the top of the screen.

Sometimes, there was no way to manually request a merge on the website for various reasons. For example, an author's name changed or had a spelling change. However, we have learned there is a way to batch upload these complex cases into a spreadsheet of authors and AUIDs provided by SciVal, another of Elsevier's products we subscribe to that specializes in analytics. Once filled in, we can upload it within SciVal and it goes to Elsevier's author team. By doing it that way, even complex merge requests are sent at once, saving time and effort.

Finally, after merge requests are sent, we still needed to determine which AUID is the chosen one going forward. At first, it was not clear without doing another manual search to check. However, we have since learned to reliably predict the outcome of all of the merges. An Elsevier representative explained that they pick the AUID with the latest publications and largest number of articles. However, if profiles are associated with the same number of publications, then they will usually pick the AUID with the most recent ones. Notably, this does not always hold true, as we have seen a few instances where the result did not go as anticipated. Hopefully, sharing this information will save time for our fellow librarians and researchers doing similar work.

Computational Framework for Large-Scale Disambiguation

To help with manual data searches and organization discussed above, our methodology attempts to automate the traditionally labor-intensive process of research assessment by utilizing data from heterogeneous sources, starting with a commercial bibliographic database (Scopus), and including institutional reporting tools (Symplectic Elements, Academic Analytics), and federal datasets (IPEDS). The development of this framework revealed significant challenges in data consistency and interoperability. There is a need for a suite of robust, fault-tolerant algorithms designed to handle the "messiness" inherent in real-world institutional data, and a call for improving the quality of data available from the sources.

Addressing Data Heterogeneity and Quality

A major obstacle in automating research assessment is that different data providers rarely use the same standards. During our collection, we frequently encountered "messy" data, including inconsistent file types, scrambled text characters, and missing information fields. To mitigate this, we developed a robust ingestion pipeline designed to be flexible rather than rigid. For example, when processing publication lists where metadata headers are malformed or missing, our system falls back to analyzing file naming conventions and directory structures to extract attribution data. This "defensive" approach ensures that valuable legacy data is not discarded due to formatting technicalities, which is a common barrier for institutions that have attempted to digitize their records.

Data Acquisition via Scopus API

To ensure comprehensive and scalable data processing, we utilized the Elsevier Scopus API rather than manual web exports. The API offers a wide variety of export options, and with batching, can extract a considerable amount of data without issue. While web-based interfaces are useful for small-scale queries, they are often ill-suited for large-scale bibliometric research due to rigid formatting and manual constraints. For example, at the time of data collection, the Scopus web export limited the number of authors displayed per paper to ten (a change quickly reversed by Scopus within weeks fortunately), which would have resulted in incomplete collaboration data for our analysis.

Automating this process at an institutional scale required a specialized approach to handle Scopus's strict traffic regulations and complex query requirements. To overcome these hurdles, we developed a Batched Query Engine designed to streamline the retrieval process:

a. Dynamic Batching: The system automatically calculates the most efficient number of AUIDs to include in a single request (usually 25–30). This maximizes the amount of data we receive at once without exceeding the technical character limits of the API web link.
b. Data Enrichment: Once the basic paper information was collected, the system performed a second "pass" to pull in deeper details, such as specific subject keywords and publication codes (CODEN). These extra layers of data allow for a much more granular analysis of research trends than a standard export would provide.

Algorithmic Resolution of Researcher IDs

The most significant barrier to accurate reporting is author ambiguity, for example, distinguishing between "J. Doe" (Engineering) and "J. Doe" (Arts). We found that relying solely on exact string matching was a weak approach. Our solution, instead, is a Repeating Canonicalization Engine that:

a. Generates Name Variants: Automatically produces all valid possible types of a researcher's name (e.g., "Doe, J.", "Doe, John", "J. Doe") to cast a wide search net.
b. Cross-References "Master" Lists: Validates potential matches against verified internal HR rosters and curated "Master" lists from Academic Analytics and other possible sources.
c. Contextual Disambiguation: Uses departmental affiliation and co-author networks as tie-breakers when multiple profiles exist.

