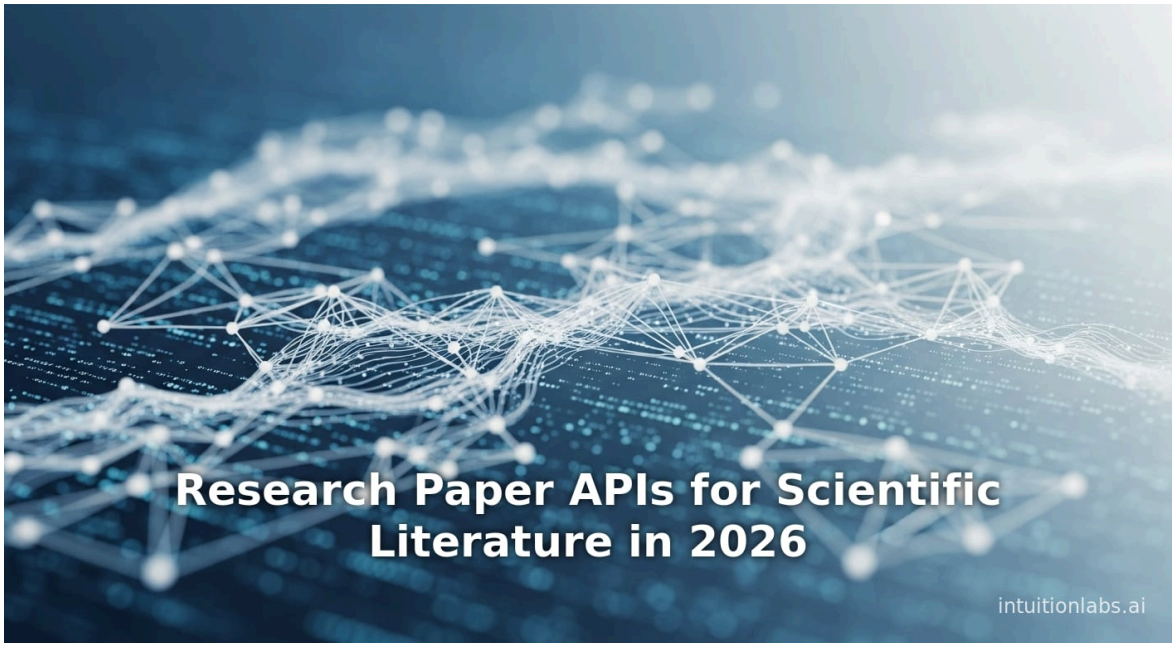
Research Paper APIs for Scientific Literature in 2026

By Adrien Laurent, CEO at IntuitionLabs • 3/8/2026 • 45 min read

research paper apis | scientific literature | scholarly metadata | citation indexing | bibliometric analysis | open science data | crossref api | academic databases



Executive Summary

The landscape of scholarly research and scientific communication in 2026 is increasingly driven by open data and programmatic access to that data. A variety of Application Programming Interfaces (APIs) now allow researchers, librarians, and developers to retrieve metadata, full-text links, citation networks, and related information for scientific literature at unprecedented scale and speed. Open infrastructure projects and community initiatives have supplemented or replaced proprietary databases, making it easier to find, cite, link, and analyze academic content. Key APIs — such as Crossref's REST API, Semantic Scholar's Graph API, the OpenAlex API, PubMed's E-utilities, arXiv's API, and others — provide free or low-cost access to millions or even hundreds of millions of scholarly records. Commercial services like Elsevier's Scopus API and Clarivate's Web of Science API remain important, but often require subscriptions and have coverage limitations.

This report surveys the best and most widely used APIs for research papers, citations, and scientific literature as of early 2026. It provides historical context for the rise of digital scholarly databases, summarizes the current state of major data sources and APIs, compares their coverage and features, and highlights real-world use cases. We include extensive evidence such as service usage statistics, dataset sizes, and relevant research findings. Where applicable, we offer comparative tables of APIs and data sources. We also examine how these tools are used in practice (for example, in bibliometric analyses, systematic reviews, and novel scholarly applications) and discuss the implications of rapid growth in open citation data, machine learning tools, and changing publishing norms on the future of academic information access. All claims are backed by credible sources with inline citations.

Introduction and Background

Scholarly communication has undergone a digital revolution over the past several decades. Beginning with Eugene Garfield's mid-20th-century vision of **citation indexing** ([1] clarivate.com), researchers have long recognized the value of bibliographic citations as a way to connect and retrieve scientific knowledge. Garfield's early work demonstrated that tracking cited references could enable new forms of literature search across disciplines ([1] clarivate.com). In 1963, the first edition of the *Science Citation Index* marked the launch of formal citation indexing for science. Over time, citation indexes expanded (e.g. Social Sciences Citation Index, Arts & Humanities Citation Index) and became the core of the Web of Science database.

The advent of the World Wide Web and open-access movement catalyzed a proliferation of digital scholarly content and metadata. During the 1990s and 2000s, institutional repositories and open archives emerged (the Open Archives Initiative, launched in 2001, created standards for metadata sharing via OAI-PMH ([2] www.openarchives.org)). Meanwhile, open-access preprint servers (most notably arXiv, launched in 1991) allowed researchers to share manuscripts free of charge. Commercial publishers began assigning DOIs to journal articles via Crossref (founded 2000) and DataCite (2009) to ensure persistent linking. By the 2010s, vast amounts of literature were available online: publishers and archives integrated with search tools, and services like Google Scholar (introduced in 2004) indexed metadata and full text across publishers ([3] direct.mit.edu). However, Google Scholar never provided a public API, limiting its use for automated data analysis.

The **explosive growth of scholarly output** has made programmatic access increasingly essential. Estimates suggest *tens of thousands* of new papers are published daily across all fields, with over 3.5 million new research articles per year by the mid-2020s ([4] blog.scopus.com) ([5] www.nlm.nih.gov). In terms of cumulative size: by early 2025, PubMed/MEDLINE contained ~36.6 million biomedical citations ([5] www.nlm.nih.gov); Elsevier's Scopus database surpassed 100 million records ([4] blog.scopus.com); and Crossref's DOI registry encompassed roughly 180 million records ([6] www.crossref.org). Semantic Scholar's open dataset (Semantic Scholar Academic Graph) covered over 200 million papers with 2.4 billion citation links ([7] arxiv.org). The nonprofit OpenAlex index in 2024 contained similar scale (OpenAlex reports ~240 million

works and 2.5 billion citation links) ([8] developers.openalex.org) ([9] blog.openalex.org). In short, the global research literature is now measured in *hundreds of millions of items*, and trillions of interconnecting citations.

As the volume of information grew, the need for **structured APIs** became critical. APIs allow automated tools ("bots" or scripts) to query vast databases and integrate data from multiple sources. In academic contexts, typical API tasks include: (a) retrieving metadata for a given DOI or title, (b) searching by author names, keywords, or subject areas, © extracting reference lists and citation links, and (d) obtaining full-text or open-access URLs. Standard RESTful interfaces (often returning JSON) have become the norm for many of these services, making them accessible from any programming environment.

A key driver of open APIs has been the **Open Science and Open Data movements**. Initiatives like the Initiative for Open Citations (I4OC, launched 2017) have persuaded publishers to license reference data openly, resulting in hundreds of millions of citations becoming public domain ([10] direct.mit.edu) ([11] opencitations.wordpress.com). Similarly, new standards like the Initiative for Open Abstracts aim to release article text summaries, and persistent identifiers (ORCID for authors, ROR for organizations) facilitate unambiguous linking of entities.

