IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

OpenAlex vs Semantic Scholar vs PubMed: Database Comparison

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-03-09 · openalex · semantic scholar · pubmed · academic databases · literature review · scholarly search engines · research tools · bibliometric analysis

Original article: https://intuitionlabs.ai/articles/openalex-semantic-scholar-pubmed-comparison



In this article

1. Executive Summary
2. Introduction and Background
3. Development History
4. Detailed Comparison
5. Conclusion

Executive Summary

In this comprehensive report, we analyze three major academic literature sources—**OpenAlex**, **Semantic Scholar**, and **PubMed**—with the goal of guiding researchers, librarians, and institutions in choosing the most appropriate resource for their needs. Each platform has distinct features, coverage, and strengths. **PubMed**, developed by the U.S. National Library of Medicine, focuses on biomedical and life sciences literature and provides a mature, curated search service with standardized Medical Subject Headings (MeSH) for precise retrieval (pubmed.ncbi.nlm.nih.gov) (academic.oup.com). **Semantic Scholar**, launched in 2015 by the Allen Institute for AI, is a broad, AI-enhanced search engine indexing research across science and technology domains; it boasts advanced features such as AI-generated summaries and citation-based rankings, and "indexes over 200 million academic papers" from diverse publishers (www.semanticscholar.org) (link.springer.com). **OpenAlex**, introduced in 2022 as the open successor to the Microsoft Academic Graph, is a fully **open-access** scholarly index covering *all* fields (sciences, social sciences, arts, humanities) with over 250 million works from 250,000 sources (link.springer.com). It provides data dumps and a hosted API built on persistent identifiers (DOI, ORCID, etc.) to facilitate large-scale analyses; API accounts receive $1/day free, with paid usage beyond that (OpenAlex pricing) (link.springer.com) (www.sciencedirect.com).

Key findings include:

- **Coverage:** PubMed contains ~40 million citations in biomedicine (pubmed.ncbi.nlm.nih.gov), Semantic Scholar ~200+ million across all STEM fields (www.semanticscholar.org) (link.springer.com), and OpenAlex ~250 million multidisciplinary works (link.springer.com). OpenAlex explicitly supports 256 fields of research (e.g. "Science, social science, arts, humanities") (wiki2.org), whereas PubMed is limited to health, biomedical, and related areas.
- **Features & Search:** PubMed offers powerful domain-specific tools (e.g. MeSH-based query mapping, clinical filters, very fast indexing of new papers) (academic.oup.com). Semantic Scholar provides modern AI-driven features (TL;DR summaries, visual influence graphs, and recommendation feeds (www.semanticscholar.org) (link.springer.com)). As checked on October 18, 2026, OpenAlex offers embedding-based semantic search through its API (OpenAlex) and model-based reranking, enabled by default on openalex.org (search documentation).
- **Data Quality & Openness:** OpenAlex promotes open science and equity (for instance, Sorbonne University replaced costly proprietary databases with OpenAlex (www.sciencedirect.com)). However, reviewers note metadata gaps (e.g. incomplete affiliation and language fields) that require cross-validation (www.sciencedirect.com). The current S2ORC dataset is released under ODC-By 1.0 (AI2). PubMed's content comes from authoritative MEDLINE indexing and repository feeds; its data are freely downloadable via NLM's FTP (no license required for use (support.nlm.nih.gov)), though it is not an "open data" offering like OpenAlex.
- **Use Cases:** For **biomedical research and clinical questions**, PubMed is usually the first choice, since it has rich domain indexing (MeSH) and filters for evidence-based medicine. Researchers in other fields (e.g. computer science, engineering) may prefer Semantic Scholar or OpenAlex for broader coverage. For bibliometric analyses and open-access infrastructure projects, OpenAlex is highly valued (e.g. the

Leiden Ranking Open Edition uses OpenAlex data (CWTS)). Several studies suggest *combining sources* yields the best results: for example, OpenAlex and Semantic Scholar have largely overlapping but complementary coverage of many references (subugoe.github.io) (www.sciencedirect.com), and experts advise using multiple databases to ensure complete literature surveys (www.sciencedirect.com).

