 ISSN:2302-8149
e-ISSN:2540-9719

A Large-Scale Open Dataset of Computer Science Research Papers (2020–2025)

[1] Zaid Mundher*, [2]Manar Talat Ahmad
[1,2]Department of Computer Science, University of Mosul, Mosul, Iraq
e-mail: zaidabdulelah@uomosul.edu.iq, manar_seyala@uomosul.edu.iq

(***received*:** 15 March 2026, ***revised*:** 16 April 2026, ***accepted*:** 19 April 2026)

Abstract

The rapid growth of publications in different fields, such as computer science, required well-structured datasets to support data-driven research. This paper presents an open large-scale dataset of computer science research papers published between 2020 and 2025, collected from Crossref metadata using the Crossref REST API. A structured keyword-based retrieval framework was developed to collect papers and their associated metadata. Preprocessing techniques, including cleaning, normalization, and validation were also made on the collected data. The introduced dataset has 4,313,328 research paper records which represents one of the largest structured collections of computer science publications for the specified period. The dataset provides comprehensive metadata fields that enable large-scale analysis, research trend identification, collaboration network exploration, and the recommendation systems development.

Keywords: computer science research papers, Crossref, large-scale dataset, REST API, Zenodo

1 Introduction

It is clearly that over the past decades, the number of scientific publications increased dramatically. Specifically, in today's technology era, the field of Computer Science gained a rise in the number of research papers. Also, having such a large number of papers helps in conducting bibliometric studies. Bibliographic data means the metadata of the research papers, such as title, authors, journals and DOI. This data may be collected using different data source, such as Scopus, Web of Science, Google Scholar and Crossref[1][2]. The main obstacle remains the lack of a structured dataset of published papers with their metadata. Having a large-scale bibliographic data can help in implementing a variety of application such as citation analysis, research trend analysis, dataset construction, and recommendation system development. To address this limitation, this study introduces a large-scale, dataset of computer science research papers derived from Crossref. Crossref is a non-profit organization which is responsible for registering and maintaining Digital Object Identifiers (DOIs) for publications [3][4]. Figure 1 illustrates the relation among Crossref, journals, and researchers.

Figure 1 Relation among crossref, journals, and researchers

Crossref can be considered as a large-scale, structured repository of scholarly metadata across many fields, providing REST API which can be used to access metadata programmatically[5][6]. This work focuses on publications between 2020 and 2025 to get recent research dynamics while ensuring substantial volume. Many data cleaning methods were implemented to the gathered data to prepare the dataset for future analysis. The resulting resource aims to support bibliometric studies, collaboration network analysis, trend detection, and intelligent recommendation systems. As a last step, the collected dataset was published on the Zenodo website. Zenodo is an online research

 ISSN:2302-8149
e-ISSN:2540-9719

repository that allows users to upload and share their publications and data. It generates and assigns a DOI to each uploaded file which make it citable and accessible easily[7][8].

2 Literature Review

The evaluation of Crossref as a bibliographic data source was addressed in several research papers such as [9][10]. Furthermore, many previous works used Crossref as a primary source to collect papers. For example, in [11], Crossref was used beside bioRxiv to build a dataset that connect preprints with their journal publications. In addition, authors of [12] describe the metadata that can be collected from Crossref and its value in the ecosystem of scholarly research. Moreover, in [13], Crossref was used beside Google Scholar and Scopus database as a data source to achieve a bibliometric analysis of higher-order thinking skill. In [14], Crossref was used to collect papers that related to AI published between 2014 and 2024. The collected data then used to implement a bibliometric analysis. On the other hand, in [15] Crossref was used as a data source to implement a bibliometric analysis of AI.

Regardless of the previous work, there is a lack of a large-scale, updated and open-access dataset specifically for Computer Science field. This work addresses this gap by providing a free dataset that enables more advanced analysis.

