

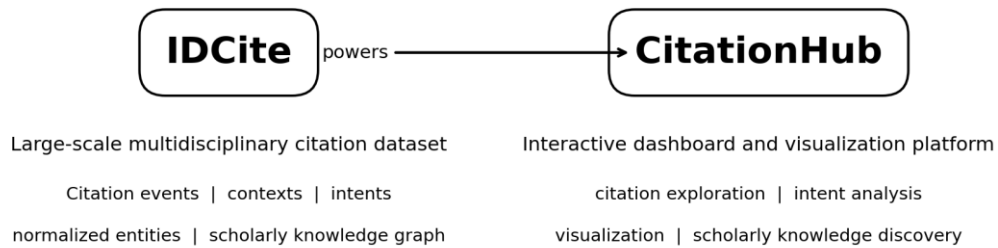
IDCite

A Large-Scale Multidisciplinary Citation Intent Dataset for Scholarly Knowledge Discovery

Project and Dataset Documentation

IDCite and CitationHub

Data foundation and interactive scholarly exploration system



IDCite provides the structured data layer; CitationHub provides the user-facing exploration and visualization layer.

Seohyun Nam

Department of Computer Engineering

Chung-Ang University
Seoul, Republic of Korea

Seohyun0315@cau.ac.kr

Zenodo DOI: 10.5281/zenodo.20796923
Version 3 documentation – August 11, 2026

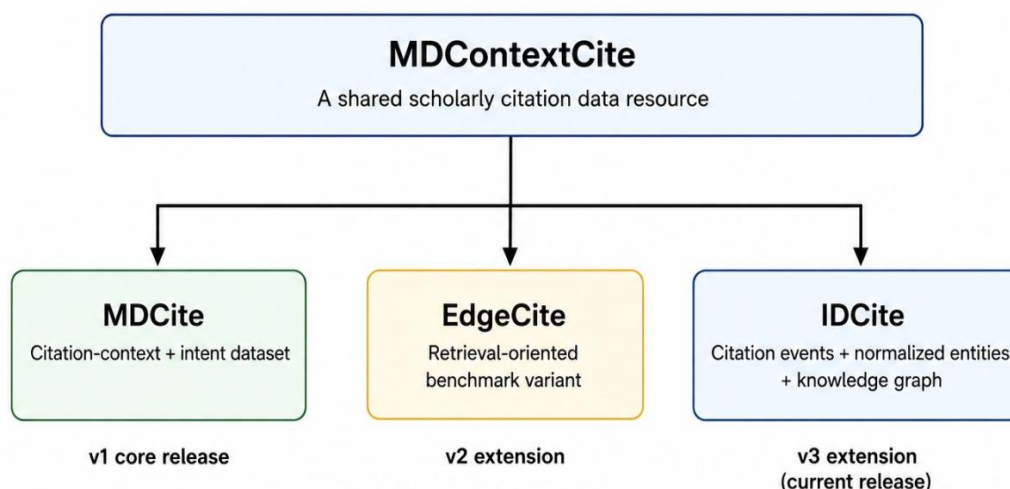
1. Executive Summary

IDCite is a large-scale multidisciplinary citation-intent dataset designed for scholarly knowledge discovery. The dataset represents citations at the citation-event level and integrates citation contexts, citation-intent annotations, bibliographic metadata, disciplinary information, normalized scholarly entities, and a supplementary ontology-ready knowledge graph.

IDCite contains 1,857,503 citation events linking 1,467,045 citing papers to 23,479 highly cited seed papers across 21 Essential Science Indicators (ESI) fields. The released resource preserves the scale, imbalance, and disciplinary heterogeneity of real-world scholarly citations. CitationHub is a separate interactive dashboard and visualization platform built on top of IDCite for exploring citation relationships, intents, disciplines, and graph-structured scholarly entities.

Version 3 scope

- Establishes IDCite as the primary Version 3 dataset release on Zenodo.
- Adds normalized citation-event tables, scholarly entity mappings, and a supplementary knowledge graph.
- Adds consolidated IDCite project and dataset documentation, including the complete 105-journal sampling frame and construction-pipeline overview.
- Clarifies the relationship between IDCite as the data layer and CitationHub as the interactive exploration and visualization system.



2. Relationship to MDContextCite and Release Lineage

Resource / release	Role in the resource evolution	Primary representation	Position
MDCite (Ver.1)	Core citation-context dataset released in the initial Zenodo version	Citation contexts and citation intents	Initial dataset release
MDContextCite (Ver.2)	A shared scholarly citation database and research resource containing MDCite and the retrieval-oriented EdgeCite variant	Citation-context database + retrieval dissemination variant	Shared citation data resource
IDCite (Ver.3)	New structured multidisciplinary dataset extending the citation resources with citation events,	Citation events + normalized entities + knowledge graph	Current dataset release

	normalized scholarly entities, and a supplementary scholarly knowledge graph		
CitationHub	Interactive dashboard and visualization platform powered by IDCite; not a Zenodo dataset version	Interactive system / application layer	Companion system

IDCite builds upon the scholarly citation data resources previously released through MDContextCite, but it should not be interpreted as a simple renaming of the previous resource.

MDContextCite was established as a large-scale, multidisciplinary citation context database and a shared scholarly research resource designed to support citation-aware scholarly information retrieval, citation intent analysis, and citation-based retrieval benchmarking. The framework was built around citation contexts - local textual spans surrounding in-text citations - as the primary unit of analysis.

Within MDContextCite, MDCite served as the core citation-context dataset. EdgeCite was subsequently provided as an additional retrieval-oriented dissemination variant created for leakage-controlled, large-scale citation retrieval experiments. Thus, MDContextCite represented the broader database and dissemination framework, while MDCite and EdgeCite served distinct research purposes within that framework.

IDCite represents a further development of these scholarly citation data resources. Rather than serving as another retrieval-oriented variant, IDCite reorganizes and extends the underlying citation information into a structured multidisciplinary dataset centered on citation events. The release links citing papers and highly cited seed papers while integrating citation contexts, citation-intent annotations, publication metadata, disciplinary information, and normalized scholarly entities.

