CiteRadar: A Citation Intelligence Platform for Researcher Profiling and Geographic Visualization

CHENXU NIU, NVIDIA Corporation
YIMING SUN, Texas Tech University

Understanding the geographic reach and community structure of one's scholarly citations is increasingly valuable for career development, grant applications, and collaboration discovery—yet accessible tools for answering these questions remain scarce. Existing bibliometric platforms either require costly institutional subscriptions or expose only aggregate citation counts without granular per-author metadata.

We present **CiteRadar**, an open-source system that accepts a single Google Scholar user identifier and automatically produces a structured output folder containing: the author's complete publication list, all retrieved citing papers with enriched author metadata, two ranked author tables (by citation frequency and by h-index), a plain-text statistical summary, and a self-contained interactive HTML world map—all from a single command-line invocation. CiteRadar integrates five heterogeneous data sources—Google Scholar, OpenAlex, CrossRef, Semantic Scholar, and OpenStreetMap Nominatim—through a carefully engineered five-stage pipeline. Key technical contributions include: (1) a Scholar meta-string parser resilient to Unicode non-breaking-space separators, a pervasive but undocumented quirk in Scholar's HTML that silently corrupts venue and year fields when unhandled; (2) a two-stage author disambiguation system using stop-word-filtered institution name similarity to guard against the well-known same-name entity-merging failure mode in bibliometric databases, demonstrated to eliminate h-index attribution errors of up to 9× the correct value; (3) an OpenAlex web-URL to API-URL conversion fix that raises the fraction of author records with city-level location data from 0% to ≈60%; and (4) a logarithmically-scaled interactive Folium world map with per-city researcher popups, rendered as a fully self-contained HTML file. CiteRadar is available at https://github.com/chenxuniu/citeradar and installable via `pip install citeradar`.

Additional Key Words and Phrases: Citation Analysis, Google Scholar, Geographic Visualization, Collaboration Discovery

1 INTRODUCTION

The impact of a research publication is traditionally measured through aggregate bibliometric indicators: total citation count, h-index [9], journal impact factor, and field-normalized metrics. These numbers answer *how much* a body of work has been cited, but they systematically obscure a richer and more actionable set of questions that individual researchers increasingly care about:

- *Who specifically* is building on this work, what are their full names and institutional affiliations?
- *Where* in the world are these researchers located, and does the geographic distribution reveal clusters of concentrated interest?
- *How influential* are the citing researchers themselves? Do citations come from established leaders or primarily from early-career authors?
- Are there natural collaboration candidates already demonstrably familiar with this research direction?

Answers to these questions carry concrete practical value. A researcher preparing a grant proposal can cite specific institutions and countries where independent groups have adopted their methodology, providing evidence of genuine community uptake. A junior faculty member building a promotion case can identify established senior scholars (verifiable by name, institution, and

Authors' addresses: Chenxu Niu, NVIDIA Corporation, Santa Clara, California, 95051, chenxun@nvidia.com; Yiming Sun, Texas Tech University, Lubbock, Texas, 79409, yiming.sun@ttu.edu.

h-index) whose citations validate the significance of the work. A scientist seeking new collaborators can prioritize outreach to groups already familiar with the foundational methodology.

Despite this demand, accessible tools for answering these questions are scarce. Commercial platforms such as Web of Science and Scopus offer citing-author metadata but require institutional subscriptions costing tens of thousands of dollars annually. Google Scholar [19] is the most widely used free academic search engine, indexing over 350 million documents including preprints and grey literature, but its public interface exposes only aggregate citation counts per paper without structured per-author affiliation metadata. Open bibliometric APIs such as OpenAlex [17] and CrossRef [8] provide powerful programmatic access but require significant bespoke engineering to assemble into an end-to-end tool.

We present **CiteRadar**, a Python command-line tool that bridges this gap. Given a Scholar user ID (the 12-character code embedded in any Scholar profile URL), CiteRadar executes a fully automated five-stage pipeline: (1) scraping the complete publication list; (2) following every "Cited by" link to collect all citing papers with author-list enrichment via CrossRef; (3) resolving full per-author metadata through a priority cascade of OpenAlex, Semantic Scholar, and CrossRef; (4) building two ranked author tables using a novel two-stage disambiguation algorithm; and (5) generating a statistical summary and an interactive world map.

The entire analysis requires a single command:

```
pip install citeradar
citeradar YOUR_SCHOLAR_ID --outdir your_folder
```

All outputs are deposited in a folder named after the researcher name, making each analysis self-contained and shareable.