3 Data Collection Methodology

The workflow of this work is shown in Figure 2 that explains the dataset construction process

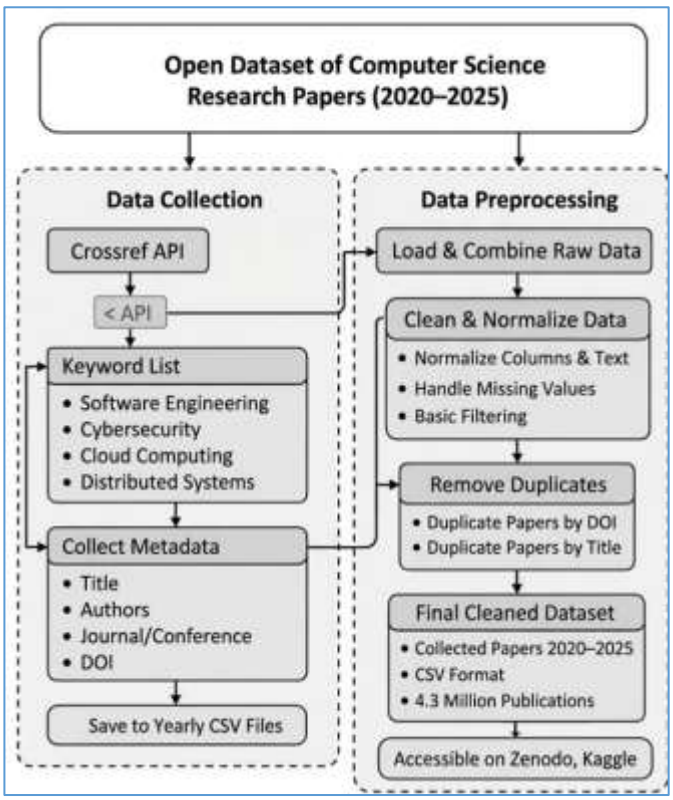

Figure 2 Overview of the proposed work

The data collection and preprocessing were implemented using Python, with libraries such as requests and pandas. More details are explained below.

3.1. Data Source

In this work, Crossref was selected a data source because of its wide coverage and its open accessibility. The data were collected using the Crossref REST API, which provides a free access publicly to the metadata for publications. Through programmatic queries, metadata elements

 ISSN:2302-8149
e-ISSN:2540-9719

including title, journal name, authors, publication date, and DOI were retrieved. Below is a sample query that explain how keywords were applied to the 'title' field to retrieve data.

https://api.crossref.org/works?
query.title=software engineering
&filter=from-pub-date:2022-01-01,until-pub-date:2022-12-31
&rows=1000
*&cursor=**
&mailto=zaid_mail

3.2. Keyword Selection

A keyword-based retrieval strategy was adopted in this work to identify publications related to the field of Computer Science. A predefined list of domain-specific keywords was built and used as search queries through the Crossref API. Table 1 presents the list of selected keywords. The retrieved metadata, along with the corresponding search keyword, were stored in CSV format for further preprocessing and analysis.

4. Data Preprocessing and Potential Uses

Different preprocessing and data cleaning techniques were applied to the collected data to prepare the dataset for future analysis.

1. Removing Duplicates

Because the keyword-based method was used, the same paper may appear in multiple queries. Therefore, duplicate records were recognized and removed using paper's DOIs.

2. Column Preprocessing

The "title" column is cleaned by removing spaces and special characters. In addition, all characters were converted to lowercase to ensure consistency. A new column ("new_title") was created to store the processed titles.

3. Removing Invalid Records

Records with incomplete or missing information were excluded from the dataset. In this step, 691,502 records were removed due to their incomplete data.

After preprocessing, the final dataset contained 4,313,328 records, representing validated Computer Science publications between 2020 and 2025. The collected dataset can be used in different research ideas and applications, such as: Bibliometric analysis, Research trend analysis, Recommendation systems, and Researcher collaboration network analysis. The dataset has been made publicly available on Zenodo and Kaggle, and it is assigned a DOI to ensure proper citation and accessibility [***doi.org/10.5281/zenodo.18011171***].

5. Results and Findings

5.1 Dataset Construction

As an initial result, approximately 7,052,374 records were collected. After doing some cleaning steps, 691,502 invalid records were removed. In addition, 2,047,544 duplicate papers were recognized and removed during the deduplication process. As a result, the final dataset consists of 4,313,328 unique research papers published between 2020 and 2025. Table 1 shows a sample output of the collected data.

