ORCID for Wikidata – Data enrichment for scientometric applications ⋆

Eva Seidlmayer[1][0000−0001−7258−0532], Jakob Voß[2][0000−0002−7613−4123], Tetyana Melnychuk[3][0000−0002−7258−2842], Lukas Galke[4][0000−0001−6124−1092], Klaus Tochtermann[4], Carsten Schultz[3][0000−0002−5984−9872], and Konrad U. Förstner[1,5][0000−0002−1481−2996]

[1] ZB MED – Information Centre for Life Sciences, Cologne and Bonn, Germany
[2] Verbundzentrale GBV, Göttingen, Germany
[3] Institute for Innovation Research, Kiel University, Germany
[4] ZBW – Leibniz Information Centre for Economics, Kiel and Hamburg, Germany
[5] University for Applied Sciences, Cologne, Germany

Abstract. Due to its numerous bibliometric entries of scholarly articles and connected information Wikidata can serve as an open and rich source for deep scientometrical analyses. However, there are currently certain limitations: While 31.5% of all Wikidata entries represent scientific articles, only 8.9% are entries describing a person and the number of entries researcher is accordingly even lower. Another issue is the frequent absence of established relations between the *scholarly article* item and the *author* item although the author is already listed in Wikidata. To fill this gap and to improve the content of Wikidata in general, we established a workflow for matching authors and scholarly publications by integrating data from the ORCID (Open Researcher and Contributor ID) database. By this approach we were able to extend Wikidata by more than 12k author-publication relations and the method can be transferred to other enrichments based on ORCID data. This is extension is beneficial for Wikidata users performing bibliometrical analyses or using such metadata for other purposes.

Keywords: Wikidata · ORCID · data enrichment · bibliometrics · scientometrics

1 Introduction

The open knowledge base Wikidata is increasingly used also to integrate and use bibliographic data [16]. Scholarly articles account for almost a third of all Wikidata items, far beyond items about humans with roughly 9% [15] of entries.

During the bibliometric research project Q-Aktiv, we retrieved social context information on authors of scientific publications from Wikidata in order to

integrate the information into our data set. Of more than 8M scholarly publications published by 39.5M authors from the 1953 till May 2019 related to the topic of *cholesterol* we were able to find about 95% in Wikidata. However, meta data for only 3% of the authors could be retrieved and used downstream for the analysis of social context. In order to avoid the common problems of author disambiguation, we used the relation between author items and publication items for identification of authors. There we determined two major reasons for this observation: Missing Wikidata items for authors and the absence of represented relations publications and their authors in Wikidata.

To overcome this shortcoming and to improve the foundation for bibliographic analysis in general, we established a workflow for matching authors and ingesting ORCID data [9]. ORCID provides a authors centric, self-curated collection of scholarly metadata linked to an unique identifier. As of 2020, the ORCID database contains more than 9.6M researcher profiles and includes almost 62M scientific publications linked to these profiles [10]. The public information of each user's record is published under an open license and can contain details on publications, education, employments, funding as well as further information.

Here we will first present the context of the bibliometric research project in which the need for an improvement of Wikidata information appeared. Afterwards, we will describe, how we harvest ORCID for information on publications and their authors, how we query Wikidata for existing items that are also listed in ORCID and how we perform the matching to items. The benefits of such an ORCID based extension of Wikidata's corpus of scholarly inform especially the inclusion of further relations will become obvious.

2 Context and application: bibliometric project Q-Aktiv

Q-Aktiv is a collaborative project of ZB MED - Information Centre for Life Sciences, ZBW – Leibniz information Centre for Economics and Kiel University. The project aims for a better understanding of developing research areas applying a bibliometric analysis. We used scientific publication data annotated with Medical Subject Headings (MeSH) vocabulary by National Library of Medicine (NLM) hosted in ZB MED database "Knowledge Environment" to map topics and relevant documents. By doing so, we were able to observe if publications of two or more distinct research fields that did not share keywords, start doing it at a later time stage. This moving of research topics towards each other can be described as *convergence*. The opposing phenomenon of research fields is termed *divergence*. As a third option, a new research topic can evolve in the intersection of converging research area.

