BIP! NDR (NoDoiRefs): A Dataset of Citations from Papers Without DOIs in Computer Science Conferences and Workshops

Paris Koloveas[1,2(✉)], Serafeim Chatzopoulos[2], Christos Tryfonopoulos[1], and Thanasis Vergoulis[2]

[1] University of the Peloponnese, Tripolis, Greece
{pkoloveas,trifon}@uop.gr
[2] IMSI, Athena RC, Athens, Greece
{schatz,vergoulis}@athenarc.gr

Abstract. In the field of Computer Science, conference and workshop papers serve as important contributions, carrying substantial weight in research assessment processes, compared to other disciplines. However, a considerable number of these papers are not assigned a Digital Object Identifier (DOI), hence their citations are not reported in widely used citation datasets like OpenCitations and Crossref, raising limitations to citation analysis. While the Microsoft Academic Graph (MAG) previously addressed this issue by providing substantial coverage, its discontinuation has created a void in available data. BIP! NDR aims to alleviate this issue and enhance the research assessment processes within the field of Computer Science. To accomplish this, it leverages a workflow that identifies and retrieves Open Science papers lacking DOIs from the DBLP Corpus, and by performing text analysis, it extracts citation information directly from their full text. The current version of the dataset contains more than 510K citations made by approximately 60K open access Computer Science conference or workshop papers that, according to DBLP, do not have a DOI.

Keywords: Citation extraction · Bibliographic metadata · Text mining

1 Introduction

A *(bibliographic) citation* refers to a conceptual (directional) link that connects a research work (usually a publication) which contains a reference to (i.e., "cites") another work (which is being "cited"). During the last decades, citations have become one of the most important types of bibliographic metadata [12]. The main reason for that is that they are often considered as proxies of scientific impact, since a citation can be interpreted as an acknowledgement for the contribution of the cited work into the citing one (although this might not always be the case [1,17]). As a result, they have been instrumental in scientometrics,

O. Alonso et al. (Eds.): TPDL 2023, LNCS 14241, pp. 99–105, 2023.
https://doi.org/10.1007/978-3-031-43849-3_9

becoming the basis for the calculation of various research impact indicators [16]. Such indicators have been used to facilitate scientific knowledge discovery (e.g., they have been used by academic search engines to help researchers prioritise their reading [15]), monitor research production [11], assist research assessment processes, and in many other applications.

Various sources of citation data have become available during the previous decades to address the needs of use-cases like the aforementioned ones. Apart from proprietary and restrictive sources, like Clarivate Analytics' Web of Science, Google Scholar and the Microsoft Academic Graph (MAG) [2], due to the raised popularity of the Open Science movement, a couple of open datasets that provide citations (e.g., OpenCitations[1], the OpenAIRE Graph[2]) have also become available during the last years. Almost all of them report citations as DOI-to-DOI pairs, failing to cover citations that involve publications for which a DOI has not been assigned. This may not be a significant problem for many disciplines, but in Computer Science, a considerable number of conferences and workshops do not assign DOIs to their papers. In addition, in this field, conference and workshop papers are peer reviewed and, historically, serve as important contributions, carrying significant weight in research assessment processes. As a result, if they are not considered during citation analyses, this can overlook an important part of scientific production and even introduce bias. In the past, Microsoft Academic Graph (MAG) was partially covering this gap by also offering citations for papers that do not have a DOI. However, since its discontinuation in December 2021, this data collection is no longer maintained and updated, thus its coverage is continuously declining.

In this work, we introduce BIP! NDR, an open dataset that aims to cover this gap, improving research assessment processes and other relevant applications within the field of Computer Science. The dataset is constructed based on a workflow that identifies and retrieves Open Science publications lacking DOIs from DBLP[3], the most widely known bibliographic database for publications from Computer Science, and then performs text analysis to extract citation information directly from the respective manuscripts. The current version of the dataset contains more than 510K citations made by approximately 60K Computer Science conference or workshop papers that, according to DBLP, do not have a DOI. We plan to frequently update the dataset so that it can become an important resource for citations in Computer Science that are missing from the most important citation datasets. This is a valuable addition to the toolboxes of scientometricians so that they can perform more concrete analysis in the Computer Science domain.

Outline. The rest of the manuscript is organized as follows: in Sect. 2 we elaborate on the technical details related to the production of the BIP! NDR dataset; in Sect. 3 we discuss the structure of the dataset; finally, in Sect. 4 we conclude the work while also discussing future planned extensions.

[1] OpenCitations: https://opencitations.net.
[2] OpenAIRE Graph: https://graph.openaire.eu.
[3] DBLP: https://dblp.uni-trier.de/.