Table 1 A sample of the collected data

Ttle	Journal	Year	DOI	Authors
Enhancing Fraud Detection in Imbalanced Datase...	Mansoura Journal for Computer and Information ...	2025	10.21608/mjcis.2025.313097.1007	Walaa salah salem; ibrahim el- hasnony; Ahmed ...
Improving the Quality of Learning Through the ...	Jurnal Penelitian Pendidikan IPA	2025	10.29303/jppipa.v11i10.10736	Tiur Malasari Siregar; Muliawan Firdaus; Trisn...

 ISSN:2302-8149
e-ISSN:2540-9719

Evaluating Data Trust in Blockchain-Based IoT ...	2025 IEEE 22nd Consumer Communications &am...	2025	10.1109/ccnc54725.2025.10975901	Rashmi Ratnayake; Madhusanka Liyanage; Liam Mu...
Dynamic label correlations and dual-semantic e...	Neurocomputing	2025	10.1016/j.neucom.2025.129371	Shaohua Teng; Ziye Fang; Zefeng Zheng; Naiqi W...
ALDEN: Dual-Level Disentanglement with Meta-le...	Proceedings of the 33rd ACM International Conf...	2025	10.1145/3746027.3754741	Yuxiong Xu; Bin Li; Weixiang Li; Sara Mandelli...

5.2 Author Statistics

The number of unique authors was calculated which was 6,561,661 authors. More author-level statistics are below:

- Average papers per author: 2.65
- Median papers per author: 1
- Maximum papers by a single author: 4,608

Furthermore, the top productive author names are shown in Table 2.

Table 2 Top productive author names

	Name	Number of papers
1.	Wei Wang	4,608
2.	Yang Liu	4,515
3.	Wei Zhang	3,712
4.	Lei Wang	3,353

It is important to notice that the above results (Table x) were conducted based on author names. This name-based approach has limitations since those different researchers may have the same name. However, Table x provides a general overview of author productivity.

5.3 Exploratory Data Analysis (EDA)

An Exploratory Data Analysis (EDA) on the collected data was implemented to get deeper insights.

A. Publication Growth

Number of papers that were published each year was also calculated to explain how research output changed over year. Figure 3 shows the distribution of papers over year.

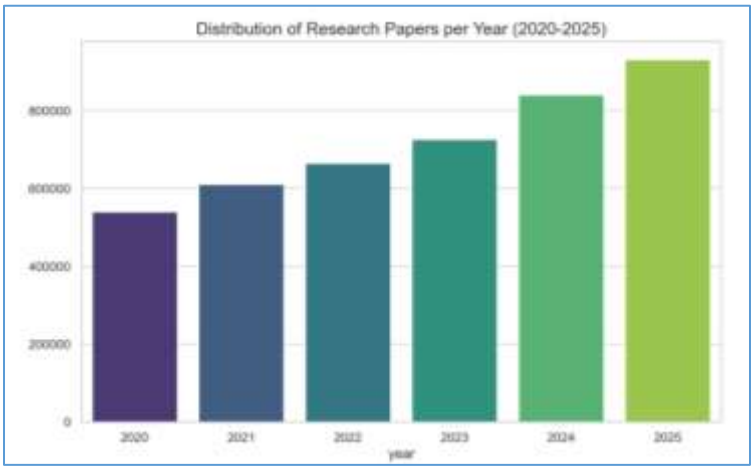

Figure 3 Distribution of research papers per year (2020-2025)

 ISSN:2302-8149
e-ISSN:2540-9719

B. Top Publishing Venues

Determining where Computer Science papers are being published was also calculated. Figure 4 shows top 10 publishing venues.