Throughout this report, we juxtapose case studies and published analyses: e.g. an analysis of 37.5 million articles found OpenAlex and Semantic Scholar each cover roughly one billion cited references with significant overlap (subugoe.github.io); reviews of search tools have shown that out of 28 academic search engines compared, only 14 met all basic search requirements (link.springer.com). The comparisons below are richly documented with the latest data and expert observations. We conclude that **no single source is universally "best"**. Instead, the choice of OpenAlex, Semantic Scholar, PubMed (or any combination thereof) should be guided by the user's subject area, query needs, and whether open data access, AI-driven search, or field-specific indexing is most important.

Introduction and Background

The exponential growth of scholarly publications in recent decades has made effective search tools essential for research and discovery (academic.oup.com) (www.sciencedirect.com). Historically, researchers relied on domain-specific indexes such as MEDLINE/PubMed for biology and medicine, and on general libraries or catalogs for other fields. In the digital age, a proliferation of academic search engines has emerged: general-purpose tools like Google Scholar, as well as specialized services. **PubMed** (launched online in 1996) is the longstanding free portal to biomedical literature maintained by the U.S. National Library of Medicine (pubmed.ncbi.nlm.nih.gov) (academic.oup.com). Semantic Scholar (deployed for public use in late 2015) was developed by the Allen Institute for AI to leverage machine learning in navigating the scientific literature (www.semanticscholar.org) (www.nature.com). OpenAlex (initially announced in 2022) is the successor to the Microsoft Academic Graph, spearheaded by OurResearch, offering a fully open catalog of papers, authors, venues, and related entities (link.springer.com).

These platforms were built in different contexts. PubMed grew out of MEDLINE, focusing on carefully curated biomedical records with controlled vocabularies (MeSH) for indexers (academic.oup.com). It has become the "primary tool" for biomedical researchers and clinicians, indexing over 5,000 health science journals (back to 1948) (academic.oup.com). In contrast, both Semantic Scholar and OpenAlex aim to cover *all* fields of research. Semantic Scholar's mission is "accelerating science using AI" (www.semanticscholar.org): it crawls open-access papers and has partnerships with publishers, using NLP to extract key concepts, citations, and even outline article content. OpenAlex, by design, ingests metadata from Crossref, Microsoft Academic Graph archives, arXiv, PubMed Central, and many institutional repositories, unifying them via DOIs and persistent identifiers (www.sciencedirect.com) (link.springer.com). Unlike trends in proprietary databases, OpenAlex and Semantic Scholar push openness: both offer large portions of their data under permissive licenses.

Figure 1 (below) conceptually illustrates the relative scopes of these systems:

- **PubMed**: 40+ million records (primarily abstracts, some links to full texts) in biomedicine (pubmed.ncbi.nlm.nih.gov). MEDLINE MeSH indexing is automated, with human review of selected citations (NLM); new articles are added rapidly (over 1 million per year (www.sciencedirect.com)).

- **Semantic Scholar**: 200+ million articles in science and tech (www.semanticscholar.org). Strong in computer science, engineering, physics, and a growing inclusion of life sciences, as illustrated by its recent biotech papers expansion (www.semanticscholar.org) (link.springer.com). Integrates citation contexts and AI analysis.
- **OpenAlex**: 250+ million works spanning all disciplines (link.springer.com). Includes not only articles and conference papers, but also books, theses, and datasets. Emphasizes global coverage, including literature from the Global South (www.sciencedirect.com) and smaller open-access venues (the "diamond" OA journals) (www.sciencedirect.com).

