A Novel Method for Resolving and Completing Authors' Country Affiliation Data in Bibliographic Records

Ba Xuan Nguyen^{1,2†}, Jesse David Dinneen^{1,3}, Markus Luczak-Roesch^{1,4}

¹School of Information Management, Victoria University of Wellington, Wellington, New Zealand

²Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam

³Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Berlin, Germany

⁴Te Pūnaha Matatini - The New Zealand Centre of Research Excellence for Complex Systems and Networks, New Zealand

Abstract

Purpose: Our work seeks to overcome data quality issues related to incomplete author affiliation data in bibliographic records in order to support accurate and reliable measurement of international research collaboration (IRC).

Design/methodology/approach: We propose, implement, and evaluate a method that leverages the Web-based knowledge graph Wikidata to resolve publication affiliation data to particular countries. The method is tested with general and domain-specific data sets.

Findings: Our evaluation covers the magnitude of improvement, accuracy, and consistency. Results suggest the method is beneficial, reliable, and consistent, and thus a viable and improved approach to measuring IRC.

Research limitations: Though our evaluation suggests the method works with both general and domain-specific bibliographic data sets, it may perform differently with data sets not tested here. Further limitations stem from the use of the R programming language and R libraries for country identification as well as imbalanced data coverage and quality in Wikidata that may also change over time.

Practical implications: The new method helps to increase the accuracy in IRC studies and provides a basis for further development into a general tool that enriches bibliographic data using the Wikidata knowledge graph.

Originality: This is the first attempt to enrich bibliographic data using a peer-produced, Web-based knowledge graph like Wikidata.

Keywords International research collaboration measurement; Bibliographic data; Country identification; Knowledge graphs; Wikidata; Open data

Abbreviations

ACM DL: ACM Digital Library; FMR: false match rate; FNMR: false non-match rate; IRC: international research collaboration; MAG: Microsoft Academic Graph

[†] Corresponding author: Ba Xuan Nguyen (E-mail: ba.nguyen@vuw.ac.nz).

Citation: Nguyen, Ba Xuan, Jesse David Dinneen, and Markus Luczak-Roesch. "A novel method for resolving and completing authors' country affiliation data in bibliographic records." *Journal of Data and Information Science*, vol. 5, no. 3, 2020, pp. 97–115. <https://doi.org/10.2478/jdis-2020-0020>

Received: Feb. 1, 2020

Revised: May 18, 2020

Accepted: Jun. 11, 2020

1 Introduction

As publications resulting from international research collaborations receive, on average, a higher number of citations (Glänzel & Schubert, 2001), researchers are incentivized to collaborate across national borders. As a result, both the number and ratio of internationally co-authored papers have risen (Fortunato et al., 2018). Research managers and policymakers are therefore interested in measuring international research collaboration (IRC; Luukkonen et al., 1993), for example, to determine if policies to boost such collaborations are effective, but as we discuss here the measurement of IRC is often not straightforward and has been relatively unexamined.

IRC measurement needs the task of country identification in bibliographic data sets because of some *de factos*. Firstly, counting instances of pairs of countries is the common approach to determine the extent of collaboration between countries. It requires mapping each researcher involving in research collaboration to the relevant countries. While this approach is simple in principle, it is complicated in practice because *identifying* countries from the researcher's description is not straightforward. Secondly, affiliation information provided in publications has been the default source to identify which countries have collaborated, although there are different alternatives according to the various possible definitions of IRC (e.g. noting the countries of authors' alma maters or their countries of birth). In other words, the convention for identifying countries in IRC measurement is to use the countries of authors' institutions at the time of publication. To demonstrate the magnitude of the convention, note that all but two of the 32 core IRC studies (Chen, Zhang, & Fu, 2018) use authors' listed affiliations. The exceptions (Beaver & Rosen, 1978, 1979) used authors' nationalities because of the narrow context of the study (identifying external collaborators with the professionalized scientific community in France). Presumably, as the availability of bibliographic data sets and affiliation data improved over time (e.g. in SCI/Web of Science), all other studies have used authors' affiliation information rather than their nationalities.

The above *de factos* in IRC measurement associate with certain complications. The task of identifying countries is non-trivial because the quality of bibliographic data sets is not often consistent or complete. For example, there are many records with varying affiliation formats (e.g. ending with country, or with state/province, or just an institution like "McGill University") or dirty data (e.g. ending in "#TAB#"). In short, measuring IRC is desirable, but currently difficult. The missing of explicit information for country identification, which is essential in IRC measurement, may affect the accuracy of the result. In these studies, co-authored publications are

counted and attributed to corresponding nations. In order to do that, these studies need to apply a string matching algorithm for country identification, which is manually coded rules to identify appropriate countries from affiliation information in bibliographic data. The result of this string matching algorithm is, therefore, far from perfect for the reason mentioned above (i.e. the absence of country identification in commonly used data sets). As a consequence, previous IRC studies using bibliographic data could only take into consideration those documents including clear information for country recognition. In short, the use of string matching is not sufficient and efforts to resolve these missing values may help to give a more detailed picture of the IRC landscape. In addition, there is no standard method for associating the affiliation data from each publication with the parent country. Methods for doing this are generally implicit or poorly described in prior works, such that every new study requires authors to make design decisions and repeat implementations for these methods.