IDCite additionally provides normalized representations of authors, affiliations, journals, scientific fields, cities, countries, and citation intents, together with a supplementary ontology-ready scholarly knowledge graph. This enables the resource to support citation-event-level analysis, entity-oriented analysis, graph analytics, and scholarly knowledge discovery in addition to citation-context and retrieval research.

Accordingly, Version 3 introduces IDCite as a new structured dataset release built upon the broader citation-data collection effort represented by the preceding releases. The previously released MDCite and EdgeCite resources remain accessible through the Zenodo version history and provide provenance for the evolution of the resource.

CitationHub is separate from this Zenodo release lineage. It is the interactive dashboard and visualization platform built on IDCite, providing a user-facing environment for exploring citation relationships, citation contexts, citation intents, disciplinary patterns, and structured scholarly entities.

2.1 Contribution and Research Scope

The principal contribution of IDCite is the integration of citation relationships with semantic evidence about how and why prior work is cited. Large scholarly infrastructures such as OpenAlex, Crossref, OpenCitations, and Semantic Scholar are essential for bibliographic and citation-link discovery, but IDCite is designed as a reusable research dataset in which citation events are jointly represented with citation contexts, citation-intent annotations, disciplinary information, and normalized scholarly entities.

This design supports research questions that are difficult to address from citation-network structure alone, including how citation purposes vary across disciplines, whether different citation intents exhibit different retrieval or recommendation behavior, how semantic citation roles relate to scholarly entities, and how contextual and structural signals can be combined for scholarly knowledge discovery.

The released resource is scoped to citation relationships among scholarly papers represented in the upstream scholarly infrastructures and the selected seed-paper corpus. IDCite is not intended as a general-purpose registry of citations to datasets, software, policy documents, or other non-paper research objects.

3. Data Sources and Construction Pipeline

IDCite integrates bibliographic, disciplinary, citation-link, and citation-context information from multiple scholarly infrastructures. The major source roles are summarized below.

Source	Role in the pipeline
Scopus	Bibliographic metadata, journal-level publication records, citation counts, and seed-paper selection support.
Web of Science / 2024 JCR	Field-aware journal grouping and selection of five Q1 journals within one representative category for each ESI field.
OpenAlex	Citation-link retrieval, DOI/title resolution, and structured metadata interoperability.
Semantic Scholar	Citation contexts and citation-linked scholarly information used for semantic enrichment.
SynIntent	Automatic citation-intent inference used to generate semantic citation-intent annotations at scale.

Construction pipeline

1. Journal-stratified sampling across 21 ESI fields and 21 representative Web of Science categories.
2. Selection of five representative Q1 journals per category (105 journals in total).
3. Independent selection of the top 5% most-cited papers within each journal to form the seed-paper corpus.
4. Citation-link harvesting and citation-event construction between citing papers and seed papers.
5. Citation-context collection and automated citation-intent annotation.
6. Normalization of publication, journal, author, affiliation, geographic, field, and intent entities.
7. Conversion of normalized entities and relations into an ontology-ready scholarly knowledge graph.

Sampling Rationale and Design Considerations

Broad disciplinary coverage is a deliberate design choice. Citation practices, terminology, publication rates, and citation density vary substantially across fields; a dataset concentrated in only a small number of domains would make it difficult to study whether citation semantics and downstream model behavior generalize across disciplines. IDCite therefore spans 21 ESI fields and uses journal-stratified sampling to preserve disciplinary heterogeneity rather than imposing an artificially balanced corpus.

Highly cited papers were selected as seed targets because they provide sufficiently dense citation neighborhoods for large-scale citation-context collection and enable repeated observation of how influential scholarly works are used across papers and fields. The top-5% threshold is applied independently within each selected journal so that citation-intensive disciplines do not determine the seed set solely through global citation counts.

This design also introduces an important scope limitation: IDCite characterizes citation behavior around highly cited seed papers and should not be assumed to represent citation behavior around moderately cited or rarely cited publications. Uneven seed-paper counts across fields are retained rather than resampled away; downstream cross-field comparisons should therefore use field-aware normalization, stratification, weighting, or sensitivity analyses where appropriate.

3.1 IDCite Construction Pipeline Overview

Figure 1 provides an overview of the IDCite construction workflow, linking the primary input sources, processing stages, and released dataset components. The pipeline summarizes multidisciplinary seed-paper selection, citation harvesting, citation-event construction, semantic citation-intent enrichment, entity normalization, and supplementary knowledge graph conversion.

