Take it Personally - A Python library for data enrichment in informetrical applications*

Eva Seidlmayer[1][0000−0001−7258−0532], Lukas Galke[2][0000−0001−6124−1092], Tetyana Melnychuk[3][0000−0002−7258−2842], Carsten Schultz[3][0000−0002−5984−9872], Klaus Tochtermann[2], and Konrad U. Förstner[14][0000−0002−1481−2996]

[1] ZB MED Information Centre for Life Sciences, Cologne, Germany
seidlmayer@zbmed.de
[2] ZBW - Leibniz Information Centre for Economics, Kiel and Hamburg, Germany
[3] Kiel University, Germany
[4] TH Köln - University for Applied Science, Cologne, Germany

Abstract. Like every other social sphere, science is influenced by individual characteristics of researchers. However, for investigations on scientific networks, only little data about the social background of researchers, e.g. social origin, gender, affiliation etc., is available.
This paper introduces "Take it personally - TIP", a conceptual model and library currently under development, which aims to support the semantic enrichment of publication databases with semantically related background information which resides elsewhere in the (semantic) web, such as Wikidata.
The supplementary information enriches the original information in the publication databases and thus facilitates the creation of complex scientific knowledge graphs. Such enrichment helps to improve the scientometric analysis of scientific publications as they can also take social backgrounds of researchers into account and to understand social structure in research communities.

Keywords: · data enrichment · informetrics · scientometrics · Python

1 Background: Author orientation of metadata in scientometrics

Research fields evolve. As part of our project Q-Aktiv, that aims to study the convergence for scientific areas, we investigated the research field of cholesterol applying network-analyses on MeSH-term indicated papers from Medline [13]. We could observe that the topic first occurred in the context of cardiovascular diseases and nutrional studies [6, 4]. Since the 1970s, an increasing number of publications using keywords concerning gynecological pathologies indicate a shifting interest in the research on cholesterol. However, who are those researchers who

* Supported by BMBF (grant numbers 01PU17013A, 01PU17013B, 01PU17013C)

had been interested in cardiovascular diseases? And, who are the researchers in recent times concentrating on gynecology? Are they a comparable group of researchers just shifting in topics? Or, does a change in the social group of researchers (e.g. due to the increasing number of women in sciences in the last decades) result in a change of research questions?

As every other social sphere, science is influenced by social structures. The outcomes of the investigations of history of sciences emphasize the social impact on scientific investigation for a long time. Already, Ludwik Fleck described social "thought collectives" and conventions in the use of language ("thought style") as a major influence to the work in medical laboratories [3]. Also, Derek de Solla Price realized distinct social groups, in science of newcomers and veterans, who show different behaviour in publishing and citing [14]. The infiltration of social norms into science, which is supposed to be objective and solely justified by reason, is also widely described in social science. According to broad-based investigation on intersectionality, we have to assume that factors such as gender, class, ethnicity and others influence behaviour in research (e.g.[2]). Further research in Psychology deals with racial privileges that, also in academic communities, lead to a majority of white privileged individuals ([11]). If we want to understand the mechanism of science, then we need to also understand the social structures that researchers are acting within.

Scientometric analyses, such as the analysis on cholesterol, usually rely on meta data provided by databases such as Web of Science, Scopus or Medline. Investigations regarding networks, citation behaviour, or social conditions of publication, in particular, would benefit from more statements related to the authors and research groups [8]. The inclusion of data sources such as Wikidata [18], ORCID [10] or CrossRef [1] would broaden the basis of informetric analysis and contribute to a consolidation of knowledge. Here we are facing the limitations of existing tools.

Our contribution to the described challenge in scientometrics is a Python library - "Take it personally" (TIP) - that aims to facilitate a more author-related view on informetric research by retrieving social information on authors of publications on a large scale. Thus, not only the single author becomes visible behind her publication, but also broader social analyses shall become possible.

We are aware that, for domain specific research, personal details might not be necessary and could even corrupt an unbiased view on disciplinary topics. However, for meta-analyses in contrast, it is important to understand the reasons for success and failure of research activities or scientific ideas.