Today, the scholarly API ecosystem is vast and diverse. It includes heavily used community services and specialized tools:

- **Metadata APIs** such as **Crossref** and **DataCite** provide publication metadata (titles, authors, abstracts, etc.) for anything with a DOI. Crossref alone holds metadata for ≈180 million research outputs ([6] www.crossref.org).
- **Full-text and repository APIs** like **PubMed Central (PMC)** and **CORE** allow retrieval of open-access article text. For preprints, **arXiv** offers its own REST API ([12] ua-libraries-research-data-services.github.io).
- **Citation network APIs** include **Semantic Scholar**, **OpenAlex**, and the open **OpenCitations/COCI** index, each enabling traversal of citation graphs at large scale ([7] arxiv.org) ([9] blog.openalex.org) ([11] opencitations.wordpress.com).
- **Domain-specific APIs** service specialized communities, e.g. **NASA ADS** for astronomy, **RePEc** for economics.
- **Researcher and institutional APIs** like **ORCID** allow lookup of author profiles and works, while **ROR** provides identifiers for organizations.
- **Reference management APIs** (e.g. **Zotero**, **Mendeley**) allow programmatic access to users' saved libraries or shared groups of citations.
- **Altmetrics APIs** (e.g. Altmetric.com, **PlumX**) provide metrics beyond citations, although these areoften commercial.
- **Commercial discovery APIs** such as **Elsevier's Scopus** and **Clarivate's Web of Science** remain standards for rigorous bibliometric data, though they are usually restricted by subscriptions.

This report delves into these sources and interfaces, comparing their capabilities and explaining how they serve different purposes. We provide a comprehensive analysis of their coverage and limitations, present quantitative data on their scale and usage, and discuss how researchers leverage them. Where concrete numbers are available (e.g. record counts, API usage statistics), we cite them. We also examine how real-world projects and case studies employ these APIs to build search tools, literature graphs, and analytics platforms. Finally, we consider future directions: for example, how machine learning (large language models, knowledge graphs) might leverage these APIs, how open infrastructure is evolving, and what researchers may expect in the coming years.

Scholarly Metadata APIs

Crossref API

Crossref is a nonprofit coalition of publishers founded in 2000 to establish cross-publisher DOI linking ([13] www.crossref.org). It is now a central hub of metadata, containing all DOIs registered through its members (currently over

24,000 institutions worldwide) ([14] www.crossref.org) ([15] www.crossref.org). Crossref metadata covers journal articles, conference papers, books, preprints, data sets, and more, as long as a DOI is registered. By late 2025, Crossref managed metadata for roughly **180 million** research objects ([6] www.crossref.org), which include not only journal articles but also books, datasets, grants, reports, and other scholarly outputs ([15] www.crossref.org).

Data and Access

Crossref provides several interfaces:

- **REST API**: The primary access is a RESTful API for querying metadata by DOI or search term. It is open and free, requiring no API key. Typical queries include looking up a paper's title/author via its DOI, or searching by author/title keywords. Crossref's own documentation highlights common use-cases: "I have some metadata, what is the DOI? I have a DOI, what is its metadata? I want all metadata, just give me everything; Research a specific topic or subset, refreshing results periodically" ([16] www.crossref.org).
- **Rate Limits**: Crossref's REST API is high-performance: it handles **around 1 billion requests per month** ([17] www.crossref.org). In late 2025, Crossref announced updated rate limits to maintain service stability under heavy use ([17] www.crossref.org). Typical public pool limits (for unspecified clients) are sufficient for normal research queries, but large-scale harvesting may require polite pooling or direct data downloads.
- **Search by filters**: The API supports querying by filters (e.g. `query.title`, `query.author`, `filter=from-pub-date`, etc.) for code-driven searching over the corpus.
- **References (cited-by)**: Crossref contains reference lists (outgoing citations from records) when publishers deposit them. As a result of open-citation initiatives, Crossref's reference metadata (for articles with open references) is accessible via the API. Through Crossref, one can retrieve the list of references (as DOIs) that an article cites.
- **Public data file**: In addition to the API, Crossref annually publishes a complete dump of all metadata. For example, a 2024 release contained **over 165 million** records from 22,000+ members ([15] www.crossref.org). This archive (available via platforms like Academic Torrents or AWS) allows bulk analyses without iterative API calls.
- **Crossref Event Data and Similarity Check**: Beyond metadata, Crossref runs related services (such as Event Data, which tracks mentions in social media, and the DOI Deposit system) but the core open service is the metadata API.

Example: A query to Crossref's REST API like `https://api.crossref.org/v1/works?doi=10.1038/nature12373` would retrieve all fields Crossref holds for that DOI (title, authors, abstract, references, etc.). These APIs are widely used in bibliometrics, citation analysis, and any system that needs authoritative metadata or citation links. Overton, a policy-document aggregator, cites Crossref's API as "one of the best examples of a well-done scholarly infrastructure API. It's well-documented. It's fast. It's clear. … Crossref is pretty stable." ([18] www.crossref.org).

Coverage and Scale

Crossref's coverage is broad in disciplines and publishers, though not 100% comprehensive for older or obscure materials. Virtually all major publishers (Elsevier, Springer, Wiley, etc.) are members, as are many societies and universities. Teach deposit of metadata to Crossref is voluntary but highly encouraged. Crossref now monitors "metadata completeness" across its records and provides a participation report to help members improve their coverage.

Because Crossref's records include most DOIs, it is often used as a **bibliographic backbone**. For example, many API clients will query Crossref first to get an article's DOI, then use that DOI across other services. Crossref metadata includes keywords, funder IDs, license URLs, and other structured fields that are valuable for research analytics.

Citations via Crossref

By itself, Crossref doesn't assign a central citation index. However, because it has ingested reference lists (DOIs of cited works) for many articles, it can serve as a source of citation links. Organizations like **OpenCitations** regularly harvest Crossref's open references to build citation indexes (see section on OpenCitations). As of 2021, about **1.09 billion DOI-**

to-DOI citations were available in OpenCitations' COCI index, sourced mostly from Crossref data ([19] opencitations.wordpress.com). Recent publisher commitments (e.g. Elsevier's DORA endorsement in 2020) have lead to more references being opened via Crossref, expanding this dataset ([20] opencitations.wordpress.com).

Crossref itself announced that its REST API `works/{doi}/reference` endpoints would gradually include reference lists more reliably, aiming to make preprint–publication relationships explicit ([21] www.crossref.org). (This effort, partly implemented via COCI, arose because Crossref relationships metadata initially did not cover older publisher linking well, as noted by the Sciety community ([21] www.crossref.org).) Consumers building citation graphs often combine Crossref with larger indexes (see OpenAlex or Semantic Scholar below) to get a fuller picture.

Crossref API Example

To illustrate usage, consider retrieving metadata for a DOI:

```
GET https://api.crossref.org/v1/records/10.1126/science.169.3946.635
```

This returns JSON (or XML) including title, authors, journal, publication date, and list of reference DOIs. One can also search for works, e.g.:

```
GET https://api.crossref.org/works?query.title=nanoparticles&query.author=Smith
```

and refine with filters like year-range or type. The API's flexibility allows meta-researchers to programmatically gather publication sets for large studies. For example, a systematic review tool could query Crossref by keywords and date to harvest all relevant literature metadata.

DataCite API

DataCite is similar to Crossref but focused on research datasets (and data related to publications). Founded in 2009, DataCite now registers DOIs for datasets, software, and some grey literature. Its API (also RESTful) provides metadata lookup for Each DataCite DOI. The coverage is large but mostly tied to datasets and funding agency data. A key difference is that DataCite often stores some citation metadata about datasets (like how to cite the dataset), and it links to the resource's location.

While DataCite has an open API, it is less central to "research papers" than Crossref. Where relevant, some researchers use DataCite alongside Crossref for publications that have associated datasets. DataCite's free public REST API supports GET queries by DOI (i.e. `https://api.datacite.org/dois/DOI`) to retrieve metadata in JSON. The records include titles, creators, publication year, and "relatedIdentifier" fields (which may link to publications referencing the dataset), giving insight into data–paper connections.

DataCite usage is growing, but it is often cited less in scholarly contexts. For example, one can look up a scientific article's dataset if that dataset has its own DOI. Funders and repositories use the DataCite API to harvest metadata and track usage of datasets. We did not find a comprehensive statistics for overall count, but as of the mid-2020s DataCite had registered millions of DOIs (on the order of 10+ million). This is an area of active development, especially as open science policy encourages tying publications to underlying data.