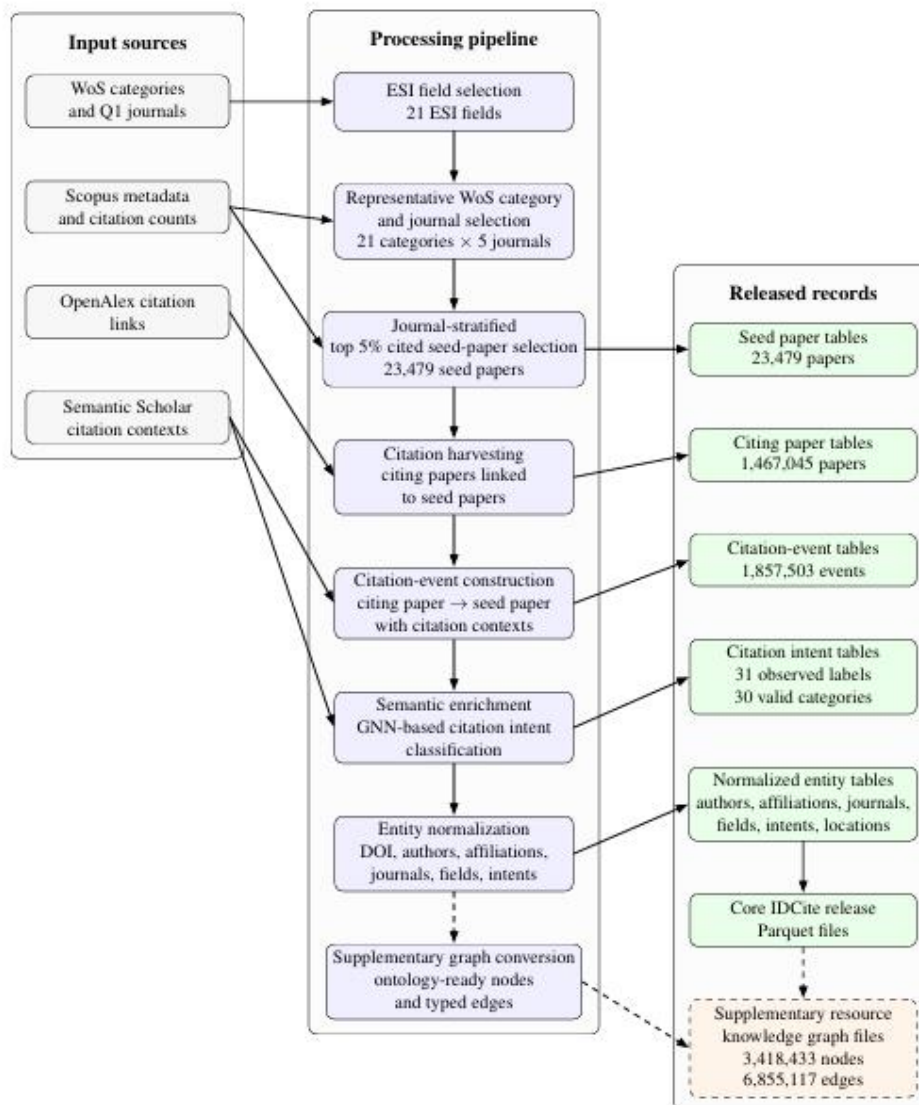


Figure 1. Overview of the IDCite construction workflow. The core release consists of citation-context, citation-intent, publication, and normalized entity tables, with an additional knowledge graph representation provided as a supplementary resource.

3.2 Citation Context and Intent Classification

Citation contexts are retrieved through the Semantic Scholar Graph API as reference-linked textual spans associated with citing papers. IDCite preserves these context spans and provenance information when constructing citation events. The repository code provides the collection and integration workflow used to connect seed papers, citing papers, citation links, and available citation-context evidence.

Citation-intent labels are generated automatically at scale using the SynIntent classifier and are released as weak semantic supervision rather than manually curated gold-standard annotations. SynIntent was trained in a supervised manner on the manually annotated MultiCite benchmark, which contains 12,653 manually annotated citation contexts from 1,193 scientific publications. On this benchmark, SynIntent achieved a Strict Accuracy of 0.67

and a Weak Accuracy of 0.80, providing empirical evidence for its use in large-scale automatic citation-intent annotation.

Accordingly, the citation-intent labels provided in IDCite should be interpreted as an operationalized measurement derived from automated inference, rather than as manually curated ground-truth annotations. This distinction is important when using IDCite for citation-intent analysis, particularly in applications requiring fine-grained interpretation of individual citation events.

4. IDCite Dataset Overview

IDCite is a large-scale multidisciplinary dataset of citation contexts and citation intents spanning 21 ESI fields. It is organized around citation events that connect a citing paper with a highly cited seed paper while preserving semantic and bibliographic evidence. This event-centric representation supports analyses that require both local citation semantics and structured scholarly relationships.

Measure	Value
Citation events	1,857,503
Citing papers	1,467,045
Seed papers	23,479
ESI fields	21
Observed intent labels	31
Knowledge graph nodes	3,418,433
Knowledge graph edges	6,855,117

Intended research uses

- Citation context analysis
- Citation intent analysis and classification
- Scientometric and bibliometric research
- Scholarly information retrieval
- Citation recommendation
- Research assessment
- Cross-disciplinary citation analysis
- Knowledge graph analytics and network analysis

5. IDCite File Inventory

File	Rows	Cols	Description
citation_events.parquet	1,857,503	20	Raw citation event records
citation_events_enriched.parquet	1,857,503	32	Citation events enriched with cited seed-paper metadata
citation_events_normalized.parquet	1,857,503	23	Citation events with normalized intent and field identifiers
citing_papers.parquet	1,467,045	7	Metadata of citing papers
citing_papers_normalized.parquet	1,467,045	8	Citing papers with normalized journal IDs
seed_cited_papers.parquet	23,479	42	Highly cited seed-paper metadata
seed_cited_papers_normalized.parquet	23,479	48	Seed papers with normalized journal, author, affiliation, city, country, and field IDs
kg_edges.parquet	6,855,117	3	Typed scholarly knowledge graph edges
kg_nodes.parquet	3,418,433	14	Scholarly knowledge graph nodes
authors.parquet	16,839	2	Author ID-name mapping
affiliations.parquet	5,271	2	Affiliation ID-name mapping
affiliation_geo.parquet	5,352	6	Affiliation, city, and country mappings
journals.parquet	46,237	2	Journal ID-name mapping
fields.parquet	21	3	Field ID-name mapping
intents.parquet	31	2	Citation intent ID-name mapping
cities.parquet	1,899	2	City ID-name mapping
countries.parquet	108	2	Country ID-name mapping

6. Core Schemas

Knowledge Graph

Citation Event Schema

Each CitationEvent connects a CitingPaper to a SeedPaper with a citation intent label. Seed papers are linked to journals, authors, affiliations, cities, countries, and research fields.

```
graph TD; SeedPaper((SeedPaper)) ---|publishedInJournal| Journal((Journal)); SeedPaper ---|locatedInCity| City((City)); SeedPaper ---|hasAuthor| Author((Author)); SeedPaper ---|locatedInCountry| Country((Country)); SeedPaper ---|hasAffiliation| Affiliation((Affiliation)); SeedPaper ---|belongsToField| Field((Field)); SeedPaper -- hasCitedPaper --> CitationEvent((CitationEvent)); CitationEvent -- hasCitingPaper --> CitingPaper((CitingPaper)); CitationEvent -- hasPrimaryIntent --> Intent((Intent));
```

Edge Types

event	→ citing	hasCitingPaper
event	→ seed	hasCitedPaper
event	→ intent	hasPrimaryIntent
seed	→ journal	publishedInJournal
seed	→ author	hasAuthor
seed	→ affiliation	hasAffiliation
seed	→ city	locatedInCity
seed	→ country	locatedInCountry
seed	→ field	belongsToField