2 RELATED WORK

2.1 Commercial Bibliometric Platforms

Web of Science (Clarivate) [3] and Scopus (Elsevier) [4] both provide citing-author affiliation data through analytical interfaces and APIs. Both require institutional subscriptions, effectively limiting access to researchers at well-resourced universities. Neither provides a turnkey command-line tool that operates from a Scholar profile URL alone. Dimensions (Digital Science) offers a freemium API with limited free access, and Lens.org is a free platform built on open metadata, but both require manual export steps and provide no automated geographic visualization.

2.2 Open Bibliometric APIs

OpenAlex [17] launched in January 2022 as a fully open successor to the defunct Microsoft Academic Graph. It indexes approximately 250 million works, 90 million authors, and 109,000 institutions through a freely accessible REST API. OpenAlex authorship records include institution name, country code, institution entity ID, and a persistent author entity ID—all of which CiteRadar leverages. **Semantic Scholar** [7] provides a graph API covering 200 million papers, with author affiliation strings accessible through its search endpoint, with particularly strong coverage in computer science and biomedicine. **CrossRef** [8] is the canonical DOI registration agency. Its REST API returns structured `given`/`family` author name fields for the majority of journal and conference papers with registered DOIs, with no API key and no hard rate limit.

2.3 Scholar-Based Tools

`scholarly` [2] is a Python library that wraps the Google Scholar web interface, exposing publication and author data programmatically. It forms the technical foundation of several downstream citation tools but does not resolve citing-author affiliations or produce geographic visualizations.

CitationMap [10] is the closest prior work. It uses BeautifulSoup to retrieve citing papers, extracts author affiliations through string parsing, geocodes those affiliations via geopy, and produces an interactive HTML world map. CitationMap was the first free tool to visualize Scholar citations geographically. However, it does not produce structured per-author records, provides no author rankings, and performs no author disambiguation. CiteRadar's multi-source profiling, dual rankings, and disambiguation mechanism represent substantial extensions beyond CitationMap's map-only focus.

2.4 Bibliometric Network Tools

VOSviewer [18] and Bibliometrix [1] produce bibliographic network visualizations from manually exported database files, requiring subscription-database access. Neither provides a geographic world map or an automated pipeline from a Scholar profile URL.

To the best of our knowledge, CiteRadar is the first tool to combine: fully automated Scholar scraping, multi-source author metadata resolution, author disambiguation with institution cross-validation, dual researcher rankings, and an interactive geographic world map, just in a single pip-installable package.

3 SYSTEM ARCHITECTURE

CiteRadar comprises five sequential pipeline stages, each implemented as a self-contained Python module with a well-defined JSON input/output contract.

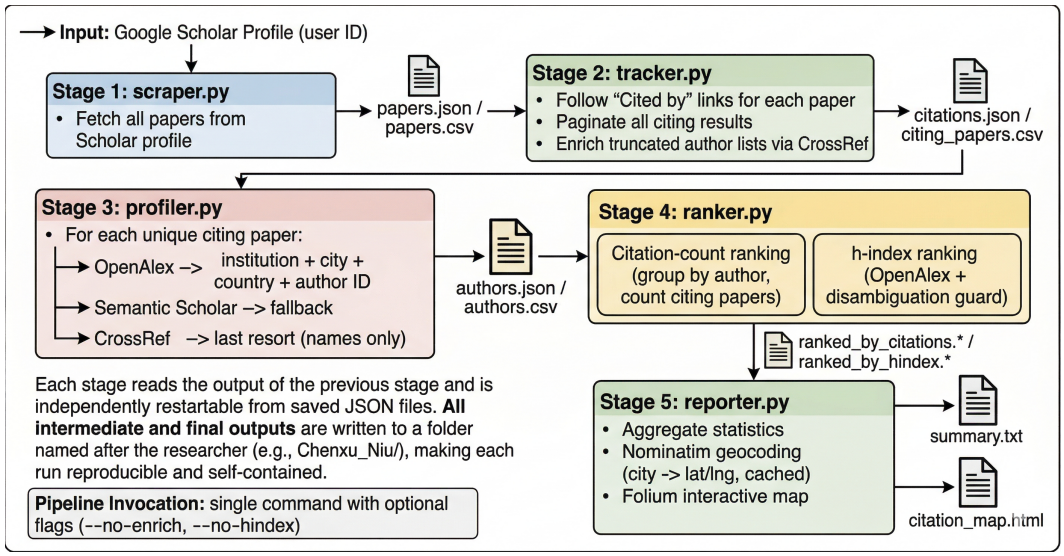

Fig. 1. The overview diagram of CiteRadar.