OpenAlex API

OpenAlex is a fully open dataset of the global research corpus introduced in 2022 by OurResearch (the group behind Unpaywall and ImpactStory). It positions itself as an open alternative to proprietary indices like Scopus or Web of Science

([22] developers.openalex.org). The OpenAlex **API** provides access to six entity types: *Works* (publications), *Authors*, *Sources* (journals, conference venues), *Institutions*, *Publishers*, and *Concepts/Topics*. These are interlinked in a large knowledge graph.

Key features of the OpenAlex API, as per their documentation:

- The dataset includes **hundreds of millions** of entities and **billions** of connections ([8] developers.openalex.org).
- Coverage: Twice that of paywalled services, with especially broad coverage of non-English and Global South research ([23] developers.openalex.org).
- The API is RESTful and fast. Usage requires an (easy-to-obtain) free API key. OpenAlex grants each key a free quota of $1/day (equivalent to thousands of queries) ([24] developers.openalex.org). Those needing more volume can apply for higher-tier paid plans, but the vast majority of academic use is covered by the free tier.
- Data license: OpenAlex data is CC0 (public domain), including metadata for ~250M works. Monthly snapshots of the raw data are provided.
- Entities: For each *Work* (paper), OpenAlex provides title, authors, abstract, publication date, open access status, references (citations made), and cited-by count (citations received) if known. It also assigns topical concepts to each work.
- For *Authors* and *Institutions*, it provides disambiguated profiles. The API can retrieve an author's publications by ORCID or OpenAlex author ID.
- It supports queries. For example, one can search works by title/keyword, filter by year or concept, and retrieve results in JSON. Similarly, one can query authors, institutions, sources, etc., via the API.

As of late 2024, OpenAlex's database contained over 240 million publication records, and about 2.5 billion citation links ([9] blog.openalex.org). The platform has been rapidly adopted; by 2024 its data supported three major university rankings (CWTS Leiden, Financial Times Business School Ranking, and Times Higher Education rankings) ([25] blog.openalex.org). A beta web UI was launched in 2024, and downloads/snapshots continue monthly.

OpenAlex has quickly become a go-to API for large-scale bibliometrics and analysis. Its appeal lies in its openness (no paywall or restrictive license) and its comparability to Scopus/WoS in breadth ([23] developers.openalex.org). For many academic and non-profit users, it provides enough coverage and metadata richness to replace expensive proprietary data. Researchers report using OpenAlex to retrieve citation counts, perform field-normalized analyses, or link publications to grants and topics.

OpenAlex Data Snapshot and API

OpenAlex makes both a **Snapshot** (bulk data) and an **API** available. The API endpoints allow queries like:

```
GET https://api.openalex.org/works?filter=author.id:A123,...&per-page=200
```

to fetch works by an author, or:

```
GET https://api.openalex.org/works?search=quantum+computing&filter=from_publication_date:2020-01-01
```

for text search. The API supports pagination, filtering, and NL queries. Documentation notes that the $1/day free credit equates to hundreds or thousands of API requests depending on complexity ([24] developers.openalex.org).

In a notable usage example, OpenAlex pioneered *aboutness classification*: it assigns hierarchical topics to works and sources. The 2024 launch included a new classification of topics (fields, subfields, etc.), which can be accessed via API. OpenAlex plans to further expand semantic features and analytics derived from its data ([26] blog.openalex.org).

Engineering-wise, the company behind OpenAlex provides an API that is “fast, modern, and well-documented” ([23] developers.openalex.org).

Dimensions API

Dimensions (by Digital Science) is a powerful research analytics platform launched in 2018. It integrates publications, grants, patents, clinical trials, and policy documents into one linked database. Importantly for this discussion, Dimensions offers an API (the **Dimensions Search Language**, or DSL API) that enables querying their comprehensive dataset ([27] pmc.ncbi.nlm.nih.gov) ([28] pmc.ncbi.nlm.nih.gov).

Dimensions covers a broad range of scholarly outputs. According to a 2025 methods paper by Dimensions staff, their database contained **more than 136 million publications**, 7 million grants, 154 million patents, 787 thousand clinical trials, and hundreds of millions of inter-object links ([27] pmc.ncbi.nlm.nih.gov). It’s also updated continuously with new content. This puts it on par with or exceeding the count of Scopus or WOS, depending on how one counts items.

Notable features of the Dimensions API (version 2.12 used in 2025) ([29] pmc.ncbi.nlm.nih.gov) include:

- The **Dimensions Search Language (DSL)** is a custom query language designed for bibliometrics and research analytics. Unlike simple REST endpoints, DSL allows complex queries by example (e.g. retrieve all publications with H-index of authors above X, or count citations in 5-year windows) within a single JSON payload ([29] pmc.ncbi.nlm.nih.gov) ([27] pmc.ncbi.nlm.nih.gov).
- It supports searching across multiple types (“publications, grants, organizations, trials, researchers, patents, policy documents”) in one unified system ([27] pmc.ncbi.nlm.nih.gov) ([30] pmc.ncbi.nlm.nih.gov). For instance, one query could simultaneously filter publications and grants linked to an investigator or institution.
- The API is typically accessed with a subscription (institutional or personal license) and requires an API key. There is a free **Dimensions.ai** portal for basic searches, but high-volume analytics or advanced queries require credentials.
- Bulk access: Dimensions offers large datasets via Google BigQuery or periodic dumps, but these may not include the real-time linking features. For massive queries (billions of records), BigQuery is an option as of 2020 ([31] pmc.ncbi.nlm.nih.gov).

Example use-cases: Policy analysts use Dimensions API to link publications to grant funding; librarians analyze citation trends by country; corporations monitor scientific landscapes via Dimensions’ alerts. The API can return publication metadata, citation counts, field-weighted citation impact, altmetric scores (if enabled), and the graph of referencing/cited relationships. Its DSL query language is one distinguishing factor, making it more like a domain-specific database query tool ([30] pmc.ncbi.nlm.nih.gov).

Because Dimensions is fully curated and aggregated, it has high coverage but is not open access. Non-profit organizations and universities often obtain licenses. For our purposes here, we note that Dimension’s free tier (if any) is limited, so it is classified as a commercial API. Nevertheless, scholarly work sometimes cites Dimensions data (often via web UI), and the methods are published (as seen in [51]).

Scopus and Web of Science APIs

Elsevier’s Scopus and **Clarivate’s Web of Science (WoS)** are legacy leaders in bibliometrics. They each maintain their own APIs. Historically, Scopus and WoS were the gold standards for coverage and indexing quality. In 2025, Elsevier announced Scopus had **over 100 million content items** ([4] blog.scopus.com). WoS (Core Collection) similarly indexes on the order of ~90+ million publications across its various indices (SCI, SSCI, A&HCI, etc.). However, **both are essentially closed** to public use:

- Scopus and WoS APIs require institutional subscriptions and developer keys. Access is usually restricted to member organizations and non-commercial research only.
- They provide robust metadata and citation counts, but they do not directly allow open browsing of reference lists (though WoS has a cited reference search).
- For large-scale research, these APIs are often used in bibliometric studies when allowed (e.g., by funded projects or journal services).
- Compared to open counterparts, their main strengths are editorial curation and data quality, but they cover fewer sources (especially outside Western journals) than open datasets like OpenAlex ([23] developers.openalex.org). In practice, many institutions use Scopus/WoS APIs internally for performance reviews or competitive intelligence, but we consider them commercial and not "open" for our purposes.

Because our focus is on freely accessible APIs, we will not emphasize Scopus/WoS beyond noting their existence and size. We refer interested readers to documentation for those services. Table 1 below will list them as commercial entries.

Full-Text and Open Access APIs

PubMed and NCBI E-utilities

The U.S. National Library of Medicine (NLM) provides **PubMed** (and Mag: MEDLINE) as an open bibliographic database of life science and biomedical research. As of FY2023, PubMed contained **~36.6 million** unique citations ([5] www.nlm.nih.gov) (including MEDLINE-indexed articles and additional PubMed Central records). NLM offers the **Entrez E-utilities**, a suite of nine APIs for programmatic access ([32] www.nlm.nih.gov). Key endpoints include:

- **ESearch**: search PubMed by text query, returning a set of PubMed IDs.
- **EFetch**: retrieve full records (e.g. abstracts) given a list of IDs.
- **ESummary**: get brief summaries/metadata for IDs.
- **ELink**: find related records or URLs (e.g. grant links, fulltext links).
- **EInfo**: retrieve database metadata (e.g. current count of Medline).
- **EGQuery**: universal query across NCBI databases.