Here we describe a method to address the above complications in identifying countries to measure IRC, and evaluate the method's magnitude, accuracy, and consistency using both general and domain-specific sets of bibliographic data.

2 Foundations and related work

Many studies have recognized the incompleteness of location information in author affiliation metadata that occurs in bibliographic records. Hinze (1994) came to the conclusion that 17% of documents for the period 1985–1993 have no easily discernible country identification. More recent studies presented similar results. Jacsó (2009) explored the extent of the absence of data elements in the Scopus and Thomson-Reuters databases, which are the two major multidisciplinary databases that exist today (Martín-Martín et al., 2018). They found that the rate of missing country data is high (e.g. there is a 34% omission rate of country metadata in Scopus, 14% in Thomson-Reuters). Regarding missing address information, Liu, Hu, and Tang (2018) concluded that more than one-fifth of publications in Thomson-Reuters' Web of Science (from 1900 to 2015) are completely missing any kind of author address metadata.

The first approach to deal with missing data in bibliographic records is to simply ignore incomplete records. This is a practical approach commonly used in IRC studies, in which researchers accept the exclusion of records without countries. Although this approach has certain advantages (i.e. simplicity and ease of use), it introduces a bias of representation because the missing data cannot be guaranteed as missing completely at random. Hence, both the missing collection and complete collection cannot be considered as representative of the whole population. Taking

the high omission rate of country identification mentioned above into consideration, this approach is undesirable.

The second approach to overcome data incompleteness is to replace missing data with substitute values using methods of data imputing. There is a wide range of imputing methods (e.g. hot-deck, cold-deck, mean and regression imputing, etc) common in social and physical sciences (Enders, 2010). All such approaches entail using existing values to infer reasonable (i.e. approximate) values, rather than identifying the correct values. However, the exact identification for countries in affiliation data of each publication is important to many measures in IRC studies. While only the total numbers of countries identified in the whole data set are required to examine the scientific sizes of countries, the relationships between countries in each publication are needed to build the IRC networks. Therefore, this approach has limited application in IRC field and is not considered in this research.

The third possible approach to tackle missing data is to leverage peer-produced open data sources such as Wikipedia or Wikidata. While there have been attempts to generate scholarly profiles from such open data sources when dedicated data is known to be present within those (Lemus-Rojas & Odell, 2018), no prior work has developed or described an algorithmic approach that aims at improving bibliographic records of author affiliation with a focus on authors' geographical location using matching heuristics, which is the focus of our work.

3 Proposed approach

3.1 Use of knowledge graphs

In previous IRC studies applying string matching, the task of country identification is done by getting pieces of data from affiliation information and matching against pre-stored lists of common phrases that act as reliable indicators for country disambiguation. The effectiveness of this commonly used method is limited by the fact that many pieces of affiliation data do not include “country” information, as described above. Our proposed approach is, therefore, to resolve and assign affiliation data to particular countries by drawing on Web-based knowledge graphs, which have multiple advantages for that use.

Various knowledge graphs have been developed to describe entities in the real world and the relationships between them. Among the knowledge graphs that are freely accessible, such as DBpedia, OpenCyc, Wikidata, and YAGO, Wikidata is overall ranked the best in data quality (Färber et al., 2018). Wikidata is a well-regarded peer-produced knowledge base that primarily solves issues of data redundancy in different language versions of Wikipedia but by now being utilized

for an increasing amount of use cases that benefit from structured background knowledge such as Google search (Pellissier Tanon et al., 2016). Wikidata is built around two main logical constructs: items and properties. While items depict entities in the real world, properties describe the items' relations (Müller-Birn et al., 2015). For example, the record "Victoria University of Wellington" in Wikidata is an instance of the item "public university" and stores a statement about the location property (i.e. the country is New Zealand). This logical paradigm allows to consider the Wikidata knowledge base as a graph (where items constitute the vertices and properties the edges). Hence it is commonly referred to as a knowledge graph and allows interesting technical use cases that make use of graph traversal through structured queries.

The success and recent widespread adoption of Wikidata have led to many studies about its data quality. There are four categories of quality dimensions addressed for this purpose: intrinsic, contextual, representational, and accessibility data quality. In details, intrinsic dimensions include accuracy, trustworthiness, consistency; contextual dimensions include relevancy, completeness, timeliness; representation dimensions include easy of understanding, interoperability; accessibility dimensions include accessibility, interlinking, and license (Piscopo & Simperl, 2019). Wikidata is highly evaluated in all these dimensions, reaching a 100% quality score with respect to the syntactic validity of RDF documents and syntactic validity of literals, and the 99% with respect to semantic validity of triples reaches (Färber et al., 2018).

It is for the aforementioned features of Wikidata that we decided to use Wikidata via its SPARQL query endpoint to complete missing country values of affiliation information in bibliographic data sets (Nguyen, Dinneen, & Luczak-Roesch, 2019).