In the Q-Aktiv project, we first analyzed the use case *cholesterol* [3]. We found that among other observations the concept of *cardiovascular diseases* - as time goes by - enters the field of *cholesterol*. Analyzing the keywords allocated to the publications, we developed a learning based similarity measure for concepts of evolving research fields. In Q-Aktiv, we calculated the similarity as cosine distance between keywords [2]. A low cosine distance represents a close

relationship between topics that are annotated to the same papers, while a high cosine distance indicates a low similarity of topics.

The learning based network analysis is based on a knowledge graph containing publications, author names and keywords. However, apart from the descriptive observation, we were so far unable to identify reasons for the observed convergence. Still, we could speculate and developed new working hypothesis for this observation: We know that the development of scientific fields is driven by people and that people actively decide address certain rearch questions. Scientists decide to take up topics, and initiate or join collaborative projects. Based on these consideration, the questions emerged who those researchers are who lead to changes of research fields? Does a change in the researchers social environment result in a change of research questions and influence convergence and divergence of scientific fields? Is science – not completely but partially – influenced by a shifting social composition of the group of researchers? Investigations regarding networks, citation behaviour, or social conditions of publication, in particular, would benefit from more information related to the authors and research groups [8]. One main difficulty is generating a solid data foundation for substantiated analysis [4]. The inclusion of additional data sources would broaden the basis of infometric analyses and contribute to a consolidation of knowledge. However, we currently face the limitations of existing tools and accessible sources. To obtain a deep understanding of the research fields development, the integration of social context information such as affiliation, gender or education into the analysis can be beneficial. Luckily, such information can be found in Wikidata.

3 Wikidata for bibliographic data enrichment

In order to introduce social context information, we haved developed a library written in Python. The library which named "Take it personally" (T.I.P.) supports enrichment of author information based on Wikidata [11,13]. We chose Wikidata as a source since it provides numerous links via publication identifiers such as DOI (Digital Object identifier), PMID (PubMed ID) and PMC (PubMed Central). In addition, Wikidata's query API support an easy retrieval of information [7]. Apart from the *scholarly article* (Q13442814, [5]) items, authors are represented by the *author* property (P50, [6]) items, which allows to switch from the publication centric view in our bibliographic data source to an author centered view. Wikidata's community approach enables individuals to correct their own entries. With regards to sensible information such as gender, Wikidata introduces queer gender declarations in addition to male and female.

Applying our T.I.P. library, we enriched several data collections based on Wikidata. Overall, we experienced a low coverage of identified authors of publications. For several data sets complied for the topic *cholesterol* consisting of more than 14M, 8M, and 99k publications, we were able to detect authors for less than 5% of the publications. As comparision: The coverage for a recent *COVID-19* data set (23k publications with 138k authors) was with 13% significantly higher.

We assume that there are two reasons for the low coverage: First, the *cholesterol set* contains older publications which are not as well covered and curated as current digitally available publications. The very recent *COVID-19* set confirms this assumption with an almost tripled coverage. A second reason are that several conditions are needed to be fulfilled for the identification of authors: On the one hand, the *publication* itself needs to be registered, on the other hand, the *author* needs to be registered in Wikidata as well. Furthermore, a *link between both items* has to be established for the retrieval of information on authors originating from the publication. While tools for manually creating these links exist [12], an automatic process would be preferable.

4 Establishing publication-author matches with ORCID

To improve the Wikidata based retrieval of social context information on scientific authors, we tried to establish a general approach that is beneficial for a broader community and does not only solve our issue.

ORCID was determined by us as data source that can help to fill the information gap. It contains a large number of researchers with connections to their publications and provides a persistent identifier for these researchers. It can only be created by the researchers themselves and is curated by the person or its institution. The entry may include information regarding the professional CV containing details on publications, education, funding, employment and memberships. Since those statements are made by the researchers themselves the data is highly trustworthy – though the list of literature does not has to be complete. The ORCID organisation publishes the public information under a Creative Commons Zero license (CC0) which allows as frictionless reuse.