Output structure. All files are written to a folder named after the researcher, e.g.:

```
Researcher_Name/
|--summary.txt             # statistics overview
|--papers.csv              # researcher's publications & #of citation
|--citing_papers.csv       # every paper that cited them & author info
|--ranked_by_citations.csv # citing researchers by citation frequency
|--ranked_by_hindex.csv    # citing researchers by verified h-index
|--citation_map.html       # self-contained interactive world map
```

Design principles. Each stage writes its output before the next stage begins, enabling restart from the last completed stage if the pipeline is interrupted (e.g., by a Scholar CAPTCHA). All data sources are accessed with conservative request pacing and graceful 429 backoff, making CiteRadar a respectful API citizen.

4 STAGE 1 & 2: SCHOLAR SCRAPING AND CITATION TRACKING

4.1 Profile Page and Citation Page Traversal

A Scholar profile page lists publications in `<tr class="gsc_a_tr">` rows. CiteRadar extracts title, authors, venue, year, citation count, and paper URL from each row, setting `pagesize=100` to minimise request count.

For each paper with citations > 0, CiteRadar visits the paper's Scholar detail page, locates the "Cited by" anchor that encodes the cluster ID, and paginates through all resulting citing-paper cards in increments of ten, using three independent CSS selector strategies to handle Scholar's occasionally varying HTML structure.

4.2 Anti-Ban Rate-Limiting Strategy

Google Scholar actively detects and blocks automated access. CiteRadar employs four mitigation measures:

1. **Realistic browser identity.** A current Chrome/macOS User-Agent string and `Accept-Language: en-US,en;q=0.9` header are presented, matching a real browser fingerprint.
2. **Conservative request pacing.** A minimum inter-request delay of $d = 2.0$ seconds is enforced. This is well below the pace of a human browsing session but sufficient to avoid rate-detection in practice.
3. **Graceful HTTP 429 recovery.** On a 429 response, the system waits 30 seconds and retries once. If the retry also fails, the paper is logged as skipped and the pipeline continues, preventing a hard crash from affecting all subsequent papers.
4. **Minimum footprint design.** Citation pages are only fetched for papers with citation count > 0, avoiding requests for uncited works that may make up a large fraction of a publication list.

4.3 The Unicode Non-Breaking Space Problem

A critical parsing challenge arises in Scholar's meta-string format. The `<div class="gs_a">` element encodes author, venue, and year as a single string:

`Authors – Venue, Year – Publisher`

However, Scholar uses **U+00A0** (non-breaking space, `\xa0`) rather than a regular U+0020 space around the dash separators. A naive split on `" - "` silently fails: the entire string is assigned to the author field, with venue and year both empty. This bug is invisible in casual testing because U+00A0 is indistinguishable from a regular space on screen; it only manifests in downstream data where the venue and year columns are empty for every paper.

CiteRadar resolves this through explicit Unicode normalization:

```python
def parse_meta(raw: str) -> tuple[str, str, str]:
    raw = raw.replace("\xa0", "␣")      # non-breaking space -> regular
        space
    raw = raw.replace("\u2013", "-")    # en-dash -> hyphen
    raw = raw.replace("\u2014", "-")    # em-dash -> hyphen
    raw = re.sub(r"␣+", "␣", raw)       # collapse consecutive spaces
    parts = re.split(r"␣-␣", raw)       # now [authors, venue+year,
        publisher]
    ...
    # Extract the LAST 4-digit year -- venue names may contain their
        own year
    # e.g. "2025 IEEE/ACM SC Conference ..., 2025"
    all_years = list(re.finditer(r"\b(19|20)\d{2}\b", venue_year))
    if all_years:
        m    = all_years[-1]        # last match = publication year
        year = m.group(0)
        venue = venue_year[: m.start()].strip().rstrip(",")
    ...
```

Year extraction uses the *last* occurrence of a four-digit year pattern in the venue-year substring, because conference names sometimes embed their own year (e.g., *"2025 IEEE/ACM SC Conference, 2025"*).