Development History

- **PubMed / MEDLINE**: PubMed started as the on-line interface to MEDLARS (a 1960s NLM initiative) and went public in 1996 (prior decades of MEDLINE print/CD coverage). It has evolved continually: for instance, in 2020 PubMed shifted its default sort to relevance, and integrated more automated indexing (www.sciencedirect.com). It remains an NLM product, free to the public. As of 2025, PubMed's **About** page states it "contains more than 40 million citations and abstracts" from biomedical and life sciences journals (pubmed.ncbi.nlm.nih.gov). Early research (Lu 2011) noted PubMed had already hit ~20 million citations by 2010 (academic.oup.com), reflecting its steady growth (roughly 4% annually (academic.oup.com)). PubMed's infrastructure includes robust APIs (Entrez E-utilities) and bulk FTP databases (annual baseline releases) (support.nlm.nih.gov).
- **Semantic Scholar**: Founded in the Allen Institute for AI (AI2) around 2015 with seed funding from Microsoft co-founder Paul Allen. Its goal was to apply AI and knowledge graphs to scientific papers. In late 2015, AI2 announced a new free "science search engine" in *Nature* (www.nature.com), and by 2017 Semantic Scholar had incorporated multiple fields including biomedical texts (techcrunch.com). Over the years it has expanded via publisher partnerships; in 2019 it indexed ~175 million papers (techcrunch.com). Semantic Scholar has also released data: e.g. the S2ORC open corpus in 2019 with 81.1 million English papers (link.springer.com), and an API for fetching citation graphs. The semantic scholar website still reports an index on "over 200 million academic papers" (www.semanticscholar.org).
- **OpenAlex**: In January 2022, OurResearch (the nonprofit behind Unpaywall) launched OpenAlex as a completely open successor to Microsoft's Academic Graph (MAG), which was sunset at the end of 2021 (link.springer.com). The name evokes the Library of Alexandria as a symbol of comprehensive knowledge. The initial collection was already on the order of 200–250 million linked records (link.springer.com). OpenAlex is funded by philanthropic grants (e.g. the Sloan and Arcadia funds) and governance by non-profit. In December 2023, Sorbonne University announced that it would discontinue its subscription to Web of Science and Clarivate's bibliometric tools in 2024 (Sorbonne University). By mid-2025, independent reviews note that OpenAlex has "established itself as a strategic open-access infrastructure in bibliometrics" (www.sciencedirect.com), although it remains under development (metadata completeness and functionality are actively researched).

Thus, each source reflects a different generation of scholarly search technology: PubMed embodies the traditional curated biomedical index; Semantic Scholar represents the modern AI-driven, full-text aware search engine; and OpenAlex represents the "open science" approach, aiming to provide a free, multi-disciplinary

knowledge graph. The table below outlines key specifications of each:

Feature	PubMed	Semantic Scholar	OpenAlex
Launch / Owner	1996 (NLM/NIH) (pubmed.ncbi.nlm.nih.gov)	2015 (Allen Institute for AI) (www.semanticscholar.org) (www.nature.com)	2022 (OurResearch nonprofit) (link.springer.com)
Primary Scope	Biomedical & health (life sciences) (academic.oup.com)	Multidisciplinary (STEM-heavy) (www.semanticscholar.org)	General scholarly works (all fields, e.g. science, social science, humanities) (wiki2.org)
Total Records	~40 million citations/abstracts (pubmed.ncbi.nlm.nih.gov)	~200+ million papers (www.semanticscholar.org)	~250+ million works (link.springer.com)
Data Types	Journal articles, books, some preprints (PMC) (pubmed.ncbi.nlm.nih.gov)	Journal articles, conference papers; includes preprints (arXiv)	Articles, books, theses, datasets, etc. (works with DOIs)
Coverage by Discipline	5,000+ biomedical journals (academic.oup.com)	Thousands of journals across science/tech (includes last yrs' med sci)	Tens of thousands of sources (250K+ sources in 4 levels taxonomy) (link.springer.com) (developers.openalex.org)
Indexing & Metadata	MEDLINE MeSH indexing is automated, with selective human quality review (NLM)	Automated text-processing; builds citation graphs, concept tags	Aggregates Crossref/MAG/PubMedCentral data; uses open IDs (DOI, ORCID, ROR) for disambiguation (www.sciencedirect.com)
Search Interface	PubMed web UI (basic text, Advanced Search builder with MeSH) (academic.oup.com); Entrez API; FTP data (support.nlm.nih.gov)	Web UI with AI features (semantic "influential citations", sub-topic TLDRs, personalized feeds) (www.semanticscholar.org) (link.springer.com); REST/GraphQL APIs; data dumps (S2ORC)	Website search with filters and reranked results (OpenAlex); API semantic search (documentation)
Ranking & Relevance	Best Match uses BM25 followed by LambdaMART; deployed in June 2017 (NCBI developers)	Uses AI/NLP to identify key contributions, "highly influential" metrics, citations context	Search results default to descending relevance_score (OpenAlex)
Open Access / License	Free to use; data downloadable without license (support.nlm.nih.gov) (NCBI may still restrict some use)	Current S2ORC dataset under ODC-By 1.0 (S2ORC documentation); reuse of API data must comply with applicable dataset licenses and underlying third-party content licenses (API license agreement).	CC0 metadata; free snapshot downloads; metered API with a $1/day account allowance (OpenAlex, pricing)
Typical Use Cases	Domain-specific lit search (clinical queries, EBM); PMID linking; systematic reviews in medicine (www.sciencedirect.com)	Broad academic discovery; interdisciplinary research; quick paper overviews (TL;DR); citation impact analysis (www.semanticscholar.org) (link.springer.com)	Bibliometric research; network analysis; meta-research; any field requiring large-scale data access; open science initiatives (www.sciencedirect.com)