The available data snapshop of 2019 contains 673k individuals who have an ORCID account and associated information. 134k of the ORCID ID could be mapped to available Wikidata items.

The rich collection of ORCID could be used to extend Wikidata by numerous further publications.

However, we decided, also after discussion with other members of the Wikidata community, to only use the data set for improving available items. The main argument was to avoid producing a large number records with minimal metadata and no links to or from other items. Bearing also the finit capacity of Wikidata in mind, we therefore focus on *matching existing items* in order to improve and condense the existing data by favouring to introduce rather relations between items than more thin items themselves. Linked data sets live from the connections, not from the actual number of items.

5 Estimating the potential of enrichments

For the purpose to estimate the potential of ORCID for matching authors and publications we extracted all publications listed in ORCID-archive-occurrence in Wikidata. The ORCID public data set for 2019 consists of a meta file called

summaries and eleven *activities* files [1]. In the first of the eleven activities files, we had been able to identify 3.8M publications by DOI, PMID, PMC, DNB (Deutsche National Bibliothek ID) and eid (Scopus ID).

For this subset 457k associated publications could be found in Wikidata but only 32k of those had author item linked to them. This means by far the larger share of 425k articles items were not connected to any author item. It is possible that the authors are recorded in the publication item but only with a plain string (Property *author name string* (P2093, [17]), not *author* (P50, [6]). Though, such information cannot be automatically translated into link to author items. At this point it would be easy to introduce the researcher who recorded the publication to ORCID as author to the Wikidata publication item. As mentioned before, the discussion with other members of the Wikidata community revealed that no new items should be created. This is why we limited the approach to authors with existing Wikidata items.

For the 32k publication items recorded with author items in Wikidata, we had to find out if the associated author information contains the known author. Since the publication is extracted from the ORCID file it is still linked to the researcher who declared it as her own or his own. This connection can be taken into account by checking whether the initial author is correctly entered to the publication item.

All together the eleven activities files bear the following content: More than 34M publications registered identified by one or more identifier such as DOI, PubMed id, PubMedCentral id, Scopus id, or DNB id and claimed by the author as own work. Based on this, it was possible to identify more than 2.6M publications registered in Wikidata. For more than 820k of those publications the authors could be detected. Unfortunately, we were facing issues due to the limitation of the Wikidata API which was returning inconsistent numbers of result items for the same requests. We tried to fix this by requesting small chunks of data and multiple passes but still could not solve the issue. This shortcoming of the Wikidata API is also reported by others.

6 Data preparation and submission

As a result of the analysis, we created CSV files containing the following identifiers: *author item identifier*, *ORCID ID*, *given name*, *family name*, *article item identifier* and *multiple item identifiers of all publication-authors*. Based on this, we created JSON templates containing the required properties. Wikidata implements a publication centered view of the semantic data. There is no property such as *is author of* but the conection starting from the publication implemented by the *has author* propert (*has author*). The property can be qualified with the author item identifier. In addition, also P1932, *has author string*, can be added and qualified with the string of the author.

Applying the tool Wikibase CLI [14] for creating a bot, we submitted the data as JSON to introduce the information to the publication entries into Wikidata. Afterwards the matching was performed. In total it was possible to includes

more than 948k articles and detect more than 792k authors which did not have an item in Wikidata yet. We found more that 47k authors were listed correctly. Using our bot we had been able to introduce more than 12k authors to the paper items as their originators.

The code for implementing the workflow as Shell and Python Scripts is deposited at Zenodo and can be retrieved at https://doi.org/10.5281/zenodo.4088048.

7 Conclusion and outlook

We have developed an efficient approach to improve the author publication links in Wikidata base on data from the ORCID database. As the ORCID researchers enter statements on their own publications, the information is trustworthy and problems of author disambiguation are avoided. Based on this established workflow we can easily introduce any other information included in the ORCID database. This can be direct statements on researcher items regarding aspects of the scientific biography and current research activity. Analogue to the approach of interweaving existing items presented here we could also introduce for example relations between organizations and researchers. It needs to be discussed in the community weather this is the kind of information that Wikidata carries further.