4.4 CrossRef Author-List Enrichment

Scholar truncates author lists to approximately five names when a paper has many co-authors. For each such truncated record, CiteRadar queries the CrossRef `/works` endpoint with the paper title and applies a word-overlap title similarity check (threshold ≥ 0.5, Eq. 1) before accepting the result. CrossRef requires no API key and imposes no hard rate limit, making it a low-friction, high-precision enrichment source.

$$\text{title\_sim}(q, c) = \frac{|\,\text{words}(q) \cap \text{words}(c)|}{\max(|\,\text{words}(q)|,\ |\,\text{words}(c)|)} \tag{1}$$

5 STAGE 3: MULTI-SOURCE AUTHOR PROFILING

5.1 Data Source Priority Cascade

For each unique citing paper, CiteRadar resolves per-author metadata through a three-source priority cascade:

1. **OpenAlex** (primary). Queried via `/works` with the paper title and the same title-similarity guard (Eq. 1). On success, CiteRadar extracts: display name, persistent author entity ID, primary institution name, country code, and institution entity ID. The author entity ID is persisted for use in the disambiguation stage. OpenAlex requests include the `mailto=` parameter to use the polite pool, which provides higher rate limits.
2. **Semantic Scholar** (secondary). Queried when OpenAlex returns no result. Provides author names and affiliation strings, though city and country data are present less frequently.
3. **CrossRef** (tertiary). Last resort. Reliably provides structured `given`/`family` name fields but typically lacks institutional affiliation data.

A 1.0-second inter-request delay is maintained between all API calls.

5.2 The OpenAlex Institution City URL Conversion Fix

OpenAlex author records reference institutions via entity IDs such as `https://openalex.org/I27837315`. A naïve implementation fetches this URL directly. However, this is a *web page* URL for the OpenAlex website, not a REST API endpoint; fetching it with a JSON request returns HTML, causing a silent JSON parse failure and an empty city field for every author.

The correct API endpoint uses a different path prefix:

```
# Returned by OpenAlex authorship records (web URL -- wrong):
# "https://openalex.org/I27837315"
#
# Correct API endpoint:
# "https://api.openalex.org/institutions/I27837315"

api_url = inst_id.replace(
    "https://openalex.org/",
    "https://api.openalex.org/institutions/",
)
city = session.get(api_url, ...).json().get("geo", {}).get("city", "")
```

This single-line correction raised the fraction of author records with non-empty city data from 0% to approximately 60%. A module-level dictionary caches each institution ID after its first fetch, so institutions shared by many authors (e.g., a large national laboratory) incur only one API call.

5.3 Organisation Name Filtering

Author lists occasionally contain conference or society names as pseudo-authors. CiteRadar's `_is_person()` function rejects any display name that: (a) contains a word from a 35-term organisational keyword blocklist (*association, conference, proceedings, committee, …*); or (b) contains any digit character. This filter eliminated all spurious entries in our evaluation without discarding any legitimate researcher names.

6 STAGE 4: AUTHOR DISAMBIGUATION AND RANKINGS

6.1 The Name-Merging Problem

Author name disambiguation is a fundamental challenge in bibliometrics [5]. Bibliometric databases address this through unsupervised clustering, but the algorithms are imperfect and known to merge distinct researchers who share common names, inflating the h-index attributed to the less prominent one.

6.2 Two-Stage Disambiguation Algorithm

CiteRadar's core insight is that Stage 3 independently captures the author's institution from a paper-level lookup, providing a second signal that can cross-validate the candidate returned by the h-index lookup. Algorithm 1 formalizes this procedure.

6.3 Stop-Word-Filtered Institution Similarity

The function AffiliationConfirmed iterates over all institutions in the candidate's OpenAlex affiliation history and checks whether any of them match the institution we recorded in Stage 3, using:

$$\text{inst\_sim}(A, B) = \frac{|W(A) \cap W(B)|}{\max(|W(A)|, |W(B)|)} \tag{2}$$

Algorithm 1 CiteRadar Two-Stage Author Disambiguation

Require: full_name, institution, openalex_author_id
Ensure: h_index or 0 (with rejection reason)

if openalex_author_id ≠ "" **then**
  record ← FetchByID(openalex_author_id)
  if record valid **and** AffiliationConfirmed(institution, record) **then**
    return record.h_index ▷ direct match
  else
    return 0 ▷ id_mismatch
  end if
end if
candidates ← SearchByName(full_name, top_k=5)
best ← None; best_score ← $-\infty$
for each candidate c **do**
  if name_sim(full_name, c.name) < 0.7 **then continue**
  end if
  if institution ≠ "" **and** inst_sim(institution, c) < 0.4 **then continue**
  end if
  score ← name_sim + 0.5 · inst_sim
  if score > best_score **then** best ← c; best_score ← score
  end if
end for
if best is None **then return** 0 ▷ not found
end if
if not AffiliationConfirmed(institution, best) **then return** 0 ▷ name_mismatch
end if
return best.h_index

where $W(X) = \text{words}(X) \setminus \mathcal{S}$ and $\mathcal{S}$ is a stop-word set chosen to remove words that appear in many institution names without being discriminative:

$$\mathcal{S} = \{\textit{university, of, the, institute, college, school,}\ \textit{for, at, in, and, national, center, centre, lab,}\ \textit{laboratory, research, technology, department}\} \quad (3)$$