Table 1. Key characteristics comparing OpenAlex, Semantic Scholar, and PubMed (headings indicate published data with references).

Detailed Comparison

The following sections examine multiple facets where OpenAlex, Semantic Scholar, and PubMed differ or align, including coverage, search features, data accessibility, and real-world performance. Where available, we cite data from official sources, recent studies, and benchmarks.

Scope and Coverage

- **Domain Coverage:** PubMed is explicitly focused on **biomedical and health-related fields**, including life sciences, behavioral sciences, chemical sciences, and engineering related to health (pubmed.ncbi.nlm.nih.gov) (academic.oup.com). It largely omits disciplines outside health (e.g. computer science, physics, humanities). In contrast, Semantic Scholar and OpenAlex target **multidisciplinary** literature. The OpenAlex documentation even shows it supports 256 distinct disciplines (from “Physical Sciences” to “Arts & Humanities”) (wiki2.org). Semantic Scholar similarly gathers papers from nearly all scientific publishers it can, though it may have less coverage in humanities and very recent niche fields.
- **Size of the Corpus:** According to official and scholarly reports, PubMed currently indexes *over 40 million* records (pubmed.ncbi.nlm.nih.gov) (as of early 2025), growing at ~1–2 million new citations per year (www.sciencedirect.com). Semantic Scholar now indexes on the order of *200+ million* papers (www.semanticscholar.org): for example, its website proudly notes “over 200 million academic papers” in its corpus. OpenAlex has *over 250 million* indexed works (link.springer.com). These numbers reflect total entries, which include some duplicates (erroneous or variant records) and certain non-article types.
- **Geographic and Language Coverage:** Because PubMed relies on NLM’s MEDLINE journal selection, its coverage is skewed towards journals indexed by MEDLINE (mostly English-language and many Western publishers). OpenAlex intentionally includes journals worldwide, including many not in MEDLINE, and early analyses show rapid uptake in the Global South (www.sciencedirect.com). For example, one review notes that adoption of OpenAlex has “increases inclusion of underrepresented literature” from non-Anglophone regions (www.sciencedirect.com). Semantic Scholar’s coverage similarly draws from global sources, but it is unclear how thorough it is outside major publishers. In all cases, English dominates, but all three include articles in other languages to some extent (PubMed has MeSH translations for some terms, OpenAlex records language fields if available (www.sciencedirect.com)).
- **Document Types:** PubMed primarily contains journal **articles and abstracts**; it does not include full texts, though it links to free PMC articles when available (pubmed.ncbi.nlm.nih.gov). It also includes some books and chapters via NCBI Bookshelf (pubmed.ncbi.nlm.nih.gov). Semantic Scholar indexes mainly peer-reviewed journal articles and conference papers, and it explicitly excludes patents; it also includes preprints (arXiv) and some book chapters. OpenAlex includes “works” broadly – journal articles, conference proceedings, books, chapters, datasets, and even data elements; essentially anything with a DOI or repository entry. This breadth makes OpenAlex valuable for cross-domain bibliographic studies, though it also means more “clutter” (records of less-standard works).