2 Related work

There had been some work on a personalizing publication data e.g. concerning gender [5]. Since our investigation on research dynamics seeks for more information than gender, we will focus on an enhancement of statements altogether with other aspects.

Accumulated in the service "Scholia", several services had been developed relying on Wikidata [12]. Scholia provides a range of statistical analysis on the scientists, papers, organizations, venues, events or topics. The project of Scholia gives a good example of analyses that can be performed with Wikidata. Furthermore, Scholia focuses on a close view on the single researcher and does not offer large scale analysis of the data on a meso level of publication networks.

3 "Take it personally" (TIP) library for Python

3.1 Overview: making the authors visible

The Python library TIP that we designed and started to implement will enable clients to retrieve information for authors, institutions, and journals. We follow a pythonic approach that eliminates the need for client-side SPARQL queries. The enrichment of bibliographic data should require not more than a single function-call. By removing these obstacles, we aim to reduce the effort that is required for conducting large-scale meta-research. We envision that a multitude of studies can profit from such a library for dynamic data enrichment.

The library's initial internal step contains the input of an identifier that allows to identify the desired item. The second step is the retrieval of features. Lastly, the retrieved attributes need to be added to the dossier of characteristics of the single instances to create complex scientific knowledge graphs.

3.2 Input and Identification: Identifiers and Wikidata as first data source

Applying identifiers, TIP will enable to create instances of three classes – authors, institutions, and journals. The assignment to items mainly depends on the presences of identifiers that allow a clear allocation of data sets.

For the library, DOIs of articles, VIAF, ISNI or ORCID-Identifiers can be used to retrieve information on authors. PubMed IDs or DOIs, can be taken to recall articles. With ISSN journals and institutions can be called.

As a first access point for the retrieval Wikidata was chosen since it supports different identifiers applied in publication data. The number of identifiers registered in the data source increases constantly. From 2018 to 2019 a growth of about 20% of all common identifiers can be found. For ORCID it is even larger with more than 300% of new entries. Furthermore, the variety of identifiers facilitate the evaluation and deep linking to other platforms [8]. Currently, in June 2019, Wikidata contains more than 57 million items [17]. According to Wikidata statistics, scholarly articles take up more than 42% of all items currently while close to 10% of the data sets cover humans [16]. We calculated the number of provided identifiers within the current Medline 2019 dataset: Medline contains more than 1,038,000 ORCIDs and nearly 203,500 ISNIs. ORCIDs, ISNIS and VIAFs have only be recorded since 2013 [9]. We sampled 5000 ORCIDs from Medline and found that 26.88% are also registered in Wikidata. Therefore, deploying Wikidata as data source for TIP-library can only be a start and needs to be supplemented by other data sources hereafter.

3.3 Retrieval of features

To perform the queries, TIP relies on the provided Wikidata-API. Using the Python library SPARQLWrapper [15], the API returns to our SPARQL formulated queries in JSON. Specific features can be requested but also the on-bloc query and the enhancement of a data dump is going to be supported. At this stage of the development the following features can be reached: for "authors": the "gender", "ORCID", "ISNI", "affiliation" and the "parents". We ask for affiliation because the working places can tell a lot about conditions of work. The choice "parents" was made due to the observation that many successful researchers come from families that include many other successful researchers. This phenomenon of "academic dynasties" can be addressed by requesting the parents of an author.

The properties describing the class "institution" contains the characteristics "country","students count", "tuition" and the "type" of organization as a research institute or a public or private university. The class "journal" combines the "country of origin", what is specific for journals within Wikidata, the "publisher", a possible "review score" and the "main subject". Other attributes can be made available in the future on the basis of Wikidata or other data sources.