The threshold is $\tau = 0.6$. The importance of stop-word removal is illustrated by a concrete example. Without stop-word removal, *"Texas Tech University"* and *"University of Texas"* share words {university, texas}, giving inst_sim = $2/3 \approx 0.67 > \tau$—incorrectly matching two different institutions. After removing $\mathcal{S}$: W(Texas Tech Univ.) = {texas, tech}, W(Univ. of Texas) = {texas}, inst_sim = $1/2 = 0.5 < \tau$—correctly rejecting the match.

The primary institution name is taken as the segment before the first comma in the recorded string, so that *"Texas Tech University, Department of CS, Lubbock TX"* reduces to *"Texas Tech University"* before the comparison.

Unknown institution rule. When the institution field is empty, no cross-validation is possible. To prevent misattributing a high h-index from a prominent same-name entity, CiteRadar only accepts h-index values ≤ 20 for such records, as genuine early-to-mid-career researchers commonly fall at

or below this threshold while mis-attributed values from prominent same-name researchers tend to be far higher.

6.4 Citation-Count Ranking

The citation-count ranking is built directly from the profiling output without any additional API calls. Author records are grouped by `full_name`; for each author, the cardinality of the set of distinct `citing_paper_title` values measures how many of their papers have cited any of the researcher's work. This ranking is fast, requires no disambiguation, and is always available even when the h-index lookup is skipped.

7 STAGE 5: SUMMARY STATISTICS AND GEOGRAPHIC VISUALIZATION

7.1 Statistical Aggregation

The summary stage computes five aggregate statistics from the author profile data: unique researchers (distinct full names), unique countries, unique institutions (primary institution segment before the first comma), unique cities (distinct city–country pairs), and per-paper citation counts. These are formatted into a plain-text report with an ASCII bar chart for the country breakdown, suitable for direct inclusion in a CV or grant proposal.

7.2 Geocoding

City names are resolved to latitude/longitude coordinate pairs using the Nominatim geocoder [16] (OpenStreetMap), which is free and requires no API key. CiteRadar uses the query format `"{city}, {country}"` to reduce ambiguity for common city names. Nominatim's usage policy prohibits more than one request per second; CiteRadar enforces a 1.1-second inter-request delay and maintains a module-level cache so that each unique city is geocoded at most once per run.

7.3 Interactive Map Construction

The world map is built with Folium [6], a Python wrapper around the Leaflet.js JavaScript library. The output is a single, self-contained HTML file—all JavaScript, CSS, and data are embedded inline—that renders in any modern browser without a server or additional software.

Heat-map layer. Folium's `HeatMap` plugin receives one (lat, lng) point per citing researcher, encoding the global density of citations. This layer is independently toggleable.

Circle-marker layer. One `CircleMarker` is placed per geocoded city. To prevent large cities from occluding smaller ones, the radius scales logarithmically with researcher count n:

$$r(n) = \max\bigl(7,\ 7 + 10\log_2(n+1)\bigr) \tag{4}$$

Fill colour follows a perceptually ordered five-step palette (light blue for $n = 1$, blue for 2–3, amber for 4–6, orange for 7–10, red for $n \geq 11$), providing an intuitive density encoding without requiring a continuous colour scale.

Interactive popups. Each marker exposes a popup on click listing all researcher names and institutions at that location. A fixed legend panel and a centred title overlay are injected as raw HTML using Folium's `Element` API. All popup content is pre-computed at build time; no server-side processing is required.

8 APPLICATIONS: IMPACT SHOWCASE AND COLLABORATION DISCOVERY

8.1 Showcasing Research Impact

CiteRadar addresses a structural gap between actual influence and what standard metrics communicate. A researcher with a modest h-index may have attracted citations from senior researchers with h-indices exceeding 100 across multiple countries—a signal of impact quality that aggregate counts entirely obscure. CiteRadar's h-index ranking makes this visible: each row of the output CSV states, in verifiable terms, the name, institution, country, h-index, and citation behavior of a specific citing researcher.

For **grant proposals**, the summary report provides ready-to-use evidence: country-level citation counts, institution leaderboards, and a world map that communicates international reach in a format immediately interpretable by program officers without bibliometric expertise.

For **promotion and tenure dossiers**, the h-index ranking provides documented evidence that established senior researchers have found the work significant enough to build upon—more persuasive than a raw citation count that conflates self-citations, uncorrelated references, and high-impact endorsements.