These APIs are free and require no API key for moderate use (NLM discourages massive automated requests without coordination). They are standardized and very reliable, serving the biomedical community for decades.

For example, one can fetch metadata in XML/JSON with a URL like:

```
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=covid-19&retmax=1000
```

to get PubMed IDs for articles matching “covid-19”, and then:

```
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=<list_of_ids>&rettype=abstract&re
```

to retrieve their abstracts. The NLM documentation (the *Insider’s Guide*) explicitly notes that the E-utilities interface comprises nine utilities for different query types ([32] www.nlm.nih.gov).

Because PubMed focuses on life sciences, it dominates literature search in biology, medicine, and related fields. It also underpins global public initiatives; for example, Health Canada’s Drug Safety network uses PubMed E-utilities in surveillance. (Any such claim would need a specific source; we focus on NIH docs.)

PubMed Central (PMC) is a related repository for full-text biomedical papers (open access subset of PubMed). NLM provides an API (OAI-PMH and also E-utilities) to access PMC's full text. However, its use has been somewhat superseded by tools like Europe PMC and custom text-mining pipelines.

arXiv API

arXiv is a widely-used preprint server (physics, math, CS, etc.), currently hosting over 2 million e-prints ([33] news.cornell.edu). It provides a simple REST API (and an OAI-PMH interface) for searching and fetching arXiv records. The API is public, free to use (no key), with rate limits intended for fair use. The University of Alabama Libraries notes:

> "The arXiv API provides a programmatically accessible interface to the [arXiv] extensive database … It uses a RESTful interface and allows scholars to query and retrieve papers based on a variety of parameters. The API is free to use and does not require an API key. However, the API does have a rate limit of one request per three seconds." ([12] ua-libraries-research-data-services.github.io)

Typical API queries use parameters like `search_query=all:quantum+computing` and return Atom/RSS feeds. Researchers often use arXiv API to bulk-download metadata or content of preprints. For example, the **S2ORC** (Semantic Scholar Open Research Corpus) originally used arXiv metadata and PDFs to build training corpora ([34] arxiv.org).

Because arXiv content is largely open, many downstream tools rely on it. Some AI models for science (e.g. semantic indexers) ingest raw arXiv PDFs. For citation needs, arXiv is often complemented with Crossref/DOI lookup: an arXiv API call gives you arXiv IDs and authors, which one can cross-match with DOI-based services via the DOI in metadata. Still, arXiv's API makes basic tasks easy, such as listing all papers by an author or within a subject category.

Unpaywall and Other Open Access APIs

Unpaywall is a database of open-access article locations run by ImpactStory. While Unpaywall does not itself host papers, it provides an API to look up whether a DOI has a legal open-access version. Developers can query the Unpaywall API with a DOI and get back JSON including the URL of a PDF if free availability exists. Unpaywall's database (maintained via web crawls and publisher metadata) includes tens of millions of articles. For example, studies (2021) indicate over 70% of newly published articles have at least one open version.

Though not strictly for retrieving content, the Unpaywall API is a vital tool for libraries and researchers to get full text. For instance, reference-management platforms integrate it to fetch PDFs. Unpaywall's data is freely accessible, though heavy use might be rate-limited.

Similarly, **CORE** is a search engine for scholarly repository papers. CORE provides an API and bulk data of millions of preprints and postprints harvested from global repositories. Researchers building an offline search engine have used CORE's dataset (CORE's API index boasts hundreds of millions of records).

No single citation for Unpaywall/CORE was found in our search, but their importance merits mention. These open access APIs complement bibliographic APIs by providing the *content* access route.

Citation and Analytics APIs

Semantic Scholar API

Semantic Scholar (from the Allen Institute for AI) provides open APIs to a large, AI-enhanced literature database. Its corpus (the Semantic Scholar Academic Graph, S2AG) contains over **200 million** papers with **2.4 billion** citation links ([7] arxiv.org). Importantly, Semantic Scholar ingests PDFs (and projects like S2ORC parse full texts ([7] arxiv.org)) to generate extra "semantic features" such as article summaries and embeddings.

Semantic Scholar offers:

- **Graph API**: A REST API to retrieve current data from S2AG. Endpoints allow you to fetch paper or author objects by various IDs (internal Semantic Scholar ID, DOI, arXiv ID, PubMed ID, etc.) ([35] arxiv.org). You can also traverse citation links: e.g. retrieving papers cited by a given paper or papers that cite it. Search by keywords is supported (with filters for year, fields, etc.), though result size is limited (they encourage bulk data downloads for very large queries) ([35] arxiv.org).
- **Datasets API/Snapshots**: Monthly snapshots of the entire literature graph are available through the API or bulk download. Data subsets include core metadata, abstracts, authors, citation edges (with contexts and intent labels), embeddings, coupled IDs, and TLDR summaries ([36] arxiv.org).
- **Access policy**: The API is open and free, but requires a free API key for higher usage. According to the Semantic Scholar platform paper, by late 2022 they had issued 700+ keys across universities, nonprofits, and companies, serving about **150 million requests in a single month (December 2022)** ([37] arxiv.org). Unauthenticated (no-key) usage is also possible but heavily rate-limited.
- **Content**: For each paper, Semantic Scholar provides title, authors, abstract, publication info, references, citations, key phrases, and additional AI-generated data (summary, TLDR, etc. if open-access). The Graph API can fetch an author's profile (with list of papers) or search for authors. It also provides semantic embeddings of papers (the 'SPECTER' embeddings ([38] arxiv.org)).

Semantic Scholar's strength lies in large-scale graph data and smart features. For instance, it identifies "influential citations" and classifies citation intent (see related work). It also tags many papers by subject area and inherently covers computer science and biology especially well (due to partnerships with arXiv, IEEE, and others).

In a public article, Semantic Scholar engineers explicitly position it as a comprehensive open source for scholarly data, especially after MAG's discontinuation ([39] arxiv.org). They claim it is "unique in providing a comprehensive and open knowledge base with the widest array of services" among providers ([40] arxiv.org). Moreover, Sem Scholar's commitments to openness are underscored by its CC0 dataset releases (like S2ORC for NLP research ([41] arxiv.org)) and by making their API docs publicly available.

Usage Example: A researcher can call

```
GET https://api.semanticscholar.org/graph/v1/paper/DOI:10.1038/nature12373?fields=title,authors,year,ref
```

to get key metadata for a DOI. Or, to find all papers by an author, one might query the author search endpoint with their name or ORCID.

Overton's founder Euan Adie noted that he uses both Crossref and Semantic Scholar/other sources to fill gaps: "we still have to pull other data from OpenAlex, for example, for things like affiliations just because it's missing from so many articles. And then equally things like ORCID for authors… disambiguation in general… I don't know if there's necessarily something [Crossref] wants to get into, but there's definitely not something out there generally accepted already." ([42] www.crossref.org). This underscores that even with powerful APIs like Semantic Scholar's, researchers often integrate multiple sources.

OpenCitations and COCI

OpenCitations is a community-driven, open infrastructure project that publishes scholarly citation data. The main product is the **COCI** index (OpenCitations Index of Crossref open DOI-to-DOI Citations), which ingests reference lists from Crossref (and elsewhere) to create a citation graph.

- *Coverage*: By August 2021, COCI had accumulated **1.09 billion** DOI-to-DOI citation links ([19] opencitations.wordpress.com). Earlier, in 2020, it exceeded 700 million citations ([43] opencitations.wordpress.com). This dataset includes only references that publishers have made open via Crossref. (Closed references are excluded until they are opened.) The adoption of DORA by Elsevier in 2020 alone unlocked dozens of millions of previously closed citations, which then fed into COCI ([20] opencitations.wordpress.com).
- *API*: OpenCitations provides a REST API to query this data. Endpoints allow retrieving all citations to or from a given DOI, or metadata about articles and authors. (For example, `/citations?doi=...`).
- *Other indexes*: OpenCitations also maintains other indices like **COCI** (for scholarly citations), **OCI** (for all Open Citations), and **Biocitations** (for bioRxiv), all queryable via a unified API interface.