Search and Retrieval Features

- **Search Interface:**
- *PubMed:* The PubMed website provides a straightforward search box with auto-Boolean (AND) operator insertion and automatic term mapping (ATM) to MeSH vocabulary (academic.oup.com). It supports advanced syntax (field tags like `[Title]`, date ranges, author searches, etc.) and filters (article type, publication date, species, etc.). For example, PubMed will automatically map a user query term to relevant MeSH descriptors to improve recall (academic.oup.com). Many users rely on PubMed's advanced filters for evidence-based medicine (e.g. clinical trials, reviews, case reports). Queries can be issued through NCBI's E-utilities API for programmatic access, and complete citation downloads are available via their FTP site (support.nlm.nih.gov).
- *Semantic Scholar:* The Semantic Scholar web interface functions more like a modern search app. It offers keyword search, but also provides results ranked by "influence" (using internal citation-based metrics). Each result card highlights key phrases and citation contexts. Notably, Semantic Scholar generates AI-driven features: a **TL;DR** summary of the paper, a "Citations" graph, and a "Topics" section. Registered users can create personalized research feeds and save libraries. There is also a public REST API (though subject to rate limits) and the ability to retrieve data through bulk datasets (e.g. the Semantic Scholar Open Corpus on AWS). The search engine supports free-text plus some filters, but it lacks formal controlled vocabularies like MeSH.
- *OpenAlex:* The API supports text queries such as `https://api.openalex.org/works?search=CRISPR` and embedding-based semantic search (OpenAlex). Model-based reranking is optional in the API and enabled by default on openalex.org (search documentation). API accounts receive $1/day of free usage, with paid additional usage (pricing).
- **Ranking and Relevance:** PubMed historically sorted by most recent first, but now defaults to "Best Match" (a relevance ranking learned from click data). Semantic Scholar uses its own relevance algorithm emphasizing "highly influential citations" (link.springer.com). OpenAlex searches default to descending relevance_score (OpenAlex). Ranked search is therefore available on all three platforms; the claimed absence of ranking should not determine platform choice.
- **Coverage Limitations:** Because OpenAlex uses automated merging of sources, it can overcount or have duplicates (openalex sometimes contains "superfluous or false entries" as one Wiki report notes (wiki2.org)). This can affect simple searches. Semantic Scholar, while powerful, is known to be missing some very recent publications until they crawl them. PubMed's coverage is controlled: if a journal is not in MEDLINE or deposited in PubMed Central, it won't appear. In practice, comparative studies show that OpenAlex and Semantic Scholar often have similar coverage of references, but each misses some entries the other has (subugoe.github.io). For example, matched source-reference counts for papers from 2015–2023 were about 982.6 million in OpenAlex versus 896.8 million in Semantic Scholar (subugoe.github.io), and OpenAlex had slightly more references on average (25.52 per article) than Semantic Scholar (23.29) (subugoe.github.io).

Data Accessibility and Licensing

- **Open Access / Licensing:** OpenAlex metadata is CC0 and its snapshot is free to download (OpenAlex); hosted API accounts receive $1/day free, with prepaid usage or annual plans for more (pricing). Semantic Scholar API reuse is limited to the rights expressly granted in its API license agreement. Before using the API for research, education or commercial work, review the standard grant, expanded-license provisions and underlying content rights. PubMed is free to query, and PubMed data (MEDLINE baseline dumps) are available on FTP; an NLM support page explicitly states “you no longer need a license” to download PubMed citations (support.nlm.nih.gov). However, PubMed’s data dump usage is subject to NLM’s terms (generally liberal, but not explicitly CC0). OpenAlex also sells annual plans with larger daily API budgets (pricing).
- **APIs and Data Exports:**
- *PubMed* provides the Entrez Programming Utilities (E-utilities) for query and retrieval of citations, authors, and abstracts. There is a standard usage limit to prevent abuse, but bulk access is possible via annual and daily XML release files (support.nlm.nih.gov). The NLM site explains how to download the entire MEDLINE dataset.
- *Semantic Scholar* offers an API (GraphQL and REST endpoints) for searching the index and retrieving paper details. It also released the S2ORC dataset on AWS (around 900GB compressed) for researchers (link.springer.com). However, not all internal fields (like their “influence scores”) are published.
- *OpenAlex* provides a RESTful public API covering eight entity types (works, authors, sources, institutions, concepts/topics, publishers, funders, countries) (developers.openalex.org). It also publishes complete snapshots (e.g. on AWS, GitHub, HuggingFace) for offline use. Notably, OpenAlex includes an API key system for higher-rate access. Costs are $1 per 1000 search queries (generous free quota exists) (developers.openalex.org).

