DBLP to Wikidata: Populating Scholarly Articles in Wikidata

Nandana Mihindukulasooriya

[1]IBM Research, New York, USA

Abstract

Scholarly data and resulting scientometrics play a vital role in the scientific community. Wikidata is a widely used knowledge graph with more than 110M entities and a comprehensive tooling ecosystem. Publishing scholarly data in Wikidata will make them more accessible and easier to integrate with existing knowledge. Such contributions will generally have to be made with the collaboration and support of the research community. This work is a small step towards that direction.

This demo introduces a tool and a method for adding our scholarly articles to Wikidata utilizing data from DBLP. We also provide authors with a tool to enhance Wikidata with associated entities, such as missing co-authors or conference proceeding entities, through a collaborative effort.

Source Repository: https://github.com/scholarly-wikidata/dblp-to-wikidata
Demo Video: http://tiny.cc/dblp-to-wikidata-demo
App URL: https://dblp-to-wikidata.streamlit.app/

Keywords

Scholarly Data, Wikidata, DBLP, Crowdsourcing, Scientometrics

1. Introduction

Given the importance of scholarly data, there are several community efforts to expose them in a semantically rich manner using Semantic Web standards; such as Semantic Web Dog Food (SWDF [1]), Scholarly data [2], and Open Research Knowledge Graph [3]. Complementary to such resources, Wikidata [4] is one of the largest crowdsourced knowledge graphs with more 110M entities and 25K active contributors[1]. Wikidata has a sustainable and user-friendly infrastructure ecosystem, including a UI tailored for crowdsourcing, an SPARQL endpoint with an easy-to-use query editor, entity linkers, and tools for search, visualization, etc. [5, 6]. By bringing Scholarly data into Wikidata, they can be more accessible and can be seamlessly integrated with existing background knowledge as well as the tooling ecosystem.

We believe it will be easier for researchers to bring their own scholarly articles to Wikidata as they have enough contextual knowledge about those articles to perform disambiguation and linking to entities such as co-authors, conference proceedings, and journals. While populating their own articles they can also populate and complete the missing entities related to those articles. The objective of this demo is to facilitate tooling and a recipe for that process.

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA

nandana@ibm.com (N. Mihindukulasooriya)
https://nandana.github.io/ (N. Mihindukulasooriya)
0000-0003-1707-4842 (N. Mihindukulasooriya)

[1] https://www.wikidata.org/wiki/Wikidata:Statistics

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2. Publication Proces

In this work, we created a web application that researchers could use to find their DBLP author IDs and extract the necessary metadata to publish their scholarly articles on Wikidata. Then, they can use the OpenRefine tool to disambiguate and link that information to Wikidata entities and transform it into a format that can be used to populate that information in Wikidata. With our tool, researchers do not need to understand the underlying details of DBLP or Wikidata or perform SPARQL queries to perform this task. Figure 1 shows an overview of the publication process which will be explained in this section.