Node Types & Attributes

SeedPaper

doi · title · journal · author · affiliation · country · field · citedby_count

CitationEvent

event_id · citing_year · primary_intent · context · is_influential

CitingPaper

doi · title · year · venue

Intent

background · uses · similarities · motivation · differences · future_work · extends

Journal

journal_name

Author

author_name · author_id

Affiliation

affiliation_name

City

city_name

Country

country_name

Field

field_name

6.1 Citation events

Field	Description
citation_event_id	Unique citation-event identifier
citing_paper_id	Identifier of the citing paper
cited_seed_paper_id	Identifier of the cited seed paper
citing_doi	DOI of the citing paper
citing_title	Title of the citing paper
citing_year	Publication year of the citing paper
citing_venue	Venue of the citing paper
primary_intent	Primary citation-intent label
all_intents	List of citation intents associated with the event
contexts	Citation-context text snippets
context_count	Number of citation contexts
intent_count	Number of intent annotations
is_influential	Whether the citation is marked as influential
has_semantic_evidence	Whether semantic evidence is available
field_id	Normalized field identifier

6.2 Seed cited papers

Field	Description
seed_paper_id	Unique seed-paper identifier
doi	DOI of the seed paper
title	Title of the seed paper
publication_name	Publication venue
creator	Main creator or first author
citedby_count	Citation count
group	Broad research group
category	Research category
journal_clarivate	Clarivate journal name
is_seed_top5pct	Whether the paper belongs to the selected highly cited group

7. Knowledge Graph Representation

IDCite includes a heterogeneous scholarly knowledge graph derived from the normalized citation-event infrastructure. Nodes represent citation events, publications, authors, affiliations, journals, cities, countries, scientific fields, and citation intents. Typed edges encode citation and entity relationships such as authorship, affiliation, publication venue, geographic association, disciplinary assignment, and citation-event linkage.

Component	Scale	Representative fields
kg_nodes.parquet	3,418,433 nodes	node_id, label, node_type, doi, publication_name, group, citedby_count, venue, year, primary_intent
kg_edges.parquet	6,855,117 edges	source, target, edge_type

The graph files are intended as a supplementary representation for graph analytics and network-oriented research. They do not replace the tabular citation-event files; rather, they expose normalized relationships in a form suited to graph representation learning and scholarly knowledge discovery.

7.1 What “Ontology-ready” Means

In IDCite, “ontology-ready” means that major scholarly entities are represented with normalized dataset-level identifiers and connected through explicit typed relations, so that the released tables can be mapped to, or extended with, a formal ontology without reconstructing entity relationships from free text. The release itself should therefore be understood as a structured graph-ready representation rather than a claim that a single external ontology has already been imposed on every entity.

This representation enables analyses that are not available from a paper-to-paper citation edge alone, including traversals from citation events to intents, journals, authors, affiliations, fields, and geographic entities; entity-level aggregation; heterogeneous graph representation learning; link prediction; and analysis of disciplinary or institutional knowledge flows.

7.2 Entity Normalization

IDCite provides normalized lookup tables for authors, affiliations, journals, fields, citation intents, cities, and countries. Normalization is used to replace repeated textual values with dataset-level identifiers and to support consistent joins across citation-event, publication, and graph tables. The normalized files should be treated as dataset-level harmonization artifacts rather than universal author- or institution-disambiguation authorities.

Where upstream identifiers or metadata are unavailable, users should retain the original provenance fields and interpret unmatched or missing values conservatively. The technical-validation workflow checks entity-table uniqueness and referential integrity so that normalization errors that would break dataset-level joins can be detected reproducibly.

8. Multidisciplinary Sampling Design

To reduce the dominance of citation-intensive disciplines, seed papers were selected through a journal-stratified design. One representative Web of Science category was selected for each of the 21 ESI fields, five journals were selected within each category, and the top 5% most-cited papers were retained independently within each journal. The resulting seed-paper counts vary by field because publication volume and citation distributions differ across disciplines. Figure 2 illustrates the resulting distribution of selected seed papers across the 21 ESI fields, highlighting the multidisciplinary coverage of IDCite as well as the field-level variation produced by the journal-stratified sampling strategy.

8.1 Journal Selection

Mapping between ESI fields and representative Web of Science categories used for seed-paper acquisition. Five representative journals were selected from each category.

ESI Field	Selected WoS Category	Q1 Journals
Agricultural Sciences	Agriculture, Dairy & Animal Science	5
Arts & Humanities, Interdisciplinary	Area Studies	5
Biology & Biochemistry	Biotechnology & Applied Microbiology	5
Chemistry	Chemistry, Physical	5
Clinical Medicine	Medicine, General & Internal	5
Computer Science	Computer Science, Artificial Intelligence	5
Economics & Business	Economics	5
Engineering	Engineering, Industrial	5
Environment/Ecology	Ecology	5
Geosciences	Geography, Physical	5
History & Archaeology	History	5
Literature & Language	Language & Linguistics	5
Materials Science	Materials Science, Ceramics	5
Mathematics	Mathematics	5
Multidisciplinary	Multidisciplinary Sciences	5
Philosophy & Religion	Philosophy	5
Physics	Physics, Applied	5
Plant & Animal Science	Plant Sciences	5
Psychiatry/Psychology	Psychology	5
Social Sciences, General	Sociology	5
Visual & Performing Arts	Film, Radio, Television	5
Total	21 Categories	105

8.2 Seed-Paper Distribution

ESI field	Seed papers
Agricultural Sciences	1412
Arts & Humanities	990
Biology & Biochemistry	440
Chemistry	2896
Clinical Medicine	2076
Computer Science	475
Economics & Business	843
Engineering	999
Environment/Ecology	561
Geosciences	410
History & Archaeology	134
Literature & Language	468
Materials Science	3686
Mathematics	876
Multidisciplinary	6516
Philosophy & Religion	165
Physics	3325
Plant & Animal Science	1936
Psychiatry/Psychology	150

