Which Publications' Metadata Are in Which Bibliographic Databases? A System for Exploration

Michael **Färber**, Christoph **Braun**, Nicholas **Popovic**, Tarek **Saier** and Kristian **Noullet**

Karlsruhe Institute of Technology (KIT), Institute AIFB, Kaiserstr. 89, 76133 Karlsruhe, Germany

Abstract

The choice of databases containing publications' metadata (i.e., bibliographic databases) determines the available publication list of any author and, thus, their public appearance and evaluation. Having all publications listed in the various bibliographic databases is therefore important for researchers. However, the average number of publications a researcher publishes per year is steadily rising, making it labor-intensive and time-consuming for authors to investigate whether all their publications are given in all bibliographic databases online. In this paper, we present *RefBee*, an online system that retrieves the metadata of all publications for a given author from the various bibliographic databases and indicates which publications are missing in which database. Our system is available online at http://refbee.org/ and supports Wikidata, ORCID, Google Scholar, VIAF, DBLP, Dimensions, Microsoft Academic, Semantic Scholar, and DNB/GNB. Our system not only can serve as assistance tool for more than 4.7 million researchers of any discipline and publication's language, but also incentivizes the usage and population of Wikidata in the scholarly field.

Keywords

scientific papers, bibliometrics, scholarly data, digital libraries, open science

1. Motivation

In light of the FAIR data principles [1] and the various open science initiatives, it is important for any researcher to have the metadata of their publications available in bibliographic databases, such as Google Scholar [2] and Semantic Scholar [3]. This ensures that the publications can be found and therefore "exist" in the academic world. In addition, it makes sure that the publications are cited and that citation-based analyses and scientific impact evaluations [4, 5] can take place. Nowadays, many bibliographic databases exist, ranging from databases of specific publishers [6], to databases of general academic use [2], to databases for specific scientific disciplines [7]. It is therefore not surprising that the choice of bibliographic databases has a considerable influence on the completeness of an author's publication list [8] and, thus, public appearance
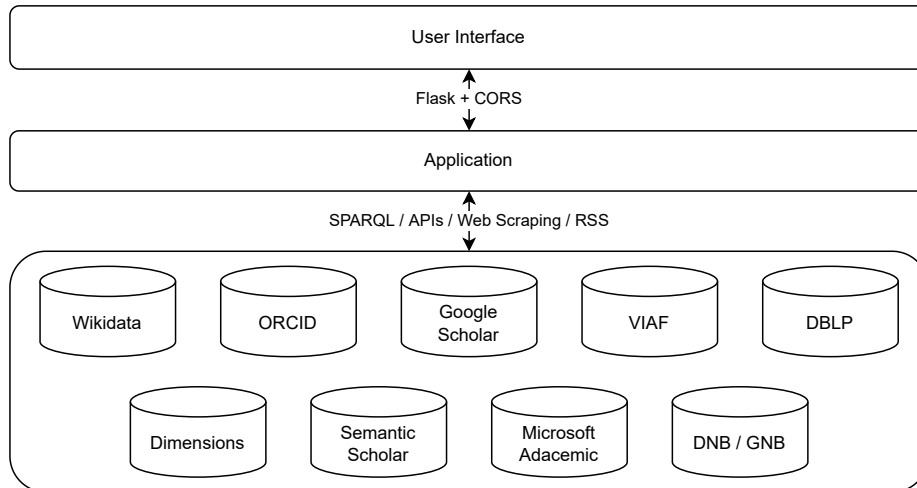
Figure 1: *RefBee*'s system architecture.

and evaluation. Furthermore, the average number of publications a researcher publishes per year is rising steadily, making it labor-intensive and time-consuming to investigate whether all their publications are given in all databases online. So far, to the best of our knowledge, an online system is missing that indicates to authors which of their publications are contained in which databases and which are not.

In this paper, we present an online system that retrieves the metadata of all publications for a given author from the various bibliographic databases and indicates which publications are missing in which database. Our system, *RefBee*, is available online at http://refbee.org/. The source code is available at https://github.com/kmdn/RefBee. Our system currently supports nine widely used databases (Wikidata, ORCID, Google Scholar, VIAF, DBLP, Dimensions, Microsoft Academic, Semantic Scholar, and DNB/GNB) and is easily extensible to further platforms. It relies on the APIs of the platforms and, thus, is always up-to-date. As Google Scholar does not provide a public API, the information from it is retrieved via web scraping. Due to the different formats of the APIs, data integration is a key aspect of our system. Our work is designed to assist researchers when dealing with their publications' metadata. In addition, our system encourages users to use and populate Wikidata as central hub in the linked open data cloud, because the users of our system are incentivized to enter missing links to Wikidata if they are not set there.