3.2 Implementation

We implemented two data enrichment steps in our method to identify the collaborating countries using the authors' affiliation data in co-authored papers. Figure 1 gives a high-level overview of our approach, which we will describe in detail in the remainder of this section.

First, we implemented algorithmic string matching using standard R packages (e.g. lists of country names and abbreviations). For records with author affiliation data containing countries' names or abbreviations, these pieces of information were matched to a predefined list of countries. We also took into account some common facts in building the algorithm. First, many affiliation data have the postcodes or the state names or city names of the US states. Therefore, additional predefined lists were also used to identify these cases. Second, some records have the names of the

Figure 1. Abstract view of the different data retrieval and processing steps in our approach.

UK components (e.g. Scotland or England). In IRC studies, the UK was usually considered as a whole entity for all its component parts so we mapped all corresponding component names to the United Kingdom. Third, there are different alternative names for China (e.g. “PRC” or “P.R.C”). All these cases mentioned above were incorporated into our method’s codes for identifying the responding countries.

There are some limitations of the above string matching step though. For some affiliation data linked to countries that do not exist (e.g. Yugoslavia or Czechoslovakia), we found no ways to map these affiliation data to the corresponding new-formed countries. Therefore, the country identification of these special cases (and others have no information about countries) relies on the success of the second step.

Second, we carried out an additional algorithm using the Wikidata query when it is necessary: for records that could not be identified having no clear country names

or state information, we then used the remaining information (e.g. university name) to query the SPARQL endpoint of Wikidata^① executing the following query^②:

```
PREFIX schema: <http://schema.org/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
SELECT ?countryLabel WHERE {
    <https://en.wikipedia.org/wiki/[AFFILIATION]> schema:about ?datalink.
    ?datalink wdt:P17 ?country.SERVICE wikibase:label
    {bd:serviceParam wikibase:language "en"}.}
}
```

This query returns English names of countries associated with the location data if there is a matching Wikidata item. For example, querying “McGill University” returns Canada. We implemented these steps in *R*, and made the source code of both the string matching and the Wikidata query implementation freely available for use in future work^③.

In rare cases where seemingly complete affiliation data does not match with a Wikidata item, each part (i.e. word) of the affiliation data is then queried, assigned a Wikidata item when possible, and those assignments are then checked for consistency. If the assignments match (i.e. are all identified with the same country), the country is regarded as successfully identified. For example, our Wikidata query could not identify the country associated with the exact phrase “Massachusetts Institute of Technology, Cambridge”, so “Massachusetts Institute of Technology” and “Cambridge” were queried. Because “Massachusetts Institute of Technology” is identified as the United States and “Cambridge” as the United Kingdom (i.e. not the same country for the record in question), the record is treated as a non-match and not assigned country data.

4 Evaluation

We utilized two distinct data sets in the evaluation of our method, wherein we assessed its magnitude, accuracy, and consistency in identifying publications across different countries. Two distinct data sets were chosen to control for potentially skewed results and interpretations resulting from significant qualitative differences that have been shown to exist in bibliographic data sets utilized for the purpose of IRC measurement (Nguyen, Luczak-Roesch, & Dinneen, 2019).

^① <https://query.wikidata.org/sparql>

^② “[AFFILIATION]” in this query gets replaced with the remaining information extracted

^③ <https://github.com/baxuan/IRCM/tree/master/Comparing-data-sets>

4.1 Preparation of data sets

The most commonly used data sets in IRC research are SCI/Web of Science and Scopus (Guerrero Bote, Olmeda-Gómez, & de Moya-Anegón, 2013; Luukkonen et al., 1993). Each set entails considerable practical challenges for researchers; for example, only 500 records can be downloaded from WoS at a time, or 2,000 from Scopus. Further, these sets (and Google Scholar) do not have as comprehensive general coverage as, for example, Microsoft Academic Graph (Paszczka 2016; Sinha et al., 2015), and may not have as complete domain-specific coverage as, for example, ACM Digital Library (ACM DL) and IEEE Xplore provide for computer science.

Therefore, we tested our method on Microsoft Academic Graph (MAG), a general scholarly bibliographic data set, and the ACM Digital Library, a scholarly bibliographic data set containing works published by the Association for Computing Machinery and primarily related to computer science. These two data sets are commonly used to study computer science research (Guha et al., 2013; Herrmannova & Knoth, 2016). Importantly, because the fields in these databases that our method evaluates and enriches (i.e. those describing authors’ affiliations) also exist in other bibliographic data sets, the method can also be applied to other bibliographic data sets (e.g. WoS and Scopus).

To make the results of our evaluations comparable across the two sets we filtered out records from MAG that were not relevant to computer science: a list of fields of study was compiled from records present in both ACM DL and MAG, and the 38 top terms (94% of papers in the overlap) were used to filter out irrelevant works. Overlapping papers were also filtered from the ACM DL set to make the sets distinct. Finally, single-author records were filtered out to identify only co-authored papers. Table 1 summarises the data sets used.

Table 1. Summary of data sets used.

Features	ACM DL	MAG
Total works	182,791	212,689,976
Unique, co-authored, computer science works	121,672	557,730

After preparing the two data sets, our enrichment method (section 3.2) was