4 Discussion

Here we presented the concept of a Python library for the large scale analysis of author information which will make it easy to extend scientometric studies by these aspects. The library itself is in its early implementation stage and uses Wikidata as it data source. A frequently expressed concern according the implementation of Wikidata, is the shortage of authors and paper records compared to publication databases. With respect to the fast expanding content mentioned above and the growing community that reflects the increasing interest in Wikidata, the problem might solve itself over time. However, newcoming authors in the scientific scene will always be difficult to record. Yet, since researchers have a general interest in being visible with their work within academia we can anticipate an increasing data resource for authors in the future. Wikidata is the suitable access point for this goal [7]. However, we are aware that other data sources as ORCID or CrossRef needs to be implemented. ORCID contains dense biographical information while CrossRef offers event data including information on social network activities.

Furthermore, the retrieval of information on the basis of the author's full name would be a desirable feature of TIP, yet it comes with all difficulties author disambiguation struggles with. The task of author disambiguation is a general difficulty for bibliometric analyses. However, the most feasible approach for TIP seems to be the self-identification of authors as it is provided by ORCID. Apart from ORCID, libraries supply entity disambiguation, for instance via ISNI or VIAF.

5 First results

TIP-library is still in an early stage of development but probably become a powerful library to compile and retrieve social data from different sources for easy analyses in scientometric investigations. By using Wikidata as a first data source that combines many common identifiers, we are currently able to address more than 26.8% of the author with ORCIDs in the current Medline 2019 snapshot. Other identifiers will complement the coverage. A general improvement of the coverage of Wikidata can also be expected due to the fast expanding amount of data sets.

Source Code: github.com/foerstner-lab/TIP-lib

References

1. Crossref: You are crossref - crossref (2019), https://www.crossref.org/
2. Degele, N., Winker, G.: Intersektionalitt als Mehrebenenanalyse (2007)
3. Fleck, L.: Entstehung und Entwicklung einer wissenschaftlichen Tatsache (1980)
4. Galke, L., Melnychuk, T., Seidlmayer, E., Trog, S., Frstner, K.U., Schultz, C., Tochtermann, K.: Inductive learning of concept representations from library- scale corpora with graph convolution. In: INFORMATIK. Gesellschaft für Informatik, Bonn (2019)
5. Iefremova, O., Wais, K., Kozak, M.: Biographical articles in scientific literature: analysis of articles indexed in web of science **117**(3), 1695–1719 (2018), https://doi.org/10.1007/s11192-018-2923-3
6. Melnychuk, T., Galke, L., Seidlmayer, E., Wustmans, M., Tochtermann, K., Förstner, K.U., Bröring, S., Schultz, C.: Analyzing scientific dynamics does machine learning help to predict scientific convergence based on bibliographic data? (2019)
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 p. 031971 (2019). https://doi.org/10.1101/031971
8. Nielsen, F.A., Mietchen, D., Willighagen, E.: Scholia and scientometrics with wikidata (2017), https://zenodo.org/record/1036595.XThTQvyxU5k
9. NLM: MEDLINE/PubMed data element descriptions (2019), https://www.nlm.nih.gov/bsd/mms/medlineelements.html
10. ORCID: ORCID - Connecting Research and Researchers (2019), https://orcid.org/
11. Phillips, L.T., Lowery, B.S.: Herd invisibility: The psychology of racial privilege p. 156162. https://doi.org/10.1177/0963721417753600
12. Scholia: Scholia (2019), https://tools.wmflabs.org/scholia/
13. Schultz, C.: Q-Aktiv, https://www.wihoforschung.de/de/q-aktiv-2178.php
14. de Solla Price, D.: Little science, big science ...and beyond (1986)
15. SPARQL-Wrapper: SPARQL endpoint interface to python (2019), https://rdflib.github.io/sparqlwrapper/
16. Wikidata: Statistical hub (2019), https://www.wikidata.org/wiki/Wikidata:Statistics
17. Wikidata: Statistics (2019), https://www.wikidata.org/wiki/Special:Statistics
18. Wikidata: Welcome to Wikidata (2019), www.wikidata.org/wiki/Wikidata:Main$_{Page}$

All online references had been lastly accessed on July 24, 2019.