2. System Design

Databases. We consider the following platforms and databases containing publications' metadata (see also Figure 1): (1) Wikidata [9], (2) Google Scholar [2], (3) Semantic Scholar [3], (4) ORCID [10], (5) VIAF [11], (6) DBLP [7], (7) Dimensions [12], (8) Microsoft Academic [13], (9) German National Library (DNB/GNB) [14]. These databases were chosen because they are among the most frequently used bibliographic databases [8] and often editable by users. Table 1

Table 1
RefBee's data sources.

Name	Type of Querying	Wikidata Property	# Links in Wikidata	Editable by Users
Wikidata	SPARQL endpoint	-		yes
ORCID	API	wdt:P496	1,743,880	yes
Google Scholar	Web scraping	wdt:P1960	51,269	yes
VIAF	API	wdt:P214	2,712,068	(yes)
DBLP	API	wdt:P2456	50,630	(yes)
Dimensions	API	wdt:P6178	69,981	no
Microsoft Academic	API	wdt:P6366	279,327	no
Semantic Scholar	API	wdt:P4012	38,171	yes
DNB/GNB	RSS	wdt:P227	1,263,475	(yes)

provides an overview of these data sources.

Wikidata is increasingly used as a hub of linked open data, containing links to various data sources. As we can see in Table 1, Wikidata is particularly rich in links to other bibliographic databases. In total, 4,672,818 people represented in Wikidata have at least one link to the eight external bibliographic databases (as of 2021-10-16; see Table 1 for the number of links per data source). In addition, publications of 26,080 authors are modeled directly in Wikidata.[1] Overall, using Wikidata allows us to cover the bibliographic information of 4,675,310 researchers worldwide. Furthermore, Wikidata excels at modeling researchers worldwide independent of their country, language, and discipline [15]. Due to these reasons, our system uses Wikidata as starting point and the interlinked information (given via the linked identifiers in Wikidata) for obtaining the bibliographic information per author.

Data Consolidation. Matching the metadata records of each publication across the various databases is an essential part of our system. This task is non-trivial due to several reasons. First, the databases all use different schemas, partially with several classes representing publications per database,[2] and support varying query languages (e.g., SPARQL, REST API, or no query language). Second, not all databases contain publications' unique identifiers such as a DOI. Third, the metadata fields can contain minute differences across the databases, such as differences in the capitalization of paper titles. We therefore match and aggregate the information for each unique publication based on its normalized title.

Implementation. We use Python to process the data in the backend and Vue.js as the frontend framework to implement our user interface. For easy reuse, our system is deployed using Docker.[3]

User Interface and User Interaction. The user starts to interact with our system by entering an author name. In case the author name is ambiguous and several people are identified in

---

[1] Wikidata contains many authors and publications. However, often these entities are not interlinked.

[2] For instance, based on related works [16], we consider the classes Q23927052, Q13442814, Q18918145, Q591041, Q55915575 for querying publications in Wikidata.

[3] https://docker.com/

Figure 2: *RefBee*'s result page when searching for publications of "Michael Färber".

Wikidata with this name, our system displays the entity descriptions, based on which the author can choose the intended person (see upper part of Figure 2). The user is then presented with the result (see lower part of Figure 2). It provides an overview of which publications are listed in which data source. Green checks, red crosses, and question marks indicate if the publications are in the respective data sources, if they are missing, and if the databases are not linked from the author's Wikidata entry.

3. Related Work

Systems Showing (Aggregated) Bibliographical Information. Running systems focus on combining bibliographical information with other information (e.g., from the authors) or allowing people to edit bibliographical information in a collaborative fashion. CloudRef [17] allows users to edit bibliographical information online. Shakya et al. [18] propose a decentralized platform for sharing bibliographic information to obtain coherent publications' metadata. Dattolo and Corbatto [19] present a system for complex visual analyses on scientific bibliographies (e.g., cascades of paper citations). All these authors do not consider differences between databases.

    Interlinking Bibliographic Databases. A few approaches for interlinking bibliographic databases exist, typically covering links between single databases. For instance, Seidlmayer et al. [20] present an approach to integrate and interlink authors and scholarly publications in Wikidata by integrating data from ORCID. The Microsoft Academic Knowledge Graph [21] models papers' metadata and is interlinked with Wikidata and DBpedia.