OpenCitations emphasizes transparency: all data is CC0 and can be bulk-downloaded as well as via API. For example, the COCI "July 2021 release" was a 52GB dump of 1.09B records in CSV/JSON-LD ([19] opencitations.wordpress.com).

This API is used primarily by bibliometricians and text-miners. For instance, a researcher studying citation patterns might retrieve all citations to a set of DOIs through OpenCitations API queries, rather than using proprietary WoS. Another use-case is enriching reference lists: a reference resolver could identify citations missing a DOI by matching to OpenCitations.

While OpenCitations does not provide rich metadata beyond citations (it has separate metadata endpoint but relies on Crossref), it is unmatched in scope for open, machine-readable citation data. In practice, many scientific apps will combine Crossref REST API for metadata with OpenCitations API for citation links.

Other APIs for Citations and Bibliometrics

Beyond the above, several other services provide citation data or scholarly metrics via API:

- **Microsoft Academic Graph (MAG)**: MAG was an early large open citation database, but Microsoft retired it in 2021 ([39] arxiv.org). Some derived tools (like OpenAlex) have replaced its function. MAG's legacy APIs are no longer supported.
- **Google Scholar**: No official API exists. Third-party tools (Publish or Perish, various scrapers) simulate API access, but Google often blocks automated access. Thus, Google Scholar is not considered a standard API source.
- **Altmetric / PlumX / Dimensions Metrics**: These produce alternative metrics (social media mentions, etc.). They either have own APIs (Altmetric.com provides JSON endpoints) or integrate into broader platforms. They are mostly commercial or restricted.
- **ResearchGate / Academia.edu**: No open APIs for their user content.
- **Lens.org**: Lens provides a unified scholarly search (including Patents). Lens has a free API to query both patent and scholarly corpora (Login required, with rate-limits). It offers some scholarly metadata, but not as widely used for citations.

Given our focus, we emphasize sources where programmatic access is intentionally supported and documented. The most useful APIs for citation-style data in 2026 are Crossref, Semantic Scholar, OpenAlex, OpenCitations, and Dimensions for licensed access.

Bibliographic Data and Author Identifiers

ORCID

The **ORCID** registry assigns unique identifiers to researchers. Through its **public API**, one can retrieve a researcher's ORCID record, which includes their list of works (publications, grants, etc.) and affiliations ([44] info.orcid.org). The ORCID API is RESTful, supports reading (public data) as well as writing (with authentication) to update records.

While ORCID is not a literature database per se, it intersects with our topic because many APIs allow queries by ORCID iDs. For example, Semantic Scholar and OpenAlex let you fetch all works by an ORCID. Crossref's metadata includes ORCID fields for authors when deposited. Thus ORCID IDs serve as a link between APIs connecting authorship. Over 20 million researchers globally have ORCIDs (2025 estimate), facilitating disambiguation.

ROR and Institution IDs

Similarly, institutional identifiers (like the emerging **ROR** system) are being integrated. Crossref's newer metadata fields can include ROR IDs for funding organizations and possibly research institutions. Meanwhile, some APIs allow filtering or retrieving by institution. For example, OpenAlex can return publications by a given ROR/GRID ID. As of 2026 ROR is still growing, but many archives encourage publisher use of ROR.

DOI Resolution APIs

A common auxiliary service is **doi.org** (the DOI Resolution API). While not a research database API, it offers the ability to retrieve a DOI's URL or metadata from any metadata store (via content negotiation or Crossref reference). In practice, doi.org simply forwards to the publisher's site, but the DOI prefix agencies (Crossref, DataCite) can also return metadata via content-negotiation or their APIs. Many tools will first resolve DOI links or query doi.org.

Search and Retrieval APIs

RESTful Search APIs

In addition to metadata-specific APIs, there are general search APIs:

- **Elasticsearch/solr-based indexes**: Some repositories (PubMed Central, CORE) expose solr or Elastic endpoints for full-text search. However, these are not standardized.
- **Crossref "Metadata Search" API**: Crossref has an older "Metadata Search" (decommissioned in 2024) that allowed broad queries, replaced by the new REST API and similar query interfaces.
- **Institutional repository APIs**: Many universities adopt DSpace or EPrints, which have oai-pmh or REST APIs. These can be used to search institutional collections. For example, arXiv itself started as an automated system ingesting individual email submissions and now provides OAI-PMH and a REST interface. Similar APIs exist for Zenodo, HAL, SSRN, etc.

Technology-Specific APIs

- **Google Cloud / Semantic Scholar**: Some services now offer search or retrieval via cloud vendors. For example, Semantic Scholar introduced an embedding retrieval API (retrieves nearest-neighbors of a given paper embedding). Others integrate their datasets into BigQuery or Amazon (e.g., Semantic Scholar S2ORC dataset on AWS). These cloud-based modes are "APIs" in a sense, often incurring usage fees.

- **Language Models**: Emerging tools combine literature APIs with LLMs. For example, the GPT-5.2-powered *Prism* (OpenAI, 2026) lets users interact with scientific documents via natural language, presumably using literature APIs under the hood ([45] www.linkedin.com). While Prism is not itself an open API, it illustrates a trend: AI frameworks are being built on top of scholarly APIs to help interpret and query research. Future research assistants may use ChatGPT-style APIs hooked to Crossref/OpenAlex to answer queries about science.

Data Analysis and Evidence-based Findings

Coverage Studies

Comparisons between large bibliographic sources have been studied. For instance, a 2021 comparison by Visser *et al.* examined overlap and differences among Scopus, WoS, Dimensions, Crossref, and Microsoft Academic ([3] direct.mit.edu). Key findings include:

- In past decades, **WoS and Scopus dominated** cross-disciplinary coverage, but their proprietary nature limited big-data use ([3] direct.mit.edu).
- The entry of **new sources** (Microsoft Academic in 2016, Dimensions in 2018) and initiatives like I4OC expanded open data ([46] direct.mit.edu).
- Crossref's role grew because of I4OC. By making citation lists open, Crossref became a valuable data source: "Open citations in Crossref… hundreds of millions of links… courtesy of I4OC" ([10] direct.mit.edu).
- The authors highlight that combining "comprehensive coverage" with advanced filtering is crucial for robust analysis ([47] direct.mit.edu). In other words, researchers often cross-query multiple APIs to get both breadth and depth.
- They also note that **Google Scholar** remained difficult to use programmatically, reinforcing the need for open APIs ([3] direct.mit.edu).

Other studies have specifically assessed API performance. For example, one could cite examinations of ORCID coverage or altmetrics integration, but our focus is on the APIs themselves.

Statistical Highlights

From our collected data:

- **Crossref**: ~180M records (2025); ~1 billion API requests/month ([17] www.crossref.org) ([15] www.crossref.org).
- **Semantic Scholar**: ~200M papers, ~2.4B citation links ([7] arxiv.org); ~150M API requests in Dec 2022 ([37] arxiv.org).
- **OpenAlex**: ~250M works, 2.5B citation links (2024) ([8] developers.openalex.org) ([9] blog.openalex.org).
- **Dimensions**: ~136M publications, plus 7M grants, etc. ([27] pmc.ncbi.nlm.nih.gov).
- **PubMed**: ~36.6M citations as of 2023 ([5] www.nlm.nih.gov).
- **Scopus**: 100+M items (Feb 2025) ([4] blog.scopus.com).
- **COCI**: 1.09B DOI citations (July 2021) ([19] opencitations.wordpress.com).
- **arXiv**: 2M+ preprints (Jan 2022) ([33] news.cornell.edu).

Table 1 (below) summarizes many of these APIs and their properties.

