Chapter 8

Author identification through and for interconnectivity: A brief history of author identification at zbMATH Open

Nicolas D. Roy

1 Introduction

Author identification, or *author disambiguation*, is the process of matching an *authorship record*, i.e. an author name string in a publication, with an entity in a database of persons. The person entity is usually defined by a set of metadata, like for example name, birthdate, affiliation, email, etc., but also by a collection of other publications authored by the person. The link between authorship record and person entity is called an *authorship assignment*.

The task of attributing a given publication to a certain author based only on person name is notoriously difficult because of the following reasons:

- *Incompleteness*. The name contained in an authorship record may be abbreviated (often the given name) or even missing (e.g. a second/middle name). In the last centuries it was also not uncommon to publish under the family name only. On the other hand, it is almost impossible to completely avoid any kind of data corruption, and that could lead to an erroneous author name in a publication.
- *Synonymy*. Different names might refer to the same individual. A name change after marriage is a common source of such name variability, but zbMATH Open provides many other examples due to the use of different historical transliteration rules, e.g. in the Russian literature.[1]
- *Homonymy*. Different persons might bear the same name. The most famous examples come probably from the Chinese language, since just the top three surnames *Wang* (王), *Li* (李), and *Zhang* (張; 张)[2] cover more than 20% of the population [2]. But European languages provide also examples like *Peter Müller*[3] or *Andrzej Nowak*[4].

The last decades have seen the emergence of various person-centered databases. As a result, interconnectivity has become a staple feature of many online services and

[1] The case of https://zbmath.org/authors/chebyshev.p-l is a particularly illustrative example of this phenomenon

[2] See for example https://zbmath.org/authors/?q=wang.wei

[3] https://zbmath.org/authors/?q=müller, peter

[4] https://zbmath.org/authors/?q=nowak,andrzej

information databases. In the domain of disambiguation of publication authorships, interconnectivity is on the one hand a very valuable by-product of the author identification, since well-identified author entities allow for a reliable matching with other similar person-based databases. But on the other hand, interconnectivity itself can constitute a powerful component of any author disambiguation system, through the harvesting of additional data relevant to author disambiguation (biographical, bibliographic, ...) from other related services, as soon as a reliable matching between both databases can be established.

2 Author identification workflow at zbMATH Open

As of October 2021, zbMATH Open indexes approximately 4.3 million documents, corresponding to about 8 million authorship records that are linked to the author database, containing currently more than 1.1 million items.

The author disambiguation at zbMATH Open is the result of a complex interaction between several algorithmic tools and user interfaces, where each module enriches the capabilities of the others.

Automatic disambiguation

Approximately 78% of the authorship records are handled by an algorithm mainly based on a name-similarity feature, which is well fitted to the particularities of the zbMATH Open dataset (taking into account for example the variability in transliterations from Cyrillic script). This name similarity is refined by time and coauthor similarity features. The time-similarity feature relies heavily on the presence of biographical metadata associated to the author entities. This biographical data is for a big part harvested from external services, as described in the next section. The automatic identification process runs daily in order to handle the newly indexed documents (ca. 600 per day), but also because, in principle, the disambiguation of a given document might have some influence on the disambiguation of other documents on the next day.

Community Author Identification Interface

Like every automatic process, the author disambiguation algorithm produces various errors: authorship records wrongly attributed to another author, publications of a given person incorrectly split into several author profiles (as a typical result of the above-mentioned 'synonymy' issue), or publications of several persons incorrectly mixed into one single profile (as a typical result of the above-mentioned 'homonymy' issue).

To circumvent that, zbMATH Open has been offering since summer 2014 an Author Identification Interface [1], accessible through the button *Edit Profile* at the top right of any author profile, and allowing every zbMATH Open user to send some correction requests. Since then, more than 8,000 user requests have been sent with this tool. Besides the possibility to correct authorship assignments or to merge together several wrongly split author profiles, the interface allows also for adding *external links* to other related services, like e.g. Wikidata, ORCID, Mathematics Genealogy Project (MGP), etc. (see Figure 1).

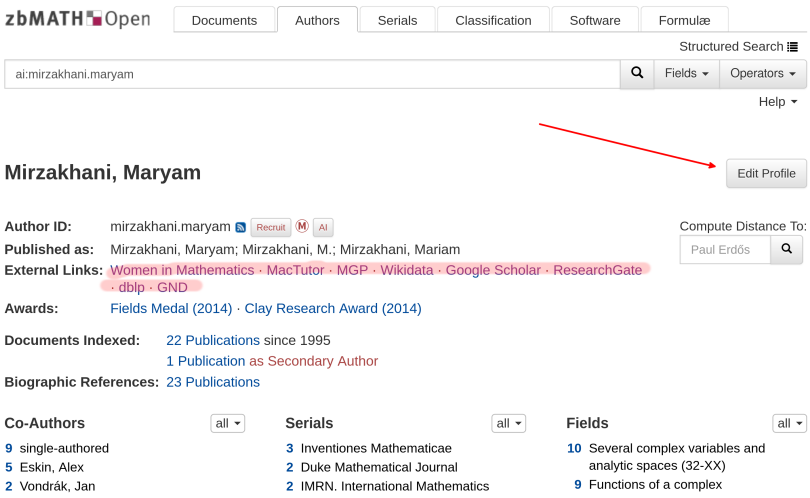

Figure 1. Maryam Mirzakhani's author profile showing several *external links* as well as the button to enter the Author Disambiguation Interface.

Every correction request is examined by the zbMATH Open Author Identification team and is visible at zbmath.org usually within 1 or 2 days.