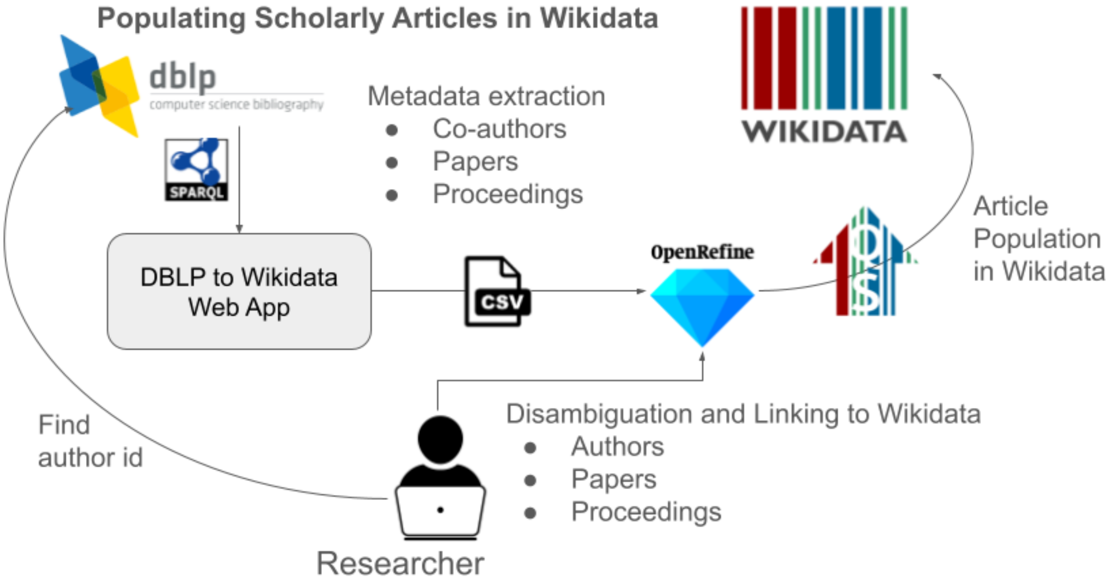

Figure 1: Overview of the process - extracting scholarly article information from DBLP, transforming, disambiguation and linking to Wikidata, and publication in Wikidata.

2.1. Finding the DBLP author ID

The main entry point for extracting the necessary information about one's research articles from DBLP [7] is through the DBLP author ID. DBLP provides a search API, and the author ID can be found by searching for the person's name. One can find your author ID using either the DBLP-to-Wikidata web application or the DBLP search API.

2.2. Extraction of metadata

Once the DBLP author ID of a person is known, the web application uses it to generate SPARQL queries following the DBLP ontology [2] to extract the metadata that will be transformed and linked to be published in Wikidata.

[2] https://dblp.org/rdf/docu/

We extract metadata for 3 main types of entities, i.e., scholarly articles, authors, and proceedings. Table 2.2 illustrated the attributes extracted from each of the entity types in DBLP related to the scholarly articles of a given researcher.

Entity	Extracted Metadata
Scholarly Article	title, co-author IDs, DBLP Publication ID, DOI, publication date, pages, language of work and proceedings ID
(Co-)Author	DBLP author ID, Open Researcher and Contributor IDentifier (ORCID), Open Research Knowledge Graph (ORKG) Identifier, Google Scholar author ID , ACM Digital Library author ID, GitHub username, and X (Twitter) username.
Proceeding	title, the DBLP publication identifier, the Digital Object Identifier (DOI), the International Standard Book Number (ISBN), publisher, series, and series volume number

Table 1
The information extracted from each of the entity types in DBLP.

2.3. Disambiguation and linking to Wikidata entities

In order to properly populate the data in Wikidata, each of the entities that we extracted from DBLP in the previous step such as scholarly articles, co-authors, and proceedings has to be disambiguated (e.g., there could be multiple people with the same name). As each researcher is supposed to process their own articles or the articles of a person they know of, they can quite easily perform this task. As shown in Table 2.2, there are several attributes that uniquely identify an entity compared to others. OpenRefine uses these attributes to initially disambiguate and to provide user suggestions. These suggestions can finally be approved or changed by the users.

It is important to note that while some of the entities we extract are already present in Wikidata, some others are not. The disambiguation step also helps avoid creating duplicate entities by linking to entities when they are present and identifying missing entities to be created.

Disambiguation is performed using “Wikidata reconciliation for OpenRefine”[3] web service. For exampple, for authors, the extracted attributes are mapped to DBLP author ID (P2456), ORCID ID (P496), ORKG ID (P10897), Google Scholar author ID (P1960), ACM Digital Library author ID (P864), GitHub username (P2037), and X username (P2002) Wikidata properties to disambiguate efficiently. If an author does not exist in Wikidata, users have the option to create an entity for that author. Once the disambiguation and linking are completed, the mapped Wikidata entities for each of the co-author are exported as a CSV file. Wikidata also allows using strings instead of entities for authors with the“author name string (P2093)” property. For proceedings, the extracted attributes are aligned to DBLP publication ID (P8978), digital object identifier (Q25670), and ISBN-13 (P212) Wikidata properties.