Social Sciences, General	227
Visual & Performing Arts	112

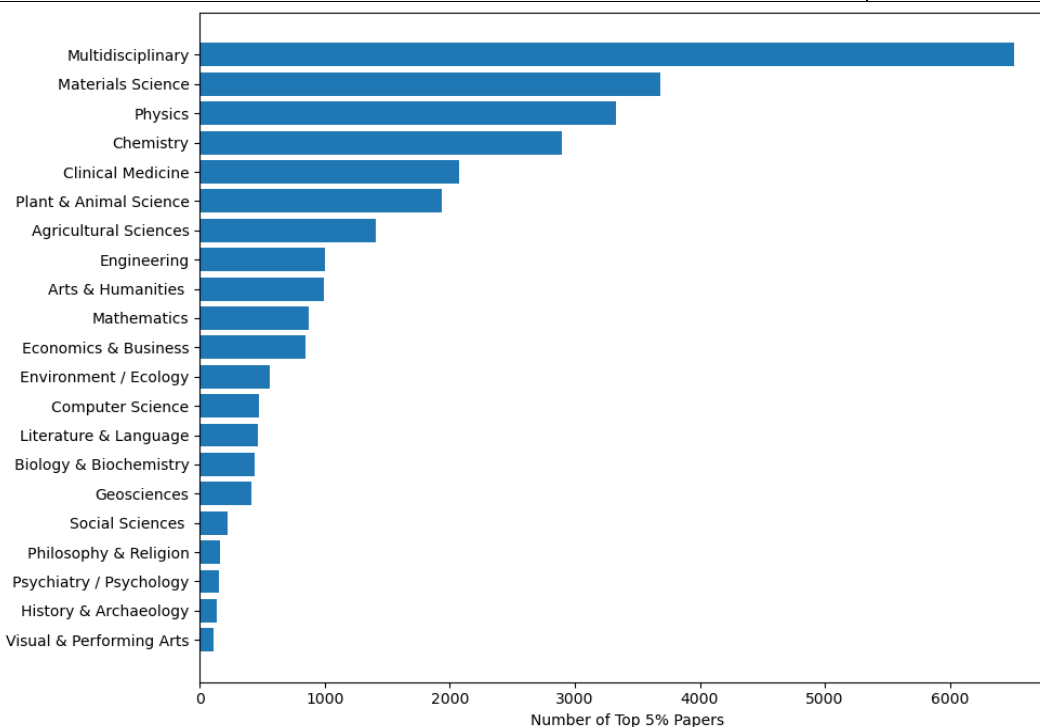
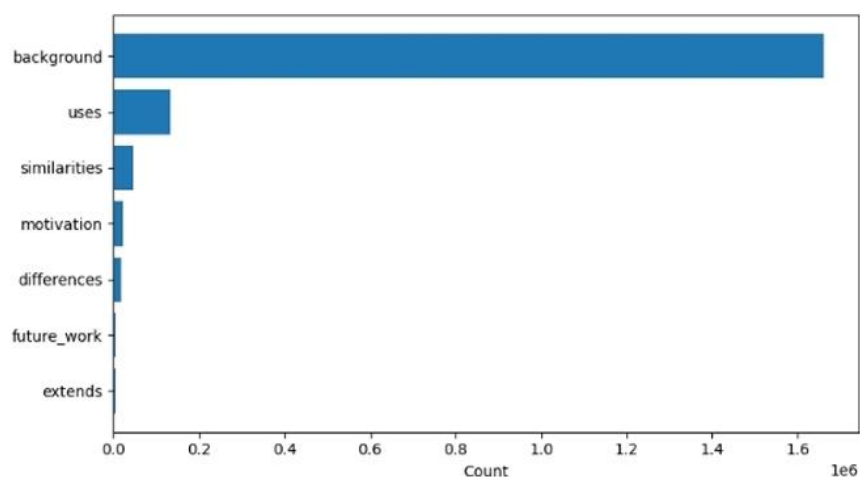


Figure 2. Distribution of the top 5% most-cited seed papers across the 21 ESI fields.

8.3 Citation Intent Distribution

Distribution of the seven canonical citation intents in IDCite. Composite intent labels were decomposed into their constituent intent categories, allowing a citation event to contribute to multiple semantic intent classes.

Intent	Count	Percentage (%)
background	1,660,899	88.20
uses	131,827	7.00
similarities	45,753	2.43
motivation	21,555	1.14
differences	15,617	0.83
future work	4,654	0.25
extends	2,813	0.15



8.4 Technical Validation and Fitness for Use

The released technical-validation workflow evaluates metadata completeness, citation-event referential integrity, citation-intent distribution and coverage, knowledge-graph integrity, and entity-table uniqueness. These checks are intended to verify the structural consistency and usability of the released files.

Structural validation should not be interpreted as equivalent to human validation of every automatically generated citation-intent label. Because the intent annotations are weak semantic supervision, users should distinguish dataset-integrity checks from annotation-quality evaluation. For applications that require gold-standard labels, manual evaluation or comparison against an appropriate externally annotated benchmark is recommended.

9. Reproducibility, Responsible Use, and Licensing

Reproducibility

The released 17 Parquet files, normalized lookup tables, and graph files are designed to support direct reuse without reconstructing the full upstream pipeline. Full reconstruction may require institutional access to Scopus as well as access to public scholarly APIs. API credentials are not distributed with the dataset.

Annotation interpretation

Citation-intent labels are generated automatically at scale rather than manually curated for every citation event. They should therefore be interpreted as model-derived semantic annotations suitable for large-scale analysis, retrieval, and machine-learning research, rather than as manually verified ground truth for every individual record.

Ethical considerations

The resource is intended for scholarly analysis and research support. Citation-based indicators should not be used as the sole basis for evaluating individual researchers, institutions, journals, research fields, or countries. Users should consider field-specific citation practices, metadata coverage, automated annotation uncertainty, and the snapshot nature of external scholarly databases when interpreting results.

License

License: Creative Commons Attribution 4.0 International (CC BY 4.0), subject to the terms and conditions associated with the underlying source services and metadata providers.

9.1 Upstream Data and Redistribution

IDCite integrates information obtained through multiple scholarly services. The CC BY 4.0 license applies to the IDCite release as stated by the repository, while access to and reuse of upstream-derived elements remain subject to the applicable terms, licenses, and conditions of the originating services. Users who reconstruct or redistribute upstream content should independently verify the current terms of Scopus, OpenAlex, Semantic Scholar, Web of Science, and Journal Citation Reports.