The takeaway: **OpenAlex offers the most open and unrestricted data access**, making it a favorite for data scientists and open-science projects. Semantic Scholar is also open but somewhat more curated, while PubMed primarily serves as a search service (with data dumps as a byproduct).

Search Performance and Case Studies

- **Relevance and Result Quality:** Several evaluations have compared these and other search tools. For biomedical queries, PubMed remains extremely popular due to its precision. Lu’s 2011 survey reviewed 28 web systems for searching biomedical literature (Lu, 2011). Semantic Scholar, while newer, has been praised for its AI-driven “influential citation” ranking, which can surface highly cited or important papers early (link.springer.com). OpenAlex also ranks search results by descending relevance_score by default (OpenAlex).
- **Coverage Example – References:** The blog analysis by Haupka (2025) compared reference lists in OpenAlex vs. Semantic Scholar for ~37.5 million articles (2015–2023) (subugoe.github.io). It found that total cited references were roughly **982.6 million** in OpenAlex versus **896.8 million** in Semantic Scholar (subugoe.github.io). On a per-article basis, OpenAlex averaged ~25.52 references while Semantic

Scholar (counting only “source references” within its corpus) averaged ~23.29 (subugoe.github.io). The distribution varies by publisher: OpenAlex had more references for major publishers like Elsevier and Springer, whereas Semantic Scholar had relatively higher counts for others like Frontiers and SAGE (subugoe.github.io). This reveals that which database has “more complete” reference data depends on the literature segment. In aggregate, no single source perfectly covers all citations; analysts often merge the two for completeness (subugoe.github.io) (www.sciencedirect.com).

- **Domain-Specific Searches:** In biomedicine, specialized search tools often augment PubMed. For instance, during the COVID-19 pandemic, Berkeley Lab developed **COVIDScholar** to help researchers search COVID-19 literature (Berkeley Lab). Meanwhile, NLM provided *LitCovid*, a curated database of COVID articles. Such examples highlight that even within a field, no single index suffices: new dedicated systems (e.g. LitCovid, CoronaCentral, COVID-SEE) were built on top of PubMed or Semantic Scholar data (www.sciencedirect.com).
- **User Studies:** Survey research indicates that users’ preferences can vary. Many clinicians and biologists still default to PubMed, citing its simplicity and trustworthiness (academic.oup.com). Computer scientists or AI researchers often mention Semantic Scholar for its visualizations and summary features. Data analysts cite OpenAlex for large-scale studies: for example, the 2025 Leiden Ranking Open Edition uses OpenAlex, while the Traditional Edition uses Web of Science (CWTS). Notably, scholarly surveys of search engines have consistently found trade-offs: one comparative analysis found only 14 of 28 popular academic search engines (including PubMed) met all key search requirements (link.springer.com), underscoring that each system has gaps.

Metadata Quality and Limitations

No database is perfect: each has known shortcomings. A 2026 review of OpenAlex notes that **open metadata** brings incomplete or inconsistent fields. For instance, some afflicted issues remain with author affiliations (sometimes missing institution names) and document type classification (www.sciencedirect.com). Language tags and funding data are also less reliable. These gaps mean caution is needed if, say, analyzing international collaborations solely via OpenAlex. Researchers have observed that OpenAlex sometimes carries “superfluous or false entries” from aggressive automated harvest (wiki2.org).