Heuristic Alignment of Institutional Entities

Linking institutional names across international databases, such as matching a "Virginia Tech" record in Scopus to "Virginia Polytechnic Institute and State University" in IPEDS, is a significant challenge. Because these databases lack a shared universal ID system, and often treat other branches as entirely separate entities, simple searches frequently return incomplete or fragmented results.

To solve this problem, we developed a "Waterfall" Matching Algorithm. The system attempts to match names through three increasingly flexible stages, ensuring we only move to a broader search if a strict one fails:

a. Clean Exact Matching: We first standardized all names by removing punctuation and unifying common abbreviations (e.g., converting both "Univ." and "U." to "University"). If the names match perfectly after this cleaning, they are linked with high confidence.
b. Word-Set Match: To account for different naming conventions, we looked for matches based on the specific words used, regardless of their order. This allows the system to recognize that "Virginia Polytechnic Institute" and "Virginia Tech" refer to the same entity.

c. Validated Fuzzy Match: For names with slight spelling variations, we used a mathematical comparison (called *Levenshtein distance*) to find the closest match. However, to prevent "false positives," we built in semantic guardrails. These are rules that stop the system from matching institutions that look similar but are functionally different, for example, ensuring a "School of Law" is never accidentally merged with a "School of Medicine."

Although not a perfect solution, it reduced false positives when matching by over 50%, and is naturally an important step in eliminating manual data handling.

Implications for Global Data Standardization

Our work highlights that while automation can significantly reduce manual labor, it is currently limited by the quality of upstream data, and producing perfect outputs will always depend on this factor. The extensive "cleanup" logic required in our code is a call for the urgent need for global adherence to persistent identifiers (like ORCID and ROR, or Research Organization Registry) at the source level. Until databases and institutes around the globe universally adopt these standards, the heuristic strategies and fault-tolerant architectures we have developed remain essential for any institution seeking comprehensive research intelligence, and are a step towards both better automation of research data handling and improving data quality from our sources.

Scopus Author ID Identification and Validation Outcomes

Faculty Roster Results

The College of Engineering (COE) faculty roster at Virginia Tech contained 731 faculty records including all types (tenure/tenure-track, instructional, administrative, etc.). Automated identification using the Elsevier Scopus Author Search API classified outcomes based on identifier uniqueness. Records returning exactly one Scopus Author ID (AUID) were treated as confirmed high-confidence matches, while records returning zero or multiple candidate profiles were classified as unresolved and routed for manual review.

Using the API workflow alone, 448 faculty records (61%) returned exactly one AUID and were classified as confirmed. In contrast, 210 records (29%) returned no Scopus author profiles under the applied query constraints and were classified as *not found*, and 73 records (10%) returned multiple candidate AUIDs and were classified as *ambiguous*. Overall, 283 records (39%) were unresolved after API retrieval and therefore required manual follow-up (Table 1; Figure 2). These results demonstrate that automated querying can provide institution-scale identifier coverage for a majority of faculty while transparently isolating uncertain cases to prevent erroneous attribution.

Table 1. Scopus Author Search API outcomes for COE faculty roster (N = 731).

Identification outcome (API)	Count	Percent
Confirmed (1 AUID returned)	448	61%
Not found (0 AUID returned)	210	29%
Multiple profiles (>1 AUID returned)	73	10%
Total	**731**	**100%**