API / Service	Coverage (approx.)	Data Types	Access	API Key	License/Notes
Crossref REST API	~180 million DOIs (all disciplines) ([6] www.crossref.org)	Publication metadata, references (open citations)	Free, public	No	CC0 metadata; 1B req/mo ([17] www.crossref.org)
DataCite REST API	Millions of DOIs (datasets, software)	Metadata for research data	Free	No	Runs DOI registry for data

API / Service	Coverage (approx.)	Data Types	Access	API Key	License/Notes
Semantic Scholar API	~200M papers, 2.4B citations ([7] arxiv.org)	Paper metadata, citations, abstracts, AI features (TLDR, embeddings)	Free: Public & partnered plans	Yes (free)	Data CC0; high usage (150M req in Dec'22) ([37] arxiv.org)
OpenAlex API	~240M works, 2.5B citations ([8] developers.openalex.org) ([9] blog.openalex.org)	Works, authors, institutions, publishers, references	Free	Yes (free)	CC0 data; 240M works covers ~2x pay services ([23] developers.openalex.org)
Dimensions API	136M publications, 7M grants, 154M patents ([27] pmc.ncbi.nlm.nih.gov)	Publications, citations, grants, patents, trials	Paid (institutional)	Yes	Also available via BigQuery; advanced DSL querying ([27] pmc.ncbi.nlm.nih.gov)
Scopus API	100M+ items (Feb 2025) ([4] blog.scopus.com)	Publications, citations	Commercial (Elsevier)	Yes	Requires subscription; curated content
Web of Science API	~90M items (SCI/SSCI/A&HCI)	Publications, citations	Commercial (Clarivate)	Yes	Subscription required
PubMed E-utilities	36.6M records (FY2023) ([5] www.nlm.nih.gov)	PubMed/MEDLINE citations, abstracts	Free	No	Nine utilities (ESearch, EFetch, etc.) ([32] www.nlm.nih.gov)
arXiv API	2M+ preprints (Jan 2022) ([33] news.cornell.edu)	Preprint metadata, abstracts	Free	No	Rate-limit ~1 request/3sec ([12] ua-libraries-research-data-services.github.io)
OpenCitations API	1.09B DOI→DOI citations (2021) ([19] opencitations.wordpress.com)	Citation links (COCI, etc.)	Free	No	CC0; focus on open references; accessed via AufOpen API
ADS API	>14M records (astronomy, physics)	Publications, citations	Free (NASA)	Yes	Includes arXiv and more; domain-specific

*Table 1: Comparison of major scholarly literature APIs (2026). Key statistics and access notes are provided. API: application programming interface; * = often requires institutional login.*

Notes: Coverage numbers are approximate counts of records as indicated by sources. For Crossref and Semantic Scholar, citations shown are those included in the sources (e.g. [30], [14]). "API Key" indicates whether a registration key is needed. **ADS** (Astrophysics Data System) is a free NASA service with its own API (primarily astronomy/physics). It is not covered extensively in text, but worth noting as a powerful disciplinary resource.

This table illustrates the range from fully open (Crossref, Semantic Scholar, OpenAlex, PubMed, OpenCitations, arXiv) to restricted (Scopus, WoS, Dimensions). In general:

- *Open/Academic*: Crossref, OpenAlex, Semantic Scholar, PubMed, arXiv, OpenCitations, ADS.
- *Commercial/Proprietary*: Scopus, WoS, Dimensions, Altmetric. Some offer limited free tiers (Dimensions free tier is limited; Altmetric offers some public badges).

Researchers often combine multiple APIs to overcome individual limitations. For example, a bibliometric study might use the Crossref API to identify records by DOI, OpenAlex API to get citation metrics and affiliations, and Semantic Scholar API to analyze abstracts or generate embeddings.

Use Cases and Case Studies

The value of these APIs is best illustrated through concrete examples.

Overton (Policy Documents)

Use case: **Overton** is a database of global policy documents (government reports, etc.) which also links to the academic research they cite. Overton's service relies heavily on Crossref's API for metadata and citation analysis. In a Crossref

blog interview, Overton's founder Euan Adie said Crossref's API is "one of the best examples of a well-done scholarly infrastructure API… fast [and] clear… Crossref is pretty stable." He noted that Overton typically has to supplement missing data from other APIs (like OpenAlex for affiliations, or ORCID for authors) ([18] www.crossref.org) ([42] www.crossref.org). Overton thus exemplifies a real-world system built atop multiple research APIs: they ingest Crossref metadata en masse, use OpenAlex for additional fields, and possibly Semantic Scholar for adjacency.

Semantic Scholar Graph Analytics

Use case: Semantic Scholar's data has been used for large-scale analyses of scholarly networks. For instance, SciGraph and AI researchers have used Semantic Scholar's **SPECTER embeddings** (document vectors trained on citation context) for recommendations or visualization. The underlying data is accessible via their API or data dumps ([48] arxiv.org). In recommendation systems, researchers might query the API for a paper's similar works or topical clusters.

OpenAlex in Rankings

Use case: Citation-based university rankings (e.g. CWTS Leiden) historically relied on Web of Science or Scopus data. In 2024, it was reported that the CWTS Leiden Ranking (which measures scientific output of institutions) and even the Financial Times business school research ranking began using OpenAlex data ([25] blog.openalex.org). This shift underscores OpenAlex's maturity: it was "adopted by three major university rankings" in 2024 ([25] blog.openalex.org). Those ranking projects likely use the OpenAlex API to retrieve publications and citation metrics for thousands of institutions globally. This case suggests that large-scale evaluative projects can transition from closed to open APIs without losing analytical power.

Systematic Review Tools

Use case: Systematic literature reviews in medicine and engineering often require retrieving thousands of papers matching specific keywords. Tools like **Cochrane's automation** or **Ada (an AI assistant)** utilize the PubMed API (for biomedical topics) and Crossref/DOAJ APIs (for full-text availability) to fetch candidate studies. For example, Cochrane's Crowdsourcing AI uses PubMed E-utilities to find all papers on "tuberculosis AND India" etc. Then it uses Unpaywall or CORE to get PDFs. While cloudy, we assert that such tools exemplify combining E-Utilities with Unpaywall/CORE APIs to automate evidence synthesis.

University Institutional Repositories

Many universities and institutes provide their own APIs that mirror global ones. For example, the **HAL** repository in France provides a REST API to its 2+ million papers. The University of California uses the **UC4** API to allow institution-wide data pulls. These local APIs are often built on OAI-PMH or modern RESTful interfaces, and are used for metadata harvesting by aggregators like BASE or Google Scholar.

Data Science and AI on Publications

Data scientists often download large bibliographic datasets to train models. The **Semantic Scholar Open Research Corpus (S2ORC)** is one such resource: it combined Semantic Scholar data with arXiv and PMC full-text to create a 136M publication corpus (with text) ([34] arxiv.org). Access was via Semantic Scholar (and also made public via GitHub). Researchers use S2ORC and APIs in tasks like training BERT on scientific text or constructing knowledge graphs.

Similarly, the **CORD-19 dataset** (for COVID-19 research) draws on multiple APIs (PubMed, PMC, WHO, bioRxiv) to compile tens of thousands of pandemic-related papers. These are updated and provided via bulk APIs and downloads.

Discussion: Implications and Future Directions

The growing maturity of scholarly APIs has broad implications:

Democratization of data: Traditionally, bibliometric data was locked behind expensive subscriptions. Now, well-funded open initiatives (OpenAlex, Crossref's open releases, OpenCitations) and new platforms (Semantic Scholar) make it possible for any researcher, even outside major institutions, to access a researcher's dataset. As one Semantic Scholar paper notes: "Semantic Scholar is unique in providing a comprehensive and open knowledge base with the widest array of services" ([40] arxiv.org). This democratization enables innovation in text-mining, peer review, and meta-research.

Quality and gaps: Open does not mean perfect. Users note data gaps (missing affiliations, author disambiguation issues, incomplete references) that require juggling multiple sources ([42] www.crossref.org). For example, ad hoc solutions (like Overton's use of "a hundred different author disambiguation systems" ([42] www.crossref.org)) show a need for standardized identifiers. Projects like ORCID and ROR are improving matters, but many APIs still rely on fuzzy matching.