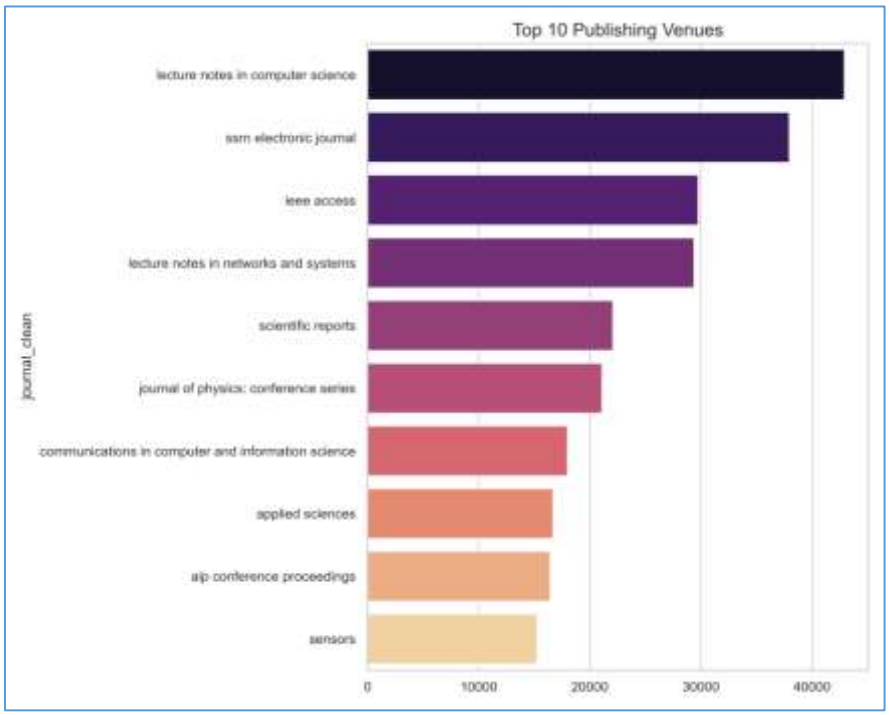

Figure 4 Top 10 publishing venues

C. Title Keyword Frequency

The word count of the most frequent words in title was visualized (Figure 5) which can help to retrieve the hottest topics.

Figure 5 Most frequent words in title

D. Author Collaboration Patterns

Number of authors per paper is calculated to estimate the level of collaboration in the Computer Science domain. Figure 6 shows the distribution of authors count per paper. By analyzing the number of authors per paper, we can see the level of collaboration in the CS field.

 ISSN:2302-8149
e-ISSN:2540-9719

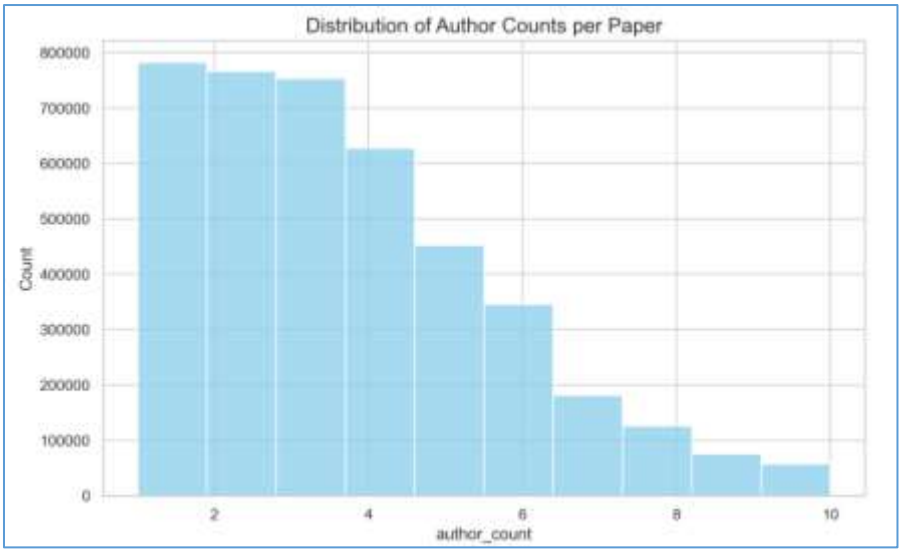

Figure 6 Distribution of author counts per paper

6. Conclusion

Motivated by the importance of data availability, this work was conducted. The main goal was to provide a large-scale open of published paper of Computer Science field. Crossref was selected as a data source to collect papers and retrieve their metadata. The Crossref REST API was used through a keyword-based approach. The experimental results confirmed the value of Crossref as a rich data source for published papers. More than four million papers were collected (with their metadata) supporting more advanced analysis and research ideas including bibliometric analysis, research trend identify, and collaboration network analysis. Furthermore, the collected dataset is publicly available on Zenodo and Kaggle.