    Analyzing the Availability of Bibliographical Information. Given the importance of bibliographical information in the bibliometrics and scientometrics fields, bibliographical

information has been studied in various regards. For instance, Nishioka and Färber [22] analyzed the availability of open citation data on the web, but do not provide a system for comparing citation data across databases and platforms. Schenkel [23] outlined in a talk how the DBLP bibliography is maintained and how it has been enhanced recently.

4. Conclusion

In this paper, we presented a system that allows scholars to investigate which of their publications are contained in which bibliographic database. Our system currently supports nine widely used databases (Wikidata, ORCID, Google Scholar, VIAF, DBLP, Dimensions, Microsoft Academic, Semantic Scholar, and DNB/GNB) and covers 4.7 million authors.

In the future, we will add bibliographic databases from further disciplines and include a component that allows users to import reference information (e.g., from DBLP) automatically into other databases. Furthermore, we plan to work on a system similar to *RefBee* that provides an overview of datasets listed in given dataset repositories (e.g., Zenodo, OpenAIRE).

References

[1] M. D. Wilkinson, M. Dumontier, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018.

[2] Google, Google Scholar, https://scholar.google.com/, 2022.

[3] AllenAI, Semantic Scholar, https://semanticscholar.org/, 2022.

[4] M. Baglioni, P. Manghi, A. Mannocci, Context-Driven Discoverability of Research Data, in: Proceedings of the 24th International Conference on Theory and Practice of Digital Libraries, TPDL'20, 2020, pp. 197–211.

[5] S. N. Kunnath, D. Herrmannova, D. Pride, P. Knoth, A meta-analysis of semantic classification of citations, Quant. Sci. Stud. 2 (2021) 1170–1215.

[6] ACM, ACM Digital Library, https://dl.acm.org/, 2022.

[7] Leibniz Center for Informatics, dblp computer science bibliography, https://dblp.org/, 2022.

[8] A. Martín-Martín, M. Thelwall, E. Orduña-Malea, E. D. López-Cózar, Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations, Scientometrics 126 (2021) 871–906.

[9] Wikimedia Foundation, Wikidata, https://wikidata.org/, 2022.

[10] ORCID Inc., ORCID, https://orcid.org/, 2022.

[11] OCLC, Inc., VIAF. Virtual International Authority File, https://www.viaf.org/, 2022.

[12] Digital Science & Research Solutions Inc., Dimensions, https://www.dimensions.ai/, 2022.

[13] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B. P. Hsu, K. Wang, An Overview of Microsoft Academic Service (MAS) and Applications, in: Proceedings of the 24th International Conference on World Wide Web, WWW'15, 2015, pp. 243–246.

[14] Bundesunmittelbare Anstalt des Öffentlichen Rechts, German National Library, https://www.dnb.de/, 2022.

[15] Y. Chikazawa, M. Katsurai, I. Ohmukai, Multilingual author matching across different

academic databases: a case study on KAKEN, DBLP, and PubMed, Scientometrics 126 (2021) 2311–2327.

[16] M. Färber, D. Lamprecht, The Data Set Knowledge Graph: Creating a Linked Open Data Source for Data Sets, Quant. Sci. Stud. 2 (2021) 1324–1355.

[17] O. Kopp, U. Breitenbücher, T. Müller, CloudRef – Towards Collaborative Reference Management in the Cloud, in: Proceedings of the 10th Central European Workshop on Services and their Composition, ZEUS'18, 2018, pp. 63–68.

[18] A. Shakya, H. Takeda, V. Wuwongse, I. Ohmukai, Sociobiblog: A Decentralized Platform for Sharing Bibliographic Information, in: Proceedings of the 2007 IADIS International Conference WWW/Internet, 2007, pp. 371–380.

[19] A. Dattolo, M. Corbatto, Assisting researchers in bibliographic tasks: A new usable, real-time tool for analyzing bibliographies, Journal of the Association for Information Science and Technology (2021).

[20] E. Seidlmayer, J. Voß, T. Melnychuk, L. Galke, K. Tochtermann, C. Schultz, K. U. Förstner, ORCID for Wikidata. Data enrichment for scientometric applications, in: Proceedings of the 1st Wikidata Workshop, Wikidata@ISWC'20, 2020.

[21] M. Färber, The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data, in: Proceedings of the 18th International Semantic Web Conference, ISWC'19, 2019, pp. 113–129.

[22] C. Nishioka, M. Färber, Evaluating the Availability of Open Citation Data, in: Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, BIRNDL@SIGIR'19, 2019, pp. 123–129.

[23] R. Schenkel, Integrating and Exploiting Public Metadata Sources in a Bibliographic Information System, in: Proceedings of the 7th International Workshop on Bibliometric-enhanced Information Retrieval, BIR@ECIR'18, 2018, pp. 16–21.