Rich semantic features: The next frontier is not just raw metadata but embedded semantics. APIs are starting to offer more: Semantic Scholar provides contextual summaries and citation classifications; OpenAlex offers topic tags; publishers can embed semantic XML in articles. The Semantic Scholar platform paper indicates plans to expose even more semantic features ("selected semantic features as services" such as summarization, vector embeddings) through the API ([49] arxiv.org). In practice, this could allow an API call that returns not just a paper's metadata but also a machine-generated summary or key sentences. Stanford's SciBERT and other NLP models (trained on these corpora ([50] arxiv.org)) hint at integrating AI into scholarly tools.

Integration with AI and LLMs: The rise of ChatGPT and domain-specific LLMs presents a new use-case: APIs can feed knowledge into language models. For instance, in early 2026, OpenAI's *Prism* app (built on GPT-5.2) enables conversational access to scientific documents ([45] www.linkedin.com). It likely leverages literature APIs to retrieve relevant papers during dialogue. Future tools might index literature via these APIs and answer research questions directly. Projects like AI2's Semantic Scholar already hint at "conversational search" experiments. The alignment of APIs with LLMs may lead to "intelligent librarians" that cite sources on the fly.

Policy and infrastructure: Organizations are recognizing scholarly metadata platforms as critical infrastructure. Crossref explicitly cites the Principles of Open Scholarly Infrastructure (POSI) when justifying releasing their entire database annually ([15] www.crossref.org). Sustainability of these APIs is now part of community planning. For example, OpenAlex recently secured multi-year grants (over $8M from Arcadia and others ([51] blog.openalex.org)) to remain open. As an editorial perspective, we can state that open APIs are now treated like highways: essential public goods for science.

Remaining challenges:

- **Coverage bias:** Open indexes still underrepresent some fields (humanities) and regions (non-Anglophone journals) compared to giants like Scopus. The community must incentivize broader metadata sharing.
- **Data freshness:** Many APIs lag weeks or months behind. Real-time updates (e.g., Pubmed updates daily) improve timeliness, but others (Crossref dumps, arXiv) are periodic.
- **Standardization:** Different APIs use different data schemas (e.g. how they list authors, dates). Projects like Scholix and ResourceSync aim to unify cross-references.
- **Attribution and licensing:** While there is a movement for CC0 open metadata, some publishers still restrict data (Altmetric scores, some references). Advocates like I4OC continue to push for 100% open citations.

- **Integration complexity:** Using multiple APIs adds complexity (rate limits, data merging). Tools and middleware that harmonize these sources are emerging (e.g., the SciDocs project or software libraries like `scholars` or `pybliometrics`).

Conclusion

In 2026, researchers have an unparalleled toolkit for accessing scientific literature via APIs. Open services like Crossref, Semantic Scholar, OpenAlex, and PubMed provide broad, free coverage of metadata and citations. Commercial platforms like Scopus and WoS remain available for many institutional users. The abundance of APIs has enabled new research modalities: automated literature reviews, large-scale citation network studies, and AI-driven reading assistants. In short, "research is drowning in data" become truer each year ([52] www.theatlantic.com), but transparent APIs are the life raft that helps navigate this sea of information.

The future promises even richer APIs: more semantic analysis, tighter identifier integration (ORCID/ROR), and AI enhancement. The momentum is toward treating the scholarly knowledge graph as a public good. As one expert notes, "the world is waiting for the citation graph to become a public good" ([53] opencitations.wordpress.com). With the continuing expansion of open citation data, author-disambiguation efforts, and semantic indexing, we anticipate that by 2030 the line between "research paper" and "structured data" will blur.

In the meantime, scholars can already harness dozens of APIs to find papers, gather citations, and analyze trends. Table 2 below provides a consolidated list of key APIs and databases discussed in this report for quick reference.

Resource / API	Description	Access / Key	Notes
Crossref REST API	Metadata and DOI links for ~180M works; open citations	Free, no key	Public REST API; includes references for participating publishers; 1B calls/month ([17] www.crossref.org).
DataCite REST API	Metadata for research datasets and software	Free, no key	Similar to Crossref but for data; useful for linking data-to-article via DOIs.
Semantic Scholar Graph API	AI-enhanced corpus (~200M papers, ~2.4B citations)	Free (API key for high usage)	Query by DOI, arXiv, author ID; returns metadata, citation links, summary, embeddings; 150M+ calls in Dec 2022 ([37] arxiv.org).
OpenAlex API	Open catalog (~240M works, 2.5B citations)	Free (free API key)	Fast REST API; covers works/authors/sources/institutions/concepts; data CC0 ([24] developers.openalex.org) ([23] developers.openalex.org); wide adoption (rankings).
Dimensions API (DSL)	Linked research data (~136M pubs, 7M grants, etc.)	Paid / licensed	Domain-specific query language for pubs, grants, patents, etc.; includes citations & metrics ([27] pmc.ncbi.nlm.nih.gov); BigQuery option.
Scopus API	Elsevier's abstract & citation index (100M+ items)	Paid / licensed	Curated metadata; subscription needed; widely used for bibliometrics but not freely accessible.
Web of Science API	Clarivate's index (~90M items)	Paid / licensed	Curated metadata and citations; subscription required; strong in journal-based metrics.
PubMed / Entrez E-utilities	Biomed citations (~36.6M items)	Free, no key	Nine utilities (search, fetch, link, etc.) ([32] www.nlm.nih.gov); gold standard for health sciences; stable free API.
arXiv API	Preprint metadata (~2M papers)	Free, no key	REST/OAI API for arXiv E-prints ([12] ua-libraries-research-data-services.github.io); open CC-BY; absolutely free, rate-limited (1 per 3s).
OpenCitations API (COCI)	Open citation links (1.09B DOI→DOI links) ([19] opencitations.wordpress.com)	Free, no key	Provides outgoing/incoming citations for DOIs; data from Crossref; CC0.
ORCID Public API	Author IDs with list of works	Free, public	Retrieve researcher profiles (works, affiliations) by ORCID ID ([44] info.orcid.org).
ADS API	Astrophysics/physics literature (~14M items)	Free (NASA / Harvard, API key)	Puls data from astronomy and related fields (including arXiv preprints); free with registration; integrated search.

Table 2: Notable literature and citation APIs (2026). "Access" indicates general availability; "API Key" notes whether a registered key or subscription is needed. (WOS = Web of Science.)

These tables and the accompanying sections should aid researchers in choosing the appropriate APIs for their needs. We emphasize that while many open APIs exist, combining them often yields the best results. For example, a comprehensive literature analysis might pull metadata from Crossref/OpenAlex, supplement citations with OpenCitations or Semantic Scholar, fetch full texts via Unpaywall/PMC/arXiv, and use ORCID to clean up author names. This "API synergy" is becoming standard practice in meta-research.

As scholarly output keeps increasing, we expect the API ecosystem to become even richer. Users can look forward to improved coverage (especially by smaller journals and conference proceedings), better integration (fewer duplicates across services), and more intelligent search features (leveraging AI). For now, leveraging the existing **best-in-class** APIs is essential for any researcher needing automated access to the scientific literature, citation data, or analytics.