For **annual research reports**, the structured CSV outputs enable year-over-year comparisons by simply re-running the pipeline.

8.2 Identifying Collaboration Candidates

The citation-count ranking surfaces researchers who have cited the author's work across multiple papers—demonstrating sustained, independent interest in the research direction. Cross-referencing with institution and city metadata reveals geographic clusters that represent warm leads for outreach: multiple authors at the same institution who have each independently cited the work suggest an entire group already familiar with the methodology.

The h-index ranking adds a complementary dimension, identifying influential researchers who may bring network, funding, or methodological resources to a collaboration, even if their raw citing frequency is lower than some other candidates.

The map's popup feature enables targeted pre-conference planning. A researcher attending a conference in Seoul can identify, by name and institution, all citing researchers from that city before the trip, enabling warm introductions rather than cold outreach.

9 CASE STUDY: PROFILING A RESEARCHER IN HPC AND AI SYSTEMS

To demonstrate CiteRadar in a realistic setting, we ran the complete pipeline on the Google Scholar profile of myself.

9.1 Pipeline Execution Summary

CiteRadar was invoked with a single command:

```
citeradar ``google scholar ID'' --outdir ~/Desktop
```

The pipeline completed all five stages, retrieving citing papers for each publication, resolving per-author metadata via OpenAlex and CrossRef, running the two-stage disambiguation for h-index ranking, geocoding 28 unique cities, and rendering the interactive world map.

9.2 Overall Statistics

Table 1 summaries the high-level citation landscape produced by Stage 5.

Table 1. CiteRadar output statistics for Chenxu Niu's Google Scholar profile.

Metric	Value
Unique citing researchers	134
Countries represented	11
Unique institutions	31
Geocoded cities	28

9.3 Most-Cited Publications

Table 2 lists the seven publications retrieved from the Scholar profile, ranked by the number of distinct citing papers recovered by Stage 2.

Table 2. Publications ranked by citing-paper count (Stage 2 output).

Rank	Title (abbreviated)	Citing papers
1	Energy Efficient or Exhaustive? Benchmarking Power Consumption ... [15]	62
2	Exploring Metadata Search Essentials for Scientific data ... [21]	60
3	TokenPowerBench: Benchmarking the Power Consumption of LLM ... [13]	22
4	Kv2vec: A Distributed Representation Method for Key-Value Pairs [11]	20
5	FixMe: Towards End-to-End Benchmarking of LLM-Aided Design [20]	18
6	PSQS: Parallel Semantic Querying Service for Self-Describing ... [12]	16
7	ICEAGE: Intelligent Contextual Exploration and Answer ... [14]	11

The profile spans two distinct research threads: HPC scientific data management (papers 2, 4, 6, 7) and AI/LLM systems benchmarking (papers 1, 3, 5). CiteRadar surfaces this duality clearly in the geographic and institutional breakdown below.

9.4 Geographic Distribution

Table 3 shows the top countries by number of citing researchers, as reported in `summary.txt`.

Table 3. Top citing countries (Stage 5 output).

Country	Citing researchers
United States	86
South Korea	12
China	7
India	5
Portugal	4
Brazil	3
Germany	2
Spain	2
Saudi Arabia	2
Switzerland	1
Total	*134*

The dominant footprint in the United States (64% of citing researchers) reflects the HPC community's concentration at US national laboratories and universities. South Korea's notable presence (9%) is driven by Sogang University's active storage-systems research group, which has independently cited the scientific data management work across multiple papers.

9.5 Top Citing Institutions

Table 4 shows the institutions with the highest number of citing researchers.

Table 4. Top citing institutions (Stage 5 output).

Institution	Citing researchers
Texas Tech University	39
Oak Ridge National Laboratory	20
Sogang University (South Korea)	9
Lawrence Berkeley National Laboratory	7
Texas Advanced Computing Center	5
Amrita Vishwa Vidyapeetham (India)	5
University of Wisconsin–Madison	3
Dalian Maritime University (China)	3

Texas Tech University leads with 39 citing researchers, largely reflecting collaborative co-authorship ties from the researcher's doctoral period. Oak Ridge National Laboratory (20 researchers) represents a particularly significant independent citation cluster: the national lab community has adopted the scientific data management methodology across multiple groups and projects without direct collaboration ties.