Semantic Scholar, being an AI application, occasionally mis-parses PDFs or omits references if they cannot be linked. Its focus on English abstracts means older or foreign-language works might be underrepresented. PubMed’s limitations are different: although highly curated, it can have a lag time (though currently only days) and it strictly omits non-biomedical journals. Moreover, PubMed provides only *abstracts*, not full text, which can leave out valuable data.

These quality issues imply that users should, as a best practice, *cross-check multiple sources*. For example, Visser et al. (2021) explicitly recommended that bibliometric analysts use several databases to ensure coverage reliability (www.sciencedirect.com). In practice, a systematic review on topics like emerging diseases often queries both PubMed and at least one broad index like Scopus or Google Scholar alongside Semantic Scholar/OpenAlex.

Future Directions and Implications

The landscape of academic search continues to change. Some anticipated trends and implications are:

- **AI and LLM Integration:** Both Semantic Scholar and PubMed communities are exploring large language models. The recent surge of ChatGPT and related tools has raised interest in conversational literature search. The eBioMedicine review notes that LLMs are being applied to biomedical retrieval tasks (www.sciencedirect.com). Future systems may combine traditional indexing with generative summaries or Q&A. Semantic Scholar, for example, has already prototype features like AI assistant (Semantic Reader) for skimming (link.springer.com). However, reliance on AI also introduces risks of hallucination and bias, so expect mixed usage (AI for suggestion, but humans verifying).
- **Open Science and Equity:** OpenAlex's success reflects a broader push for open science infrastructure. Institutions and funders are increasingly demanding open data and tools. The Sorbonne case (www.sciencedirect.com) is emblematic: scholars no longer rely exclusively on commercial citation databases. OpenAlex and Semantic Scholar embody this ethos by providing free alternatives. This democratization may reduce knowledge access gaps between wealthy and resource-limited institutions. However, it also means sustainability depends on continued non-profit funding. Researchers may also shift citation practices if open indexes reveal different trends (e.g. more visibility for non-Western research).
- **Competition and Consolidation:** While we focus on these three, the search domain is dynamic. Google Scholar remains dominant in many fields (with >400 million records, unindexed here), and new players (Dimensions.ai, TheLens, SciLit) compete. Over time, we may see mergers or partnerships. For example, the Leiden Ranking Open Edition uses OpenAlex data (CWTS). It's plausible that tools will interlink: e.g. one could build an interface that queries both OpenAlex and Semantic Scholar APIs and merges results on the fly.
- **Ecosystem and Interoperability:** The use of standard identifiers (DOI, ORCID, ROR) in OpenAlex (www.sciencedirect.com) and increasingly in Semantic Scholar (via Microsoft Academic identifiers in its early data) fosters interoperability. In the future, we expect richer linking: for example, linking from a PubMed abstract to OpenAlex topics or Semantic Scholar summaries. Efforts like NISO's growing bridge between IR and library data suggest these systems will not remain silos.
- **Research on Research:** Academics are also critically studying these tools themselves. The literature is growing on altmetrics, on search biases, on how well these indexes represent the "true" scholarly record. We cite several such studies throughout (see References). This means that any recommendation we make now may evolve. For instance, if OpenAlex's known metadata issues get resolved (improvements are already underway), its standing relative to legacy DBs will improve further.

In sum, the future of literature search is toward hybrid, multi-system approaches augmented by AI. No single database will likely achieve 100% coverage or perfect search. Users should select tools based on their specific needs: **PubMed** for authoritative biomedical queries, **Semantic Scholar** for broad AI-enhanced discovery, and **OpenAlex** for open-data-intensive tasks and bibliometrics. Researchers should also stay alert to new

developments: e.g. NIH's NIH Open Citation Collection or Europe PMC enhancements, which may add new facets. As one systematic review on search tools put it, researchers and developers must "keep current" by mixing sources and adopting emerging technologies (academic.oup.com) (link.springer.com).