Soon there will be the release of the new ORCID data set for 2020 and we will continue our integration of such data into Wikibase.

Acknowledgements

This work was supported by BMBF of Germany within the program *Quantitative Wissenschaftsforschung* under grant numbers 01PU17013A, 01PU17013B, and 01PU17013C.

The work was worked out in major parts within the fellowship program *Freies Wissen (Free Knowledge)* 2019 / 2020 by Wikimedia Germany, Stifterverband, and Volkswagenstiftung.

References

1. Figshare: ORCID Public Data File 2019, https://doi.org/https://doi.org/10.23640/07243.9988322.v2
2. Galke, L., Melnychuk, T., Seidlmayer, E., Trog, S., Förstner, K. U., Schultz, C. & Tochtermann, K., (2019). Inductive Learning of Concept Representations from Library-Scale Bibliographic Corpora. In: David, K., Geihs, K., Lange, M. & Stumme, G. (Hrsg.), INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft. Bonn: Gesellschaft für Informatik e.V.. (S. 219-232). DOI: 10.18420/inf2019_26
3. Melnychuk, T., Galke, L., Seidlmayer, E., Förstner, K., Tochtermann, K., Schultz, C.: Development of Similarity Measures from Graph-Structured Bibliographic Metadata: An Application to Identify Scientific Convergence. Manuscript submitted for publication for Scientometrics (2020)

4. Mihaljević, H., Tullney, M., Santamaría, L., Steinfeldt, C.: Reflections on Gender Analyses of Bibliographic Corpora. Frontiers in Big Data **2**(29) (2019), from https://www.frontiersin.org/article/10.3389/fdata.2019.00029/full, https://doi.org/10.3389/fdata.2019.00029
5. https://www.wikidata.org/wiki/Q13442814
6. https://www.wikidata.org/wiki/Property:P50
7. Mitraka, E., Waagmeester, A., Burgstaller-Muehlbacher, S., Schriml, L. M., Su, A. I., Good, B. M.: Wikidata: A platform for data integration and dissemination for the life sciences and beyond. BioRxiv (2015), from https://doi.org/10.1101/031971, https://doi.org/10.1101/031971
8. Nielsen, F. A., Mietchen, D., Willighagen, E.:Scholia and scientometrics withwikidata. https://zenodo.org/record/1036595.XThTQvyxU5k (2017)
9. ORCID for Wikidata GitHub Repository, https://github.com/EvaSeidlmayer/orcid-for-wikidata. Last accessed 29 Jul 2020
10. ORCID: ORCID Statistics, https://orcid.org/statistics. Last acccessed 08 Oct 2020
11. Seidlmayer, E., Galke, L., Melnychuk, T., Schultz, C., Tochtermann, K., Förstner, K. U.: Take it Personally - A Python library for data enrichment in infometrical applications. In: SEMANTICS Posters & Demos 2019, Karlsruhe (2019)
12. Smith, A.: Author Disambiguator. 2020 https://author-disambiguator.toolforge.org/
13. TIPlib, Github Repository, https://github.com/foerstner-lab/TIPlib. Last acccessed 12 Aug 2020
14. Wikibase-CLI, GitHub Repository, https://github.com/maxlath/wikibase-cli. Last accessed 29 Jul 2020
15. Wikidata Statistics 2020, https://www.wikidata.org/wiki/Wikidata:Statistics. Last accessed 29 Jul 2020
16. Wyatt, L., Ayers, P., Proffitt, M., Mietchen, D., Seiver, E., Stinson, A., Taraborelli, D., Virtue, C., Tud, J., Curiel, J.: WikiCite Annual Report, 2019–20 (2020) https://doi.org/10.5281/zenodo.3869810
17. https://www.wikidata.org/wiki/Property:P2093