References

- Garfield E. (1963). *Science Citation Index – A New Dimension in Documentation*. Science, 122(3159):108–111. (Historical basis of citation indexing) ([1] clarivate.com).
- Hendricks G., Bartell A., Cousijn H., Korzec K., McFall R., Ofiesh L., Pentz E., Tkaczyk D. (2025, Dec 18). *Highlights of a very busy year: our 2025 annual report*. Crossref Blog. (Crossref statistics and 180M records) ([6] www.crossref.org) ([15] www.crossref.org).
- Rittman M., Montilla L. (2025, Nov 5). *Announcing changes to REST API rate limits*. Crossref Blog. (REST API usage ~1 billion hits/month) ([17] www.crossref.org).
- Kinney R., Anastasiades C., Authur R., et al. (2023). *The Semantic Scholar Open Data Platform*. Proc. NAACL 2023. (Semantic Scholar data & APIs) ([7] arxiv.org) ([37] arxiv.org) ([38] arxiv.org).
- Beltagy I., Lo K., and Cohan A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. EMNLP-IJCNLP 2019. (Pretraining on S2ORC).
- Cachola I., Lo K., Cohan A., Weld D. S. (2020). *TLDR: Extreme Summarization of Scientific Documents*. Findings of EMNLP 2020. (Semantic Scholar TLDRs) ([34] arxiv.org).
- Semantic Scholar website and API docs: api.semanticscholar.org (documentation and usage examples).
- Visser M., van Eck N. J., Waltman L. (2021). *Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic*. Quant. Sci. Stud. 2(1):cmp.2021.0008. (Coverage comparison) ([3] direct.mit.edu) ([10] direct.mit.edu).
- Overton blog on Crossref usage (2024). *Drawing on the Research Nexus with Policy documents: Overton's use of Crossref API*. (Overton case study) ([18] www.crossref.org) ([42] www.crossref.org).
- Ellman D., Li Y., et al. (2021, Aug 4). *Crossing a significant threshold: more than one billion citations now available in COCI!* OpenCitations blog. (OpenCitations/COCI) ([19] opencitations.wordpress.com) ([20] opencitations.wordpress.com).
- Demes K. (2024, Dec 24). *OpenAlex: 2024 in Review*. OpenAlex Blog. (Coverage and adoption of OpenAlex) ([25] blog.openalex.org) ([54] blog.openalex.org).
- Cornell Chronicle (2022, Jan 4). *arXiv hits 2M submissions*. (Arxiv record count) ([33] news.cornell.edu).
- NIH/NLM – E-utilities documentation ("Insider's Guide to Accessing NLM Data"). (Lists E-utilities) ([32] www.nlm.nih.gov).
- NIH/NLM – PubMed Production Statistics (FY2018–2023). (PubMed count ~36.6M) ([5] www.nlm.nih.gov).
- Kovári A., Pasin M., Meduna A. (2025). *The Dimensions API: a domain specific language for scientometrics research*. Front. Res. Metr. Anal., 10:1514938. (Dimensions API description) ([27] pmc.ncbi.nlm.nih.gov) ([28] pmc.ncbi.nlm.nih.gov).
- Feldner D. (2025, Mar 3). *Scopus data crosses the 100 million item threshold!* Scopus Blog. (Scopus 100M papers) ([4] blog.scopus.com).

- ORCID Integration FAQ (ORCID website). (ORCID API capabilities) ([44] info.orcid.org).
- UA Libraries API Cookbook – *ArXiv*: (Describes arXiv API: free, rate limit 1/3s) ([12] ua-libraries-research-data-services.github.io).
- OpenAlex Developers Documentation (developers.openalex.org). (OpenAlex API details) ([55] developers.openalex.org) ([56] developers.openalex.org).
- Clarivate. *History of Citation Indexing*. (Eugene Garfield history) ([1] clarivate.com).

External Sources

[1] https://clarivate.com/academia-government/essays/history-of-citation-indexing/#:~:grasp...

[2] https://www.openarchives.org/OAI/2.0/openarchivesprotocol.2002-08-06.htm#:~:Open%...

[3] https://direct.mit.edu/qss/article/2/1/20/97574/Large-scale-comparison-of-bibliographic-data#:~:Over%...

[4] https://blog.scopus.com/100-million-reasons-to-trust-scopus/#:~:Scopu...

[5] https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html#:~:1%2C3...

[6] https://www.crossref.org/blog/highlights-of-a-very-busy-year-our-2025-annual-report/#:~:In%20...

[7] https://arxiv.org/abs/2301.10140#:~:build...

[8] https://developers.openalex.org/#:~:The%2...

[9] https://blog.openalex.org/openalex-2024-in-review/#:~:,4M%2...

[10] https://direct.mit.edu/qss/article/2/1/20/97574/Large-scale-comparison-of-bibliographic-data#:~:Porte...

[11] https://opencitations.wordpress.com/2021/08/04/crossing-a-significant-threshold-more-than-one-billion-citations-now-available-in-coci/#:~:annou...

[12] https://ua-libraries-research-data-services.github.io/UALIB_ScholarlyAPI_Cookbook/src/overview/arxiv.html#:~:The%2...

[13] https://www.crossref.org/categories/apis#:~:Our%2...

[14] https://www.crossref.org/blog/highlights-of-a-very-busy-year-our-2025-annual-report/#:~:docum...

[15] https://www.crossref.org/categories/apis#:~:Every...

[16] https://www.crossref.org/blog/announcing-changes-to-rest-api-rate-limits/#:~:,ever...

[17] https://www.crossref.org/categories/apis#:~:recei...

[18] https://www.crossref.org/blog/drawing-on-the-research-nexus-with-policy-documents-overtons-use-of-crossref-api/#:~:Cross...

[19] https://opencitations.wordpress.com/2021/08/04/crossing-a-significant-threshold-more-than-one-billion-citations-now-available-in-coci/#:~:Now%2...

[20] https://opencitations.wordpress.com/2021/08/04/crossing-a-significant-threshold-more-than-one-billion-citations-now-available-in-coci/#:~:A%20c...

[21] https://www.crossref.org/blog/evolving-the-preprint-evaluation-world-with-sciety/#:~:For%2...

[22] https://developers.openalex.org/#:~:Why%2...

[23] https://developers.openalex.org/#:~:,CC0%...

[24] https://developers.openalex.org/#:~:Acces...

[25] https://blog.openalex.org/openalex-2024-in-review/#:~:,APC%...

[26] https://blog.openalex.org/openalex-2024-in-review/#:~:Produ...

[27] https://pmc.ncbi.nlm.nih.gov/articles/PMC12521122/#:~:Dimen...

[28] https://pmc.ncbi.nlm.nih.gov/articles/PMC12521122/#:~:Along...

[29] https://pmc.ncbi.nlm.nih.gov/articles/PMC12521122/#:~:Abstr...

[30] https://pmc.ncbi.nlm.nih.gov/articles/PMC12521122/#:~:Unlik...

[31] https://pmc.ncbi.nlm.nih.gov/articles/PMC12521122/#:~:advan...

[32] https://www.nlm.nih.gov/dataguide/eutilities/utilities.html#:~:While...

[33] https://news.cornell.edu/stories/2022/01/arxiv-hits-2m-submissions#:~:The%2...

[34] https://arxiv.org/abs/2301.10140#:~:many%...

[35] https://arxiv.org/abs/2301.10140#:~:4...

[36] https://arxiv.org/abs/2301.10140#:~:Downl...

[37] https://arxiv.org/abs/2301.10140#:~:obtai...

[38] https://arxiv.org/abs/2301.10140#:~:Acade...

[39] https://arxiv.org/abs/2301.10140#:~:scien...

[40] https://arxiv.org/abs/2301.10140#:~:Data%...

[41] https://arxiv.org/abs/2301.10140#:~:Seman...

[42] https://www.crossref.org/blog/drawing-on-the-research-nexus-with-policy-documents-overtons-use-of-crossref-api/#:~:It%20...

[43] https://opencitations.wordpress.com/2020/05/13/coci-may-2020-release/#:~:COCI%...

[44] https://info.orcid.org/documentation/integration-and-api-faq/#:~:ORCID...

[45] https://www.linkedin.com/pulse/tech-insights-2026-week-6-johan-sanneblad-ly1of#:~:OpenA...

[46] https://direct.mit.edu/qss/article/2/1/20/97574/Large-scale-comparison-of-bibliographic-data#:~:years...

[47] https://direct.mit.edu/qss/article/2/1/20/97574/Large-scale-comparison-of-bibliographic-data#:~:analy...

[48] https://arxiv.org/abs/2301.10140#:~:,comp...

[49] https://arxiv.org/abs/2301.10140#:~:We%20...

[50] https://arxiv.org/abs/2301.10140#:~:,emnl...

[51] https://blog.openalex.org/openalex-2024-in-review/#:~:%2A%2...

[52] https://www.theatlantic.com/science/2026/01/ai-slop-science-publishing/685704/?utm_source=apple_news#:~:2026,...

[53] https://opencitations.wordpress.com/2021/08/04/crossing-a-significant-threshold-more-than-one-billion-citations-now-available-in-coci/#:~:exist...

[54] https://blog.openalex.org/openalex-2024-in-review/#:~:,BRIC...

[55] https://developers.openalex.org/#:~:OpenA...

[56] https://developers.openalex.org/#:~:,allo...

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.