3 Interaction with the scientific publication landscape

Matching processes

The external links provided by community users through the public author identification interface are complemented by several matching tools, based on name similarity but also on various other features adapted to the data available in the considered services. For example, the ORCID matching (ca. 30,000 matched items) highly relies on DOIs, while the matching with MGP (ca. 50,000 matched items) is based on a *collaborator similarity* between the publication coauthors on the zbMATH Open side and the PhD advisors and students on the MGP side.

Data harvesting and spreading tools

The connection with other relevant services[5] not only increases the visibility of the respective services and the Internet presence of the authors,[6] but also allows for very valuable mutual data enrichment and data quality improvement. This is achieved through several automatic data harvesting and spreading tools:

- *Data spreading tools*. An automatic process checks every night the presence of new wikidata IDs in the zbMATH Open author database, and incorporates the involved zbMATH Open author IDs into the corresponding Wikidata profile when necessary.
- *Data harvesting tools*. Every night, the external services linked to any of the zbMATH Open author entities are queried for the presence of any new relevant information, like biographical data (birth year, PhD year, etc.), scientific information (awards, etc.), or other external links.

Integration of harvested data into author disambiguation

The additional data automatically harvested from partner services is then incorporated into the corresponding zbMATH Open author entity, and it is subsequently used to improve the performance of the algorithmic author identification (particularly the time-similarity feature).

Figure 2 is a hypothetical but realistic example of how the combination of matching processes and data harvesting can lead to major changes in the author disambiguation, as described below:

- Assume that at the beginning, many papers authored by a person named 'María López' published in the years 1978, 1983, 1990, 2001, 2011, 2013, and 2017, are grouped together by the author identification algorithm in a profile with id `lopez.maria`.
- Suppose that in Step 1 a matching with the ORCID database would return an ORCID entity with ID `1234-5678-9876-5432` because of name similarity and DOI concordance with the later papers (2011–2017).
- The data harvesting tool would then in Step 2 query the Wikidata API and may find a suitable Wikidata person entity with ID `Q12345678`, together with a MGP entity with ID 343434.

[5]The currently supported services are: http://www.mathnet.ru, https://mathscinet.ams.org, https://dblp.uni-trier.de, https://www.wikidata.org, https://orcid.org, https://portal.dnb.de, https://www.idref.fr, https://www.mathgenealogy.org, https://www.researchgate.net, https://scholar.google.com, https://arxiv.org, https://mathoverflow.net, https://celebratio.org, https://mathshistory.st-andrews.ac.uk, https://www.agnesscott.edu/lriddle/women/alpha.htm,

[6]Currently there are more than 220,000 external links in zbMATH Open author profiles

- In Step 3, the MGP database would be queried and the PhD year 2012 of this entity 343434 would be fetched and incorporated into the zbMATH Open author profile `lopez.maria`.
- This important biographical data would then be exploited in the next run of the author disambiguation algorithm, and would lead to the exclusion of the earlier papers (1978–2001) from the profile `lopez.maria` for time incompatibility. This would result in a splitting of the original author profile into two separate author entities with same name 'María López', where the first one (`lopez.maria.1`) would contain the later papers (2011–2017), the ORCID, Wikidata and MGP IDs and the PhD year, whereas the second author entity (`lopez.maria.2`) would gather the older papers (1978–2001).

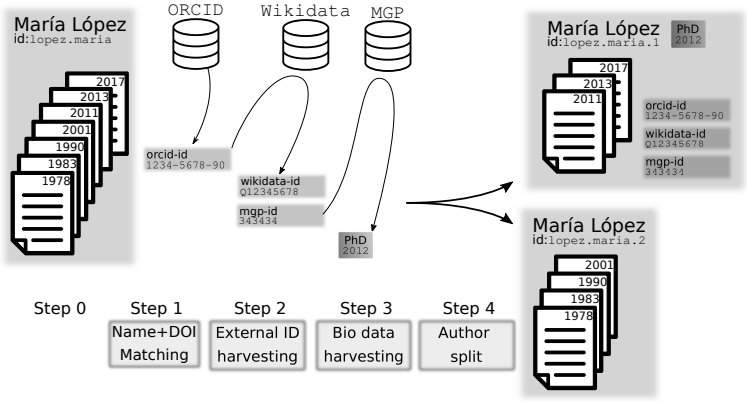

Figure 2. Integration of different harvested data into the author disambiguation.

4 Future developments

The current implementation of the author disambiguation algorithm at zbMATH Open has somehow reached its limit. In particular, the constant adjustment and fine-tuning of the name-similarity procedure to the peculiarities of the zbMATH Open dataset (e.g. the high name variability occurring in transliterations from Russian), has led to a very heuristic and possibly over-fitted algorithm, which is difficult to maintain and improve.

Acknowledging this observation(s) was the starting point of the project *ScAD* (Scalable Author Disambiguation for Bibliographic Databases)[7] in cooperation with *Schloß Dagstuhl* and the *Heidelberg Institute for Theoretical Studies*, launched in

[7] https://www.dagstuhl.de/en/about-dagstuhl/projects/author-disambiguation/

2015, whose legacy is the development of a new framework for author disambiguation (called *KafkAdam*), fully parallelized, highly modular and scalable. It will allow to improve the quality of the author disambiguation through the easy integration of new similarity features (topic analysis, citations, etc.), the possible use of machine learning, and a more efficient and more automated interconnection with other services.

References

[1] H. Mihaljević-Brandt and N. Roy, zbMATH author profiles: Open up for user participation. *Eur. Math. Soc. Newsl.* **93** (2014), 53–55

[2] Wikipedia, Chinese name. https://en.wikipedia.org/wiki/Chinese_name visited on 14 March 2024