API credentials and proprietary database access are not distributed with IDCite. Because upstream scholarly infrastructures are continuously updated, IDCite should be interpreted as a snapshot corresponding to the November 2025 collection period.

9.2 Code and Persistent Reproducibility

The construction and validation code is maintained in the MDCite GitHub repository, while CitationHub is maintained separately as the interactive dashboard and visualization platform powered by IDCite. Version 3 documentation and archived Zenodo files provide a persistent reference point for the released data snapshot; GitHub repositories provide the actively maintained software layer.

10. Access, Citation, and Versioning

Resource	Identifier / location
Zenodo concept DOI	10.5281/zenodo.18410049
Version 1 DOI (MDCite)	10.5281/zenodo.18410050
Version 2 DOI (MDContextCite / EdgeCite)	10.5281/zenodo.18536895
Version 3 DOI (IDCite)	10.5281/zenodo.20796923
IDCite processed dataset mirror	https://huggingface.co/datasets/Daniel0315/IDCite
IDCite construction pipeline	https://github.com/kecau/MDCite
CitationHub interactive system	https://citation-hub-website.vercel.app

IDCite citation

```
@dataset{idcite2026,  
  title={IDCite: A Large-Scale Multidisciplinary Citation Intent Dataset for Scholarly Knowledge Discovery},  
  author={Nam, Seohyun},  
  year={2026},  
  publisher={Zenodo},  
  doi={10.5281/zenodo.20796923}  
}
```

Version 3 release note

Version 3 introduces IDCite as a new structured multidisciplinary dataset release built upon the scholarly citation resources developed through the preceding Zenodo releases. IDCite is not a simple rename of MDContextCite: MDContextCite served as the broader citation-context database and dissemination framework for MDCite and EdgeCite, whereas IDCite reorganizes and extends the resource around citation events, normalized scholarly entities, and a supplementary ontology-ready scholarly knowledge graph. The Version 3 title therefore reflects the expanded data representation and broader analytical scope of the current release. Previous MDCite and EdgeCite releases remain accessible through the Zenodo version history. CitationHub remains a separate interactive dashboard and visualization platform powered by IDCite.

Appendix A. Complete Journal Sampling Frame

Supplementary Table A1 lists the 105 journals used for seed-paper acquisition based on the 2024 Journal Citation Reports (JCR). Five journals were selected for each of 21 representative Web of Science categories, one category per ESI field.

ESI Field	WoS Category	Journal Name	Edition
Agricultural Sciences	Agriculture, Dairy & Animal Science	Annual Review of Animal Biosciences	SCIE
Agricultural Sciences	Agriculture, Dairy & Animal Science	Animal Nutrition	SCIE
Agricultural Sciences	Agriculture, Dairy & Animal Science	Journal of Animal Science and Biotechnology	SCIE
Agricultural Sciences	Agriculture, Dairy & Animal Science	Journal of Dairy Science	SCIE
Agricultural Sciences	Agriculture, Dairy & Animal Science	Animal Frontiers	SCIE
Arts & Humanities, Interdisciplinary	Area Studies	European Security	SSCI
Arts & Humanities, Interdisciplinary	Area Studies	Journal of Chinese Political Science	SSCI
Arts & Humanities, Interdisciplinary	Area Studies	China Journal	SSCI
Arts & Humanities, Interdisciplinary	Area Studies	East European Politics	SSCI
Arts & Humanities, Interdisciplinary	Area Studies	Post-Soviet Affairs	SSCI
Biology & Biochemistry	Biotechnology & Applied Microbiology	Nature Reviews Drug Discovery	SCIE
Biology & Biochemistry	Biotechnology & Applied Microbiology	Nature Biotechnology	SCIE
Biology & Biochemistry	Biotechnology & Applied Microbiology	Trends in Biotechnology	SCIE
Biology & Biochemistry	Biotechnology & Applied Microbiology	Journal of Nanobiotechnology	SCIE
Biology & Biochemistry	Biotechnology & Applied Microbiology	Biotechnology Advances	SCIE
Chemistry	Chemistry, Physical	Nature Catalysis	SCIE
Chemistry	Chemistry, Physical	Nature Materials	SCIE
Chemistry	Chemistry, Physical	Joule	SCIE
Chemistry	Chemistry, Physical	Interdisciplinary Materials	ESCI
Chemistry	Chemistry, Physical	Advanced Materials	SCIE
Clinical Medicine	Medicine, General & Internal	Lancet	SCIE
Clinical Medicine	Medicine, General & Internal	New England Journal of Medicine	SCIE
Clinical Medicine	Medicine, General & Internal	Nature Reviews Disease Primers	SCIE
Clinical Medicine	Medicine, General & Internal	JAMA-Journal of the American Medical Association	SCIE

Clinical Medicine	Medicine, General & Internal	BMJ-British Medical Journal	SCIE
Computer Science	Computer Science, Artificial Intelligence	Foundations and Trends in Machine Learning	ESCI
Computer Science	Computer Science, Artificial Intelligence	Nature Machine Intelligence	SCIE
Computer Science	Computer Science, Artificial Intelligence	IEEE Transactions on Pattern Analysis and Machine Intelligence	SCIE
Computer Science	Computer Science, Artificial Intelligence	Information Fusion	SCIE
Computer Science	Computer Science, Artificial Intelligence	AI Open	ESCI
Economics & Business	Economics	Energy Economics	SSCI
Economics & Business	Economics	Quarterly Journal of Economics	SSCI
Economics & Business	Economics	Journal of Financial Economics	SSCI
Economics & Business	Economics	American Economic Review	SSCI
Economics & Business	Economics	Annual Review of Economics	SSCI
Engineering	Engineering, Industrial	Journal of Manufacturing Systems	SCIE
Engineering	Engineering, Industrial	Journal of Industrial Information Integration	SCIE
Engineering	Engineering, Industrial	Reliability Engineering & System Safety	SCIE
Engineering	Engineering, Industrial	Technovation	SCIE
Engineering	Engineering, Industrial	International Journal of Production Economics	SCIE
Environment/Ecology	Ecology	Trends in Ecology & Evolution	SCIE
Environment/Ecology	Ecology	Nature Ecology & Evolution	SCIE
Environment/Ecology	Ecology	Global Change Biology	SCIE
Environment/Ecology	Ecology	Annual Review of Ecology Evolution and Systematics	SCIE
Environment/Ecology	Ecology	ISME Journal	SCIE
Geosciences	Geography, Physical	ISPRS Journal of Photogrammetry and Remote Sensing	SCIE
Geosciences	Geography, Physical	Landscape and Urban Planning	SCIE
Geosciences	Geography, Physical	Geography and Sustainability	ESCI
Geosciences	Geography, Physical	GIScience & Remote Sensing	SCIE
Geosciences	Geography, Physical	Journal of Geovisualization and Spatial Analysis	ESCI
History & Archaeology	History	Journal of Economic History	AHCI
History & Archaeology	History	Past & Present	AHCI, SSCI
History & Archaeology	History	Nations and Nationalism	SSCI