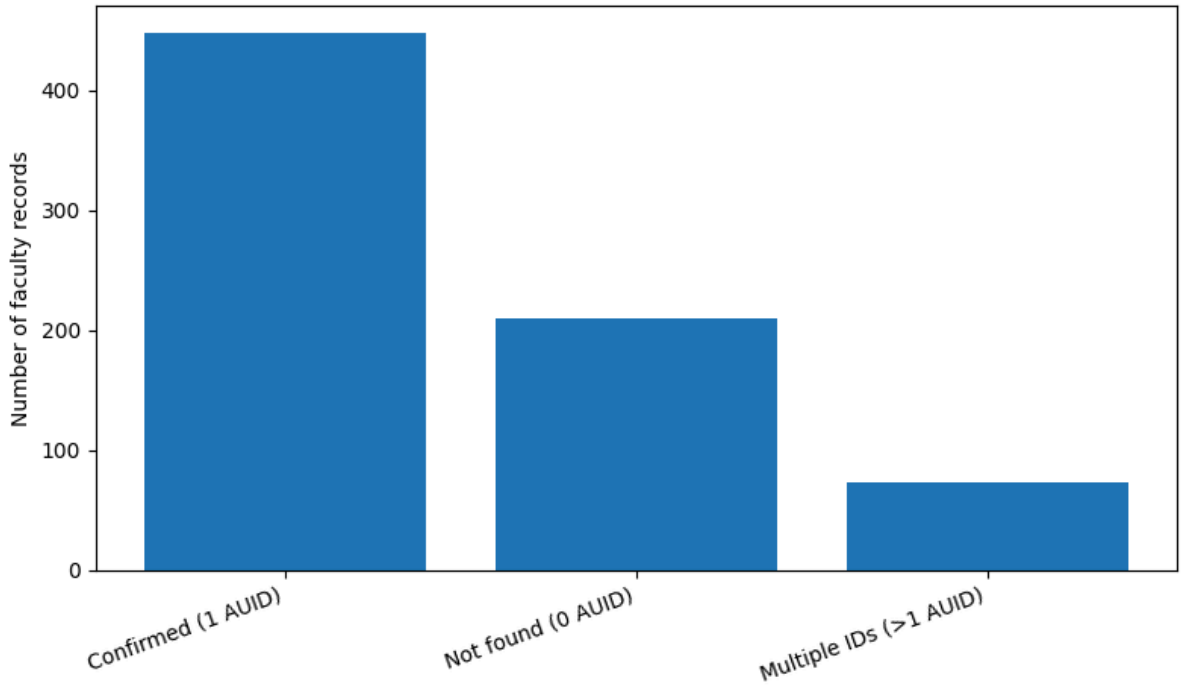

Figure 2: Scopus author search API outcomes (N=731). This bar chart represents graphically how many COE faculty had a high confidence match, no match, or multiple matches utilizing the API to find AUIDs.

Manual verification was subsequently conducted for all unresolved cases to either identify the correct AUID or confirm that no Scopus profile existed. Following manual investigation and reconciliation, the number of faculty with a confirmed unique AUID increased to 617 (84%), while 114 faculty records (16%) were confirmed to have no AUID (Table 2). In total, manual verification recovered 169 additional confirmed identifiers, raising confirmed coverage from 61% to 84% (Figure 3).

Table 2. Final Scopus Author ID coverage after manual verification (N = 731).

Final identifier status	Count	Percent
Confirmed unique Scopus Author ID (AUID)	617	84%
Confirmed none (no Scopus Author ID)	114	16%
Total	**731**	**100%**