2 Dataset Production Workflow

In this section, we discuss the BIP! NDR dataset production workflow and we elaborate on the technical details of its various components. The source code of the production workflow is available as open source on GitHub[4]. A high-level overview of the workflow is depicted in Fig. 1.

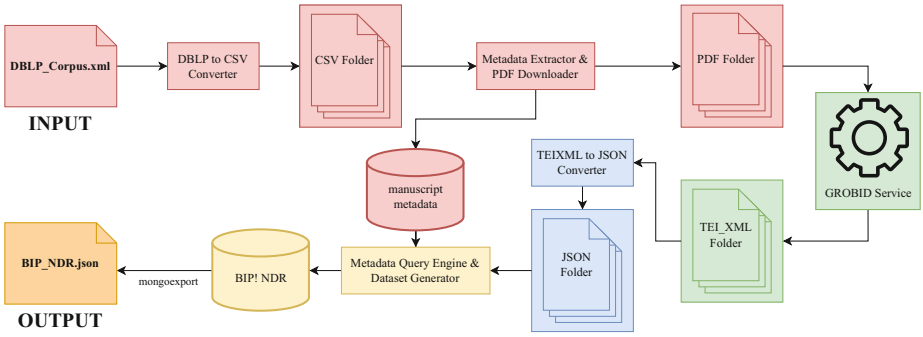

Fig. 1. A high-level overview of the dataset production workflow.

The main input to the workflow is the DBLP Corpus, which we use to collect URLs hosting Open Access manuscripts from the field of Computer Science, focusing on those that do not have a DOI. We collect these manuscripts so that we will be able to extract citations from the respective PDF files. DBLP [5, 6] consolidates scholarly metadata from several open sources which cover the Computer Science field and is largely manually curated and frequently updated.

As a result, this collection is ideal for our purposes. Our analysis shows that out of the approximately 320K open access conference publications, approximately 260K do not have a DOI. These publications are the ones that we aim to cover through the evolution of our dataset. The current version of our dataset (v0.1) [4] is based on the November 2022 Monthly Snapshot of DBLP [13]. The DBLP Corpus comes in XML format with all the bibliographic entries together in a single file. Therefore, as a first step, we use `dblp-to-csv`[5] to split the corpus into separate CSV files, grouped by publication type. We further process these CSV files to (a) extract manuscript metadata and store them in a document-oriented database, and (b) follow the included links to download the PDF files of Open Access papers. These operations ensure that the structured manuscript metadata from the DBLP Corpus are easily accessible to our workflow for querying and further processing.

For the next step of our workflow, we needed a tool to extract information from the PDF files while maintaining the headers, structure and sectioning

[4] BIP! NDR repository: https://github.com/athenarc/bip-ndr-workflow.
[5] dblp-to-csv: https://github.com/ThomHurks/dblp-to-csv.

of the manuscript. After a thorough evaluation of the literature regarding the tools used for reference extraction from PDFs, we concluded that based both on surveys [9,14], and prominent works that required extensive bibliography parsing [7,10], GROBID [8] is currently the best tool for the task. GROBID converts the PDF files to the TEI XML publication format[6]. Apart from the PDF extraction capabilities, GROBID offers a consolidation option to resolve extracted bibliographical references using services like biblio-glutton[7] or the CrossRef REST API[8]. We apply this consolidation option to our workflow, and GROBID sends a request to the Crossref web service [3] for each extracted citation. If a core of metadata (such as the main title and the first author) is correctly identified, the system retrieves the full publisher's metadata. These metadata are then used for correcting the extracted fields and for enriching the results. We utilize this output to potentially identify the DOI of a publication and attempt to match it with a DBLP entry.

The TEI XML files that GROBID produces are useful for identifying the structure of a manuscript, but are very verbose and are not convenient to process in large volumes. For that reason, we have created a *TEIXML to JSON Converter* that turns the files into JSON format. This conversion process involves extracting relevant information from the TEI XML files and mapping it to the corresponding JSON structure. The resulting JSON files are smaller in size and are compatible with a wide range of tools for processing.

At this point, we have reached the core functionality of our workflow, the process of *querying the DBLP metadata* for the bibliographic references of the papers in our collection. This process queries the manuscript metadata database for each document in the JSON Folder. For each document, we parse the reference list and we first check if a DOI exists in a publication entry. If it exists, we query our database based on the DOI. If a result is returned, we store the `dblp_id`, the `doi`, as well as, the `bibliographic_reference` extracted from the JSON file. Otherwise, we query based on the publication title. On a positive result, we store the previously mentioned fields to the dataset entry. If neither the publication title nor the DOI return a positive result, the publication could not be found in our DBLP metadata, so we store only the `doi` and `bibliographic_reference` from the JSON file. This process ultimately creates the "BIP! NDR" collection which constitutes our dataset.