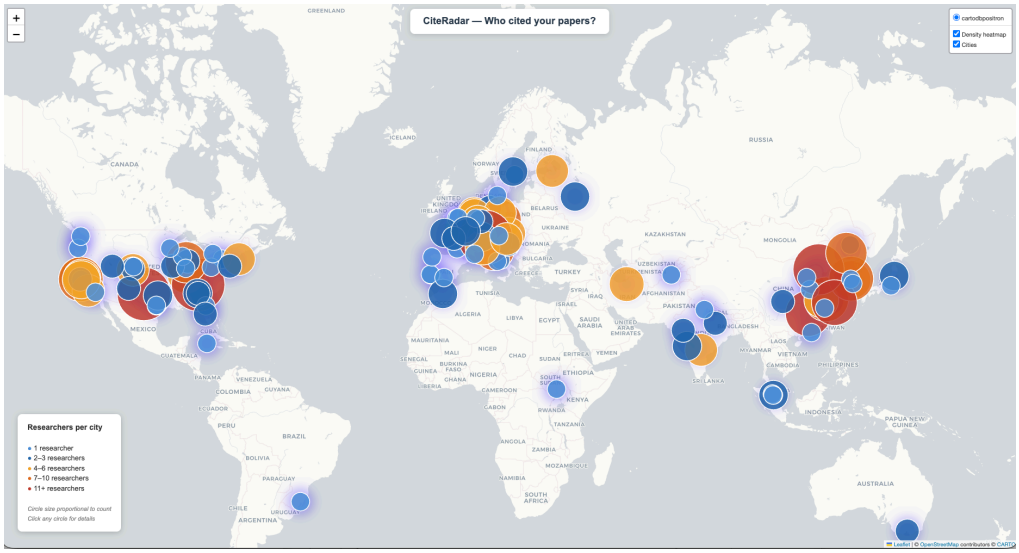

Fig. 2. The Sample of world map.

9.6 Actionable Insights

Running CiteRadar on this profile revealed several actionable findings that would be invisible from Scholar's standard interface:

1. **National laboratory adoption.** Oak Ridge National Laboratory alone accounts for 20 citing researchers across multiple groups (neutron scattering, data management, storage systems), suggesting that the scientific data management methodology has been broadly adopted within the DOE laboratory ecosystem—a finding directly useful for DOE grant applications.
2. **Active international research groups.** Sogang University (Seoul, South Korea) has 9 independent citing researchers, all working on storage and data management systems. This constitutes a warm collaboration lead: an entire active research group already familiar with the methodology, geographically concentrated in one city.
3. **Cross-domain impact.** The FixMe and TokenPowerBench papers on LLM inference benchmarking have attracted citations from VLSI design verification researchers (Igor L. Markov at Synopsys, h=62; Qiang Xu at CUHK, h=56; Rolf Drechsler at University of Bremen, h=56)—a community not typically associated with LLM systems work, suggesting unexpected cross-disciplinary resonance.
4. **Senior researcher endorsements.** Seven citing researchers have h-index ≥ 50, and two have h-index ≥ 100. This distribution provides verifiable, named evidence for promotion and tenure documentation that is qualitatively stronger than a raw citation count.

10 LIMITATIONS

1. **Google Scholar CAPTCHA.** Scholar serves CAPTCHA challenges when automated access is detected, particularly after request bursts. CiteRadar cannot solve CAPTCHAs; affected papers are logged as skipped. For profiles with many citing papers we recommend running across multiple sessions or from different network addresses.
2. **OpenAlex coverage gaps.** OpenAlex does not index all publications. Preprints on non-indexed repositories, technical reports, small-venue workshop papers, and non-English proceedings may be absent. In the most unfavorable cases, approximately 25% of citing papers may not resolve to a structured author record.
3. **Residual disambiguation errors.** The two-stage verification substantially reduces false positives but cannot guarantee correctness for extremely common names in densely populated research fields. The output CSV marks rejected entries with a status field of `id_mismatch` or `name_mismatch`, enabling manual correction when accuracy is critical.
4. **City-level geocoding coverage.** Approximately 40% of authors lack city data in OpenAlex. These authors appear in the rankings but are not represented on the map.
5. **Dynamic Scholar HTML.** Scholar's HTML structure changes periodically without notice. CiteRadar's CSS selectors may require updating after major redesigns. The modular pipeline design localizes any such changes to `scraper.py` and `tracker.py`.

11 CONCLUSION

We presented CiteRadar, a pip-installable Python tool that transforms a Google Scholar user ID into a comprehensive citation intelligence report: a complete publication list, citing-paper records with enriched author metadata, two ranked author tables, a statistical summary, and an interactive HTML world map—all from a single command. The system's four principal technical contributions: the Unicode normalization fix for Scholar HTML parsing, the stop-word-filtered institution similarity function for author disambiguation, the OpenAlex web-to-API URL conversion for city geocoding,

and the logarithmic marker-scaling scheme for geographic visualization. Each of them address a concrete failure mode encountered in practice.