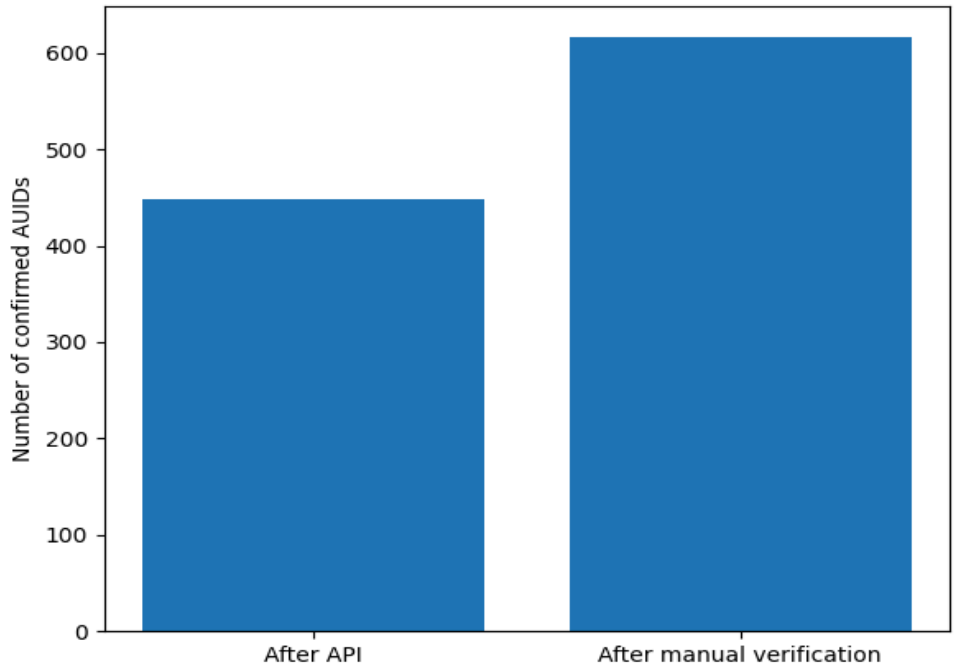

Figure 3: Confirmed Scopus author ID coverage improvement. After manual work, over 600 of the 731 total faculty had confirmed AUIDs.

Focusing specifically on the unresolved queue produced by the API workflow (N = 283), manual review resolved 169 cases (60%) to a confirmed unique AUID. The remaining 114 cases (40%) were verified as legitimately lacking a AUID (Table 3; Figure 4). Manual investigation revealed those without AUIDs were not in areas that required or focused on publishing, such as communications, assistants, research or program coordinators, information technology, business and operations directors, academic advisors, or Alumni relations. However, if any person's AUID has been incorrectly omitted, we anticipate making corrections as needed as we learn of them. These findings indicate that unresolved API results frequently reflect ambiguity or metadata limitations rather than true absence of Scopus coverage, and that structured human validation substantially improves institutional completeness without compromising identifier accuracy.

Table 3. Manual review outcomes for unresolved API cases (N = 283).

Manual resolution outcome	Count	Percent of API cases
Resolved to confirmed unique AUID	169	60%
Confirmed none (no Scopus Author ID)	114	40%
Total resolved	**283**	**100%**

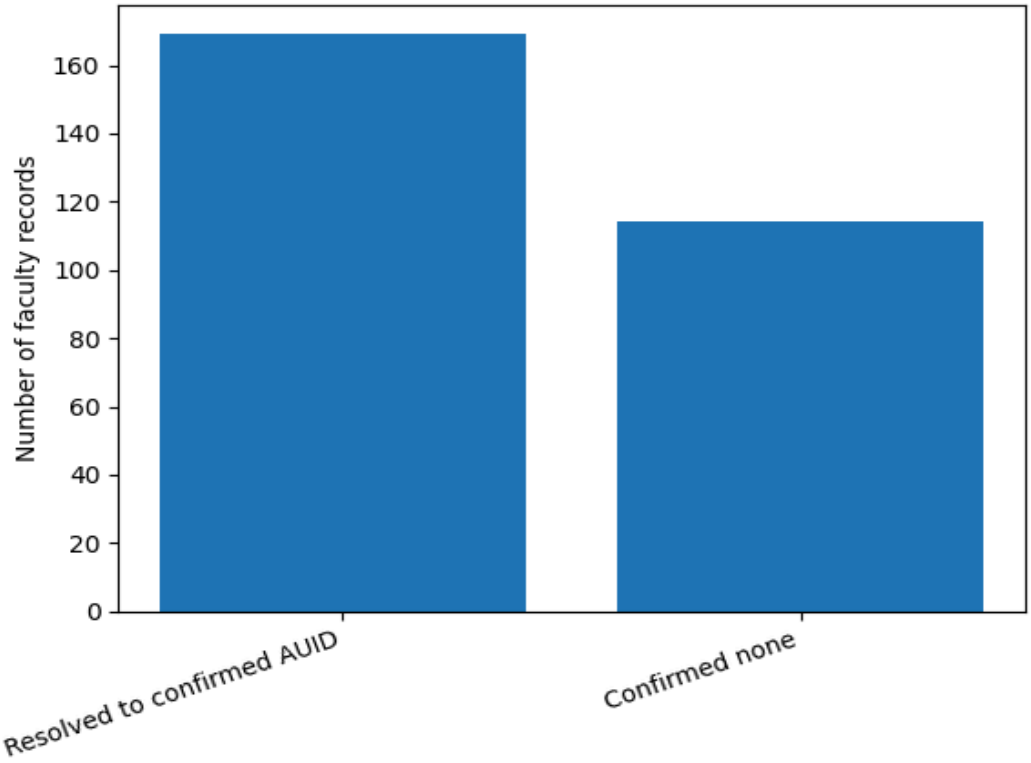

Figure 4: Manual review outcomes for unresolved cases (N=283). This bar chart represents graphically how many COE faculty were confirmed to have an AUID (over 160) or none (over 110) compared to the initial API work.