The final step involved using the *mongoexport* utility to export the "BIP! NDR" collection from MongoDB into the final JSONL file. The exported file served as the culmination of the dataset generation process, providing a structured collection of scholarly data ready for research and analysis.

[6] TEI XML format: https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html.
[7] biblio-glutton: https://github.com/kermitt2/biblio-glutton.
[8] Crossref API: https://www.crossref.org/documentation/retrieve-metadata/rest-api/.

```
{
   "_id":{
      "$oid": "6460a56bda929a01210c1b57"
   },
   "citing_paper": {
      "dblp_id": "conf/ecsa/GasperisPF21"
   },
    "cited_papers": [
      {
         "dblp_id": "journals/sigpro/AlbusacCLVL09",
         "doi": "10.1016/j.sigpro.2009.04.008",
         "bibliographic_reference": "J. Albusac, J. Castro-Schez, L. Lopez-Lopez, D. Vallejo, L. Jimenez-Linares,
             A supervised  learning approach to automate the acquisition of knowledge in surveillance systems,  Signal
             Processing 89 (2009) 2400-2414.  doi:https://doi.org/10. 1016/j.sigpro.2009.04.008, special Section: Visual
             Information Analysis for Security."
      },
      {
         "dblp_id": "journals/cssp/Elhoseny20",
         "doi": "10.1007/s00034-019-01234-7",
         "bibliographic_reference": "M. Elhoseny, Multi-object detection and tracking (modt) machine learning model for real
             -time video surveillance systems, Circuits, Systems, and Signal Processing 39 (2020) 611-630.
             doi:10.1007/s00034-019-01234-7."
      },
      {
         "doi":"10.23919/IRS.2019.8768102",
         "bibliographic_reference":"F. Opitz, K. Dästner, B. v. H. z. Roseneckh-Köhler, E. Schmid, Data analytics and machine
             learning in wide area surveillance systems, in: 2019 20th International Radar Symposium (IRS), 2019, pp. 1-10.
             doi:10.23919/IRS.2019.8768102."
      },
      {
         "dblp_id":"journals/rfc/rfc3411",
         "bibliographic_reference":"D. Harrington, R. Presuhn, B. Wijnen, An architecture for describing simple network
             management protocol (snmp) management frameworks, 2002. doi:10.17487/RFC3411."
      }
   ]
}
```


Fig. 2. Data structure of the BIP! NDR dataset.

3 The BIP! NDR Dataset

In this section we present the structure of the dataset along with some basic statistics of the current version. The dataset is formatted as a JSON Lines (JSONL)[9] file where each line contains a valid JSON object. This file format enables file splitting and data streaming as the dataset grows in size. An indicative record (in JSON format) of the BIP! NDR dataset is depicted in Fig. 2.

Each JSON object has the following three main fields:

1. `_id` – the unique identifier of each entry
2. `citing_paper` – an object holding the `dblp_id` of each citing paper
3. `cited_papers` – an array that contains the objects that correspond to each reference found in the text of the `citing_paper`. Each object of the array may contain some or all of the following fields:
   (a) `dblp_id` – the dblp_id of the cited paper
   (b) `doi` – the doi of the cited paper
   (c) `bibliographic_reference` – the raw citation string as it appears in the citing paper

Note that not all the aforementioned fields in (3) are required for a `cited_paper` to be valid. Specifically, one of the `dblp_id` or `doi` identifiers is required for a cited paper to be added in the collection. Finally, the `bibliographic_reference` exists in all cited_paper objects since it is extracted directly from the PDF files of each citing paper in the dataset.

[9] JSON Lines data format: https://jsonlines.org/.

Table 1 summarises some statistics about the BIP! NDR dataset. In particular, 59,663 full texts from Open Access papers were parsed. A total of 1,054,107 references were evaluated, and among them, 511,842 references were successfully matched with corresponding keys from the DBLP database. Additionally, 366,106 DOIs were successfully matched with these DBLP keys. Finally, there were 22,569 DOIs that could not be matched with any DBLP key, indicating that they have not been indexed by DBLP.

Table 1. Statistics of BIP! NDR dataset (current version).