Conclusion

The choice of literature source ("OpenAlex vs Semantic Scholar vs PubMed") ultimately depends on the user's domain and goals. PubMed remains the gold standard for **biomedical searches**, offering concentrated coverage and tools tailored to health sciences (academic.oup.com). Semantic Scholar excels in **multidisciplinary exploration**, leveraging AI to unearth connections and summaries across science and technology (www.semanticscholar.org) (link.springer.com). OpenAlex shines in **bibliometric research and open science initiatives**, providing the raw data needed for large analyses and covering fields often overlooked by PubMed (link.springer.com) (www.sciencedirect.com). OpenAlex's open metadata should be distinguished from its metered API service: accounts receive $1/day free, with paid additional usage (pricing).

Empirical studies suggest that a combined strategy often works best. Academics conducting comprehensive literature reviews frequently consult both PubMed and an AI-powered general engine (like Semantic Scholar or Google Scholar) to ensure no major paper is missed. Bibliometricians use OpenAlex together with one or more commercial databases to cross-validate metrics. In fact, experts like Visser et al. explicitly advise **complementing** one source with others to achieve full coverage (www.sciencedirect.com).

In conclusion, there is no one-size-fits-all answer. We advise: start with the source closest to your field (PubMed for biomedicine, Semantic Scholar or OpenAlex for broader fields), but be prepared to use others if needed. Regularly consult multiple platforms, especially for critical tasks like systematic reviews or policy-making literature scans. The landscape is evolving rapidly – new features (AI summaries, translational search) and new players (LLM-based tools) are emerging. Staying informed of these changes, as documented in recent analyses (www.sciencedirect.com) (subugoe.github.io), will help users leverage each platform's strengths and mitigate its weaknesses.

References: Inline references above provide detailed sources for all data and claims (e.g. PubMed stats (pubmed.ncbi.nlm.nih.gov), OpenAlex adoption (www.sciencedirect.com), comparative studies (subugoe.github.io) (link.springer.com)). See the cited literature for in-depth evaluations, case studies, and expert opinions. These include peer-reviewed reviews, official documentation pages, and meta-analyses that underpin every key point in this report.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://pubmed.ncbi.nlm.nih.gov/about
2. https://academic.oup.com/database/article/doi/10.1093/database/baq036/460587
3. https://www.semanticscholar.org/about/team
4. https://link.springer.com/article/10.1007/s13278-025-01476-1
5. https://help.openalex.org/access/pricing

6. https://www.sciencedirect.com/science/article/pii/S1751157726000337?dgcid=rss_sd_all
7. https://wiki2.org/en/ImpactStory
8. https://help.openalex.org/api/semantic-search
9. https://help.openalex.org/api/searching
10. https://github.com/allenai/s2orc
11. https://support.nlm.nih.gov/knowledgebase/article/KA-03504/en-us
12. https://open.leidenranking.com/resources
13. https://subugoe.github.io/scholcomm_analytics/posts/s2_reference_analysis/s2_reference_analysis.html
14. https://www.nature.com/articles/nature.2015.18703
15. https://support.nlm.nih.gov/knowledgebase/article/KA-05326/en-us
16. https://www.sciencedirect.com/science/article/pii/S2352396424000239
17. https://techcrunch.com/2017/10/17/allen-institute-for-ais-semantic-scholar-adds-biomedical-papers-to-its-ai-sorted-corpus
18. https://techcrunch.com/2019/10/23/ai2s-semantic-scholar-expands-to-cover-178-million-papers-in-all-scientific-disciplines
19. https://www.sorbonne-universite.fr/en/news/sorbonne-university-unsubscribes-web-science
20. https://developers.openalex.org/guides/key-concepts
21. https://help.openalex.org/access/website-basic
22. https://pmc.ncbi.nlm.nih.gov/articles/PMC6112631
23. https://www.semanticscholar.org/product/api/license
24. https://help.openalex.org/access/overview
25. https://api.semanticscholar.org/license
26. https://developers.openalex.org/guides/searching
27. https://newscenter.lbl.gov/2020/04/28/machine-learning-tool-could-provide-unexpected-scientific-insights-into-covid-19
28. https://www.leidenmadtrics.nl/articles/the-cwts-leiden-ranking-2025-more-open-more-inclusive-more-informative

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.