[3] https://wikidata.reconci.link/

2.4. Populating Wikidata

Finally, schema mapping in Open Refine is used to map scholarly articles and their authors to Wikidata using Wikidata properties and qualifiers. The OpenRefine schema with the mapping to Wikidata properties is available here[4]. The Schema mapping tool also performs validation, reports if there are any issues and provides a preview of changes to the Wikidata. Once everything is verified, the changes can be pushed to Wikidata through Open Refine.

3. Implementation

The application is created as a simple Web app in Python using the Streamlit[5] framework. The Web app used the DBLP author search API [6] and the SPARQL endpoint [7] to extract the data. Open Refine is used for entity disambiguation and linking to Wikidata. Open Refine is also used to map the data into Wikidata properties and create the Wikidata edits needed for the updates. The source code and the Wikibase schema mappings are available in the Github repository.

4. Conclusions and Future Work

This work introduces a tool and a method for extracting personal scholarly article data from DBLP and adding them to Wikidata. Researchers can use this process to enhance Wikidata by contributing missing scholarly articles, researchers, and proceedings through crowd participation. A part of this work was used to populate papers from Semantic Web conferences to Wikidata as described in [8].

One of the limitations of this work is that the user has to move between the Web application and OpenRefine by transferring data using files. If OpenRefine is integrated within the web application using API level integration, this burden can be reduced. Such integration is planned as a future work.

There are several pieces of useful information that cannot be directly extracted from DBLP, such as the “main subject (P921)” and “cites work (P2860)”. We intend to use other sources and automatic extraction tools, including large language models, to populate this information.

The availability of information about scholarly articles in Wikidata will enable scientometrics [9] use cases. Furthermore, it will enable other practical use cases such as automatically generating lists of articles of a given authors or bibliographies in an automated manner or paper recommendations based on links in the KG.

We believe that easy-to-use tools will help make scholarly data more accessible by making them available in the sustainable and well-established Wikidata infrastructure. This work is a small contribution towards that goal.

[4] https://github.com/scholarly-wikidata/dblp-to-wikidata/blob/main/open_refine_schemas/scholarly_article_schema.json

[5] https://streamlit.io/

[6] https://dblp.org/search/author/api

[7] https://sparql.dblp.org/sparql

References

[1] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, Semantic web conference ontology-a refactoring solution, in: European semantic web conference, Springer, 2016, pp. 84–87.

[2] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, Conference linked data: the scholarlydata project, in: The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15, Springer, 2016, pp. 150–158.

[3] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. E. Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving Access to Scientific Literature with Knowledge Graphs, Bibliothek Forschung und Praxis 44 (2020) 516–529.

[4] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[5] D. Diefenbach, M. D. Wilde, S. Alipio, Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph, in: The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings, volume 12922 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 631–647.

[6] L. Rossenova, P. Duchesne, I. Blümel, Wikidata and wikibase as complementary research data management services for cultural heritage data, in: Wikidata 2022: Wikidata Workshop 2022, Proceedings of the 3rd Wikidata Workshop 2022 co-located with the 21st International Semantic Web Conference (ISWC2022), 2022.

[7] M. Ley, The dblp computer science bibliography: Evolution, research issues, perspectives, in: International symposium on string processing and information retrieval, Springer, 2002, pp. 1–10.

[8] N. Mihindukulasooriya, S. Tiwari, D. Dobriy, F. A. Nielsen, T. R. Chhetri, A. Polleres, Scholarly Wikidata: Population and Exploration of Conference Data in Wikidata using LLMs, in: 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), 2004.

[9] S. Kirrane, M. Sabou, J. D. Fernández, F. Osborne, C. Robin, P. Buitelaar, E. Motta, A. Polleres, A decade of semantic web research through the lenses of a mixed methods approach, Semantic Web 11 (2020) 979–1005. doi:10.3233/SW-200371.