History & Archaeology	History	Twentieth Century British History	AHCI
History & Archaeology	History	Postcolonial Studies	AHCI, SSCI
Literature & Language	Language & Linguistics	Transactions of the Association for Computational Linguistics	AHCI
Literature & Language	Language & Linguistics	Computer Assisted Language Learning	AHCI
Literature & Language	Language & Linguistics	ReCALL	AHCI
Literature & Language	Language & Linguistics	Computational Linguistics	AHCI
Literature & Language	Language & Linguistics	Language Teaching	AHCI
Materials Science	Materials Science, Ceramics	Journal of Advanced Ceramics	SCIE
Materials Science	Materials Science, Ceramics	Journal of the European Ceramic Society	SCIE
Materials Science	Materials Science, Ceramics	Ceramics International	SCIE
Materials Science	Materials Science, Ceramics	Journal of the American Ceramic Society	SCIE
Materials Science	Materials Science, Ceramics	Journal of the Korean Ceramic Society	SCIE
Mathematics	Mathematics	Acta Numerica	SCIE
Mathematics	Mathematics	Acta Mathematica	SCIE
Mathematics	Mathematics	Annals of Mathematics	SCIE
Mathematics	Mathematics	Journal of Numerical Mathematics	SCIE
Mathematics	Mathematics	Advances in Nonlinear Analysis	SCIE
Multidisciplinary	Multidisciplinary Sciences	Nature Reviews Methods Primers	SCIE
Multidisciplinary	Multidisciplinary Sciences	Nature	SCIE
Multidisciplinary	Multidisciplinary Sciences	Science	SCIE
Multidisciplinary	Multidisciplinary Sciences	Innovation	ESCI
Multidisciplinary	Multidisciplinary Sciences	Exploration	ESCI
Philosophy & Religion	Philosophy	Ethics and Information Technology	AHCI
Philosophy & Religion	Philosophy	Journal of Critical Realism	ESCI
Philosophy & Religion	Philosophy	Philosophical Review	AHCI
Philosophy & Religion	Philosophy	Constructivist Foundations	AHCI
Philosophy & Religion	Philosophy	Philosophy Compass	AHCI
Physics	Physics, Applied	Nature Reviews Physics	SCIE
Physics	Physics, Applied	Nature Materials	SCIE

Physics	Physics, Applied	Nano-Micro Letters	SCIE
Physics	Physics, Applied	Nature Photonics	SCIE
Physics	Physics, Applied	Interdisciplinary Materials	ESCI
Plant & Animal Science	Plant Sciences	Annual Review of Plant Biology	SCIE
Plant & Animal Science	Plant Sciences	Molecular Plant	SCIE
Plant & Animal Science	Plant Sciences	Trends in Plant Science	SCIE
Plant & Animal Science	Plant Sciences	Nature Plants	SCIE
Plant & Animal Science	Plant Sciences	Annual Review of Phytopathology	SCIE
Psychiatry/Psychology	Psychology	Annual Review of Psychology	SCIE
Psychiatry/Psychology	Psychology	Psychological Bulletin	SCIE
Psychiatry/Psychology	Psychology	Psychotherapy and Psychosomatics	SCIE
Psychiatry/Psychology	Psychology	Annual Review of Clinical Psychology	SCIE
Psychiatry/Psychology	Psychology	Advances in Methods and Practices in Psychological Science	SCIE
Social Sciences, General	Sociology	Annual Review of Sociology	SSCI
Social Sciences, General	Sociology	Annals of Tourism Research	SSCI
Social Sciences, General	Sociology	Sociological Methods & Research	SSCI
Social Sciences, General	Sociology	American Sociological Review	SSCI
Social Sciences, General	Sociology	Sociological Theory	SSCI
Visual & Performing Arts	Film, Radio, Television	Positif	AHCI
Visual & Performing Arts	Film, Radio, Television	Media Psychology	AHCI
Visual & Performing Arts	Film, Radio, Television	Television & New Media	AHCI
Visual & Performing Arts	Film, Radio, Television	Critical Studies in Television	AHCI
Visual & Performing Arts	Film, Radio, Television	Continuum-Journal of Media & Cultural Studies	AHCI

Appendix B. CitationHub System

CitationHub is a separate interactive dashboard and visualization platform built on top of IDCite. It provides a user-facing environment for exploring influential seed papers, citing papers, authors, citation intents, disciplinary patterns, geographic distributions, citation networks, and the structured scholarly knowledge graph derived from IDCite. The platform also provides search, analytics, knowledge-graph exploration, and SPARQL-based querying capabilities. CitationHub should therefore be understood as an application and exploration layer powered by IDCite rather than as a separate Zenodo dataset version.