References

[1] M. Yıldız and T. K. Yılmaz, "*Bibliometric Analysis in Scientific Research using R: A Review of Scopus and Web of Science Databases*," *Journal of Data Applications*, pp. 31–46, 2024, DOI: 10.26650/JODA.1462396.

[2] M. Juliardi and I. Malik, "*Bibliometric Analysis of Data Science: Trends, Contributions, and Research Developments*," *West Science Interdisciplinary Studies*, Vol. 1, pp. 365–375, 2023, DOI: 10.58812/wsis.v1i07.81.

[3] Chudlarský, Tomáš & Dvorak, Jan. (2020). *Can Crossref Citations Replace Web of Science for Research Evaluation? The Share of Open Citations*. Journal of Data and Information Science. 5. 10.2478/jdis-2020-0037.

[4] Pentz, Ed. (2022). *Role of Crossref in Journal Publishing Over the Next Decade*. Science Editing. 9. 53-57. 10.6087/kcse.263.

[5] Lammey, Rachael. (2019). *How Publishers Can Work with Crossref on Data Citation*. Science Editing. 6. 166-170. 10.6087/kcse.165.

[6] Visser, Martijn & van Eck, Nees Jan & Waltman, Ludo. (2021). *Large-Scale Comparison of Bibliographic Data Sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic*. Quantitative Science Studies. 2. 1-37. 10.1162/qss_a_00112.

[7] Garrido, Irene & Loureiro, Maria & Gutleber, Johannes. (2025). *The Value of an Open Scientific Data and Documentation Platform in a Global Project: The Case of Zenodo.* In The Economics of Big Science 2.0 (pp. 181–200). Springer. https://doi.org/10.1007/978-3-031-60931-2_14

[8] Sicilia, M. & Barriocanal, Elena & Sánchez-Alonso, Salvador. (2017). *Community Curation in Open Dataset Repositories: Insights from Zenodo*. Procedia Computer Science. 106. 10.1016/j.procs.2017.03.009.

[9] van Eck, Nees Jan & Waltman, Ludo. (2022). *Crossref as a Source of Open Bibliographic Metadata*. 10.31222/osf.io/smxe5.

 ISSN:2302-8149
e-ISSN:2540-9719

[10] Liang, Zhentao & Mao, Jin & Lu, Kun & Li, Gang. (2021). *Finding Citations for PubMed: A Large-Scale Comparison between Five Freely Available Bibliographic Data Sources*. 10.48550/arXiv.2111.00172.
[11] Badalova, Fidan & Sienkiewicz, Julian & Mayr, Philipp. (2026). *PreprintToPaper dataset: Connecting bioRxiv Preprints with Journal Publications*. Scientific Data. 13. 10.1038/s41597-026-06867-3.
[12] Hendricks, Ginny & Tkaczyk, Dominika & Lin, Jennifer & Feeney, Patricia. (2020). *Crossref: The Sustainable Source of Community Owned Scholarly Metadata*. Quantitative Science Studies. 1. 1-14. 10.1162/qss_a_00022.
[13] Deda, Yohanis Ndapa. (2023). *Bibliometric Analysis of Higher-Order Thinking Skills based on Google Scholar, Crossref, and Scopus Database*. 127-136. 10.23917/varidika.v35i2.23223.
[14] Pirmanto, Dovel. (2025). Analisis *Bibliometrik Artificial Intelligence* pada *Database Crossref* (2014 – 2024). Shaut Al-Maktabah : Jurnal Perpustakaan, Arsip dan Dokumentasi. 17. 87-106. 10.37108/shaut.v17i2.2409.
[15] Borrego, Ángel & Ardanuy, Jordi & Arguimbau, Llorenç. (2023). *Crossref as a Bibliographic Discovery Tool in the Arts and Humanities*. Quantitative Science Studies. 4. 1-17. 10.1162/qss_a_00240.