Statistic	#
Total Files Parsed	59,663
Total References Evaluated	1,054,107
DBLP Keys Matched	511,842
DOIs Matched with DBLP Key	366,106
DOIs without DBLP Key	22,569

4 Conclusions

We presented BIP! NDR, a dataset created using text analysis techniques on the DBLP database to extract citation information from the full text of the Open Access papers that do not have an assigned DOI. The dataset offers over 500K citations from Computer Science papers that do not have DOIs, addressing a significant limitation of widely used citation datasets in the field, that fail to cover them. As a result, it enables more comprehensive and accurate research assessment in Computer Science. In the future, we plan to improve the workflow so that it can identify more Open Source publications and to extend the dataset so that it can offer additional metadata for each citation (e.g., a class according to a citation classification algorithm).

Acknowledgements. This work was co-funded by the EU Horizon Europe projects SciLake (GA: 101058573) and GraspOS (GA: 101095129).

References

1. Abu-Jbara, A., Ezra, J., Radev, D.: Purpose and polarity of citation: towards NLP-based bibliometrics. In: Proceedings of the 2013 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 596–606 (2013)

2. Färber, M., Ao, L.: The Microsoft academic knowledge graph enhanced: author name disambiguation, publication classification, and embeddings. Quant. Sci. Stud. **3**(1), 51–98 (2022). https://doi.org/10.1162/qss_a_00183
3. Hendricks, G., Tkaczyk, D., Lin, J., Feeney, P.: Crossref: the sustainable source of community-owned scholarly metadata. Quant. Sci. Stud. **1**(1), 414–427 (2020). https://doi.org/10.1162/qss_a_00022
4. Koloveas, P., Chatzopoulos, S., Tryfonopoulos, C., Vergoulis, T.: BIP! NDR (NoDoiRefs): a dataset of citations from papers without DOIs in computer science conferences and workshops. https://doi.org/10.5281/zenodo.7962020
5. Ley, M.: The DBLP computer science bibliography: evolution, research issues, perspectives. In: Laender, A.H.F., Oliveira, A.L. (eds.) SPIRE 2002. LNCS, vol. 2476, pp. 1–10. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45735-6_1
6. Ley, M.: DBLP: some lessons learned. Proc. VLDB Endow. **2**(2), 1493–1500 (2009)
7. Lo, K., Wang, L.L., Neumann, M., Kinney, R.M., Weld, D.S.: S2orc: the semantic scholar open research corpus. In: ACL (2020)
8. Lopez, P.: GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 473–474. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04346-8_62
9. Meuschke, N., Jagdale, A., Spinde, T., Mitrović, J., Gipp, B.: A benchmark of pdf information extraction tools using a multi-task and multi-domain evaluation framework for academic documents. In: Sserwanga, I., et al. (eds.) Information for a Better World: Normality, Virtuality, Physicality, Inclusivity. LNCS, vol. 13972, pp. 383–405. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28032-0_31
10. Nicholson, J.M., et al.: scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. Quant. Sci. Stud. 1–17 (2021). https://doi.org/10.1162/qss_a_00146
11. Papastefanatos, G., et al.: Open science observatory: monitoring open science in Europe. In: Bellatreche, L., et al. (eds.) TPDL/ADBIS -2020. CCIS, vol. 1260, pp. 341–346. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-55814-7_29
12. Peroni, S., Shotton, D.M.: Opencitations, an infrastructure organization for open scholarship. Quant. Sci. Stud. **1**(1), 428–444 (2020). https://doi.org/10.1162/qss_a_00023
13. The DBLP Team: DBLP computer science bibliography. Monthly snapshot release of November 2022. https://dblp.org/xml/release/dblp-2022-11-02.xml.gz
14. Tkaczyk, D., Collins, A., Sheridan, P., Beel, J.: Machine learning vs. rules and out-of-the-box vs. retrained: an evaluation of open-source bibliographic reference and citation parsers. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 99–108 (2018)
15. Vergoulis, T., et al.: BIP! Scholar: a service to facilitate fair researcher assessment. In: Aizawa, A., Mandl, T., Carevic, Z., Hinze, A., Mayr, P., Schaer, P. (eds.) JCDL '22: The ACM/IEEE Joint Conference on Digital Libraries in 2022, Cologne, Germany, 20–24 June 2022, p. 42. ACM (2022). https://doi.org/10.1145/3529372.3533296
16. Vergoulis, T., et al.: BIP! DB: a dataset of impact measures for scientific publications. In: Leskovec, J., Grobelnik, M., Najork, M., Tang, J., Zia, L. (eds.) Companion of The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, 19–23 April 2021, pp. 456–460. ACM / IW3C2 (2021). https://doi.org/10.1145/3442442.3451369
17. Yousif, A., Niu, Z., Tarus, J.K., Ahmad, A.: A survey on sentiment analysis of scientific citations. Artif. Intell. Rev. **52**(3), 1805–1838 (2019)