Discussion and Implications for University Priorities

These results highlight a deliberate trade-off between automation and data integrity in institutional-scale author identification. Conservative API-based matching successfully identifies the majority of faculty while minimizing false-positive assignments; however, it also produces a sizable unresolved queue due to name ambiguity, fragmented author profiles, and inconsistent affiliation metadata. Human-in-the-loop verification proved essential for improving completeness, recovering nearly 60% of unresolved cases and raising confirmed AUID cases to 84%, which aligns with findings from previous studies [5], [12].

For institutional bibliometric dashboards, this validated identifier set provides a practical foundation for automated metric retrieval and reporting. Confirmed AUIDs can be integrated into downstream data pipelines with high confidence, while unresolved and "confirmed none" cases should remain explicitly flagged. Maintaining these categories prevents inaccurate attribution, supports transparency around coverage limitations, and ensures that faculty-level metrics used for assessment and decision-making are derived only from verified identifiers.

These findings are conditioned by the completeness and consistency of Scopus metadata, including variability in author name representation and institutional affiliation reporting. API retrieval outcomes may therefore reflect external profile quality rather than workflow performance alone. In addition, manual verification is resource-intensive and introduces some dependence on reviewer judgment and the availability of external evidence. Finally, results reflect one disciplinary context (engineering faculty) and may not fully generalize to other colleges with different publication patterns and Scopus indexing coverage.

Metrics derived from these AUIDs have been formulated into a dashboard for university administrators such as department heads and shared with them recently. Once the data is reviewed by those who know their departments or units best, opportunities will arise for administrators to ask questions such as "why is J. Smith missing from my department's metrics?" or "what is an h5-index and why does C. Brown have such a high/low number?". We hope that these will be productive discussions that help encourage faculty or at least a designee in their department to better manage their data.

As we discussed above, AUIDs were selected as a starting point for this work out of the four sources (ORCID, Scopus, Academic Analytics, and Elements). As Academic Analytics is primarily curated not by us, but Academic Analytics' team and only includes research faculty, we have already begun initial steps towards work with ORCID and Elements data.

Conclusions

The Scopus Author ID work took our team close to a year to process the over 6,000 faculty at Virginia Tech, finishing data for the 2024-2025 academic year after it concluded. We have tentatively set a regular update interval with our Office of Research and Innovation to twice a year for updating the faculty roster (and monthly for metrics). We found from the Fall 2025 semester faculty data that there is approximately 10 percent turn-over between the old and new data. This is understandable as faculty come and go from institutions, especially in today's workplace where staying at a single institution for one's whole career is no longer expected.

As our College of Engineering is responsible for much of our grant funding and national rankings, how engineering faculty manage their profiles becomes of high importance for missions like Global Distinction. Unless our institution finds opportunities to enforce profile management, these will continue to be messy data. As of writing, we in the libraries are becoming even more essential to our Office of Research and Innovation due to this work – no one else on campus is better equipped to handle bibliographic and bibliometric data.

Although data from Scopus works quite well for engineering, there are limitations for other types of research output. This especially affects the humanities and performing arts for example who measure scholarly productivity differently. Plays, books, musical scores, and even research outputs like computer code are not well represented by our current data set. We hope to begin mitigating these limitations via other data sources. We also recognize that other data sources like Web of Science are of interest for rankings, however it is not used in current ranking methodology Virginia Tech prioritizes and was not included at this time.

As we enter the next phases of our project, our advice to other libraries who support engineering faculty is to consider how your institution is viewing ranking, metrics, and other indicators of academic "success." How much are your engineering faculty adding to your institution's rankings? Do more of them need to manage their profiles? Are your administrators making decisions based on incomplete data? What tools do you already have at hand? A great place to start can be with ORCID iDs as they are now required (or at least highly encouraged) for many applications from federal grants to journal articles. ORCID does have a public API as well, which can provide data on your institution's engineering faculty such as how many have a public ORCID iD when paired with a faculty roster list. Our process that we highlight in this paper can be applied to other IDs – considering name disambiguation, manual versus API searches, and much more.