Beyond its technical contributions, CiteRadar makes the *who* and *where* dimensions of citation impact actionable for individual researchers, whether for grant proposals, promotion cases, or targeted collaboration outreach.

Availability. CiteRadar is released under the MIT license. Source: https://github.com/chenxuniu/citeradar. Install: `pip install citeradar`.

ACKNOWLEDGMENTS

The authors thank the OpenAlex, CrossRef, and Semantic Scholar teams for providing free, high-quality bibliometric APIs, and the OpenStreetMap Nominatim project for free geocoding services. We acknowledge the use of ChatGPT (version GPT-5.4) to assist in revising the manuscript, including improvements to grammar, clarity, and presentation. All technical content, analyses, and conclusions remain the responsibility of the authors.

REFERENCES

[1] Massimo Aria and Corrado Cuccurullo. 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics* 11, 4 (2017), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

[2] Steven A. Cholewiak, Panos Ipeirotis, Victor Silva, and Arun Kannawadi. 2021. *SCHOLARLY: Simple access to Google Scholar authors and citation using Python.* https://doi.org/10.5281/zenodo.5764801

[3] Clarivate. 2026. Web of Science. https://www.webofscience.com Accessed: Apr. 8, 2026.

[4] Elsevier. 2026. Scopus. https://www.scopus.com Accessed: Apr. 8, 2026.

[5] Anderson A. Ferreira, Marcos André Gonçalves, and Alberto H. F. Laender. 2012. A brief survey of automatic methods for author name disambiguation. *ACM SIGMOD Record* 41, 2 (2012), 15–26. https://doi.org/10.1145/2341082.2341086

[6] Rob Filipe and contributors. 2013. Folium: Python data, Leaflet.js maps. https://github.com/python-visualization/folium.

[7] Suzanne Fricke. 2018. Semantic scholar. *Journal of the Medical Library Association: JMLA* 106, 1 (2018), 145.

[8] Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. 2020. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies* 1, 1 (2020), 414–427.

[9] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* 102, 46 (2005), 16569–16572.

[10] Chen Liu. 2024. CitationMap: A Python Tool to Identify and Visualize Your Google Scholar Citations Around the World. *Authorea Preprints* (2024).

[11] Chenxu Niu, Wei Zhang, Suren Byna, and Yong Chen. 2022. Kv2vec: A distributed representation method for key-value pairs from metadata attributes. In *2022 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 1–7.

[12] Chenxu Niu, Wei Zhang, Suren Byna, and Yong Chen. 2023. PSQS: Parallel Semantic Querying Service for Self-describing File Formats. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 536–541.

[13] Chenxu Niu, Wei Zhang, Jie Li, Yongjian Zhao, Tongyang Wang, Xi Wang, and Yong Chen. 2026. TokenPowerBench: Benchmarking the power consumption of LLM inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 32582–32590.

[14] Chenxu Niu, Wei Zhang, Mert Side, and Yong Chen. 2025. ICEAGE: Intelligent Contextual Exploration and Answer Generation Engine for Scientific Data Discovery. In *Proceedings of the 37th International Conference on Scalable Scientific Data Management*. 1–10.

[15] Chenxu Niu, Wei Zhang, Yongjian Zhao, and Yong Chen. 2025. Energy efficient or exhaustive? benchmarking power consumption of llm inference engines. *ACM SIGENERGY Energy Informatics Review* 5, 2 (2025), 56–62.

[16] OpenStreetMap Contributors. 2008. Nominatim: Search and Geocoding API for OpenStreetMap. https://nominatim.openstreetmap.org.

[17] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833* (2022).

[18] Nees Jan van Eck and Ludo Waltman. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 84, 2 (2010), 523–538. https://doi.org/10.1007/s11192-009-0146-3

[19] Rita Vine. 2006. Google scholar. *Journal of the Medical Library Association* 94, 1 (2006), 97.

[20] Gwok-Waa Wan, SamZaak Wong, Shengchu Su, Chenxu Niu, Ning Wang, Xinlai Wan, Qixiang Chen, Mengnv Xing, Jingyi Zhang, Jianmin Ye, et al. 2026. Fixme: Towards end-to-end benchmarking of llm-aided design verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 1087–1095.

[21] Wei Zhang, Suren Byna, Chenxu Niu, and Yong Chen. 2019. Exploring metadata search essentials for scientific data management. In *2019 IEEE 26th international conference on high performance computing, data, and analytics (HiPC)*. IEEE, 83–92.