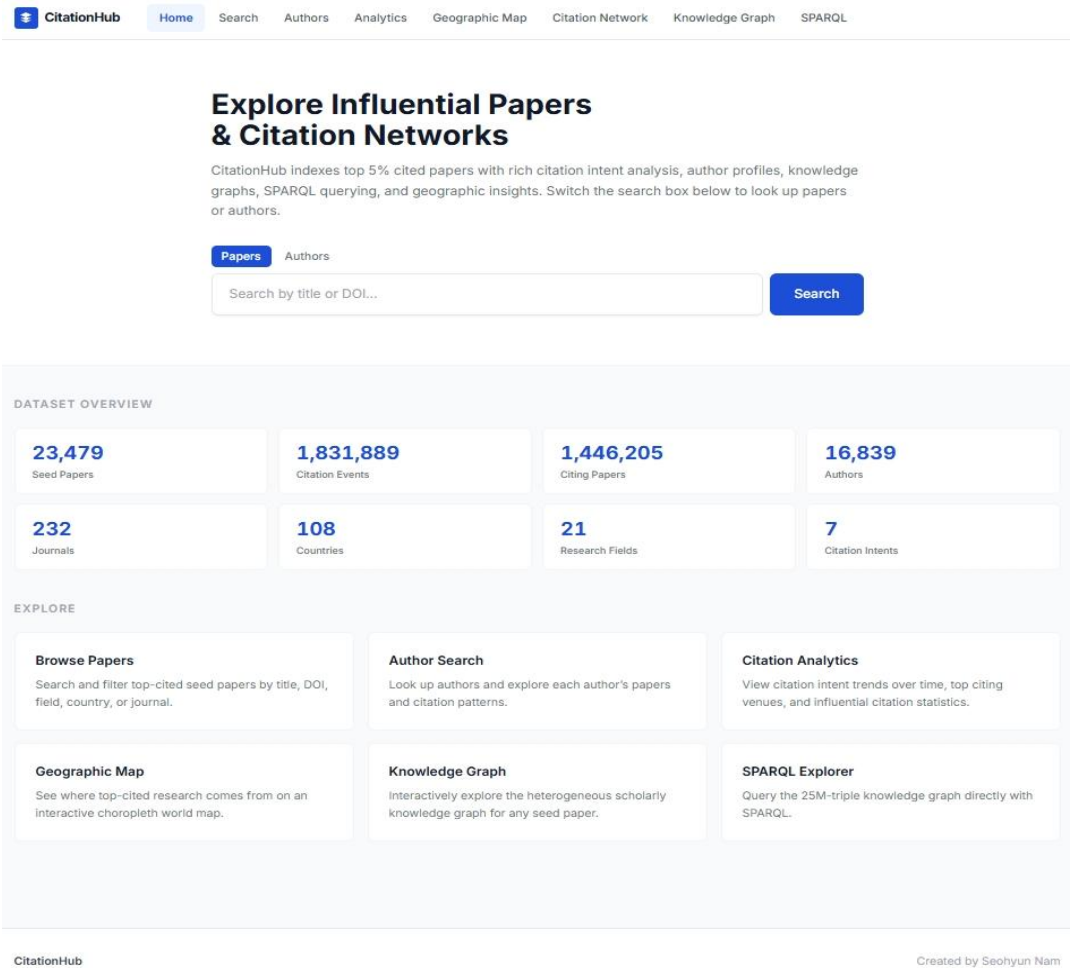


Figure B1. CitationHub home interface. The system provides interactive access to IDCite-based paper and author search, citation analytics, geographic exploration, citation networks, knowledge-graph exploration, and SPARQL querying.

Bibliography and List of References

- [1] M.D. Wilkinson, et al., The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data* 3 (2016) 160018. <https://doi.org/10.1038/sdata.2016.18>.
- [2] P. Knoth, et al., CORE: A global aggregation service for open access papers, *Sci. Data* 10 (2023) 366. <https://doi.org/10.1038/s41597-023-02208-w>.
- [3] S. Peroni, D. Shotton, OpenCitations, an infrastructure organization for open scholarship, *Quant. Sci. Stud.* 1 (2020) 428-444. https://doi.org/10.1162/qss_a_00023.
- [4] G. Hendricks, D. Tkaczyk, J. Lin, P. Feeney, Crossref: The sustainable source of community-owned scholarly metadata, *Quant. Sci. Stud.* 1 (2020) 414-427. https://doi.org/10.1162/qss_a_00022.
- [5] J. Priem, H. Piwowar, R. Orr, OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, *arXiv* (2022). <https://doi.org/10.48550/arXiv.2205.01833>.
- [6] W. Ammar, et al., Construction of the literature graph in Semantic Scholar, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Demonstrations*, Association for Computational Linguistics, New Orleans, LA, USA, 2018, pp. 84-91. <https://doi.org/10.18653/v1/N18-3011>.
- [7] K. Lo, L.L. Wang, M. Neumann, R. Kinney, D.S. Weld, S2ORC: The Semantic Scholar Open Research Corpus, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 4969-4983. <https://doi.org/10.18653/v1/2020.acl-main.447>.
- [8] A. Cohan, W. Ammar, M. van Zuylen, F. Cady, Structural scaffolds for citation intent classification in scientific publications, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Minneapolis, MN, USA, 2019, pp. 3586-3596. <https://doi.org/10.18653/v1/N19-1361>.
- [9] A. Lauscher, et al., MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting, in: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Seattle, WA, USA, 2022, pp. 1875-1889. <https://doi.org/10.18653/v1/2022.naacl-main.137>.
- [10] D. Jurgens, S. Kumar, R. Hoover, D. McFarland, D. Jurafsky, Measuring the evolution of a scientific field through citation frames, *Trans. Assoc. Comput. Linguist.* 6 (2018) 391-406. https://doi.org/10.1162/tac_l_a_00028.
- [11] S. Teufel, A. Siddharthan, D. Tidhar, Automatic classification of citation function, in: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Sydney, Australia, 2006, pp. 103-110. <https://aclanthology.org/W06-1613/>.
- [12] T.-A. Phan, K.-H.N. Bui, J.J. Jung, Understanding citation intents by generative intent model based on heterogeneous graph neural network, *Inf. Process. Manag.* 63 (2026) 104743. <https://doi.org/10.1016/j.ipm.2026.104743>.
- [13] A. Hogan, et al., Knowledge graphs, *ACM Comput. Surv.* 54 (2021) 1-37. <https://doi.org/10.1145/3447772>.
- [14] M.E. Rose, J.R. Kitchin, pybliometrics: Scriptable bibliometrics using a Python interface to Scopus, *SoftwareX* 10 (2019) 100263. <https://doi.org/10.1016/j.softx.2019.100263>.

End of IDCite Version 3 Project and Dataset Documentation