In collaboration with the Division of Scholarly Integrity and Research Compliance, members of the Elements Implementation Team have started to roll out a new ORCID requirement for faculty who are actively publishing and/or applying for and receiving external grants [13]. The primary purpose of this project is to strengthen research security, protect Virginia Tech researchers by ensuring accurate attribution of scholarly work, and strengthen author name disambiguation.

The team has identified two primary groups for outreach:

- Faculty who are actively publishing but have not claimed or connected an ORCID iD in Elements
- Faculty who are actively submitting grant proposals or receiving awards but are not actively publishing and have not claimed an ORCID iD

With this targeted population identified, the team is now developing a coordinated strategy for communication, outreach, implementation, and compliance tracking over the coming months. Widespread ORCID adoption helps reduce the risk of misattribution, including instances in which researchers' names are improperly added to publications to meet authorship requirements despite no actual contribution.

Data & Code Availability:

A sample code package can be found here: https://github.com/shehryarkh4n/RII-HUB. This is intended to help other institutions do similar work as our full version is only usable for Virginia Tech.

References:

[1] "Higher Education Research and Development (HERD) Survey 2024 | NSF - National Science Foundation." Accessed: Jan. 20, 2026. [Online]. Available: https://ncses.nsf.gov/surveys/higher-education-research-development/2024

[2] "Global Distinction FAQs." Accessed: Apr. 23, 2026. [Online]. Available: https://www.vt.edu/global-distinction/faqs.html

[3] "The Best Colleges for Undergraduate Engineering," US News & World Report. Accessed: Apr. 27, 2026. [Online]. Available: https://www.usnews.com/best-colleges/rankings/engineering-doctorate

[4] "University Rankings Data: A Closer Look for Research Leaders," www.elsevier.com. Accessed: Apr. 24, 2026. [Online]. Available: https://www.elsevier.com/academic-and-government/the-university-rankings-data

[5] J. Fuhr and C. Monnin, "Researcher profile system adoption and use across discipline and rank: A case study at the University of Manitoba," *Quant. Sci. Stud.*, vol. 5, no. 3, pp. 573–592, Aug. 2024, doi: 10.1162/qss_a_00319.

[6] L. Hourlay, "Improving the visibility of the institution, researchers, and publications by introducing specific identifiers (PIDs)," *J. EAHIL*, vol. 19, no. 4, pp. 2–6, Dec. 2023, doi: 10.32384/jeahil19579.

[7] L. Zhang and C. Li, "Investigating Science Researchers' Presence on Academic Profile Websites: A Case Study of a Canadian Research University," *Issues Sci. Technol. Librariansh.*, no. 95, Sep. 2020, doi: 10.29173/istl51.

[8] V. Vasan *et al.*, "The Effect of Multiple Scopus Profiles on the Perceived Academic Productivity of Neurosurgeons in the United States," *World Neurosurg.*, vol. 171, pp. e500–e505, Mar. 2023, doi: 10.1016/j.wneu.2022.12.056.

[9] "Scopus content | Elsevier," www.elsevier.com. Accessed: Jan. 20, 2026. [Online]. Available: https://www.elsevier.com/products/scopus/content

[10] "Academic Analytics FAQ," Frequently Asked Questions: About Our Products. Accessed: Jan. 20, 2026. [Online]. Available: https://www.academicanalytics.com/frequently-asked-questions

[11] "The Elements Platform," Symplectic. Accessed: Jan. 20, 2026. [Online]. Available: https://www.symplectic.co.uk/products/symplectic-elements/
[12] V. Aman, "Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates," *Scientometrics*, vol. 117, no. 2, pp. 705–720, Nov. 2018, doi: 10.1007/s11192-018-2895-3.
[13] E. Mazure, "Virginia Tech ORCID Requirement for Faculty." Accessed: Apr. 24, 2026. [Online]. Available: https://guides.lib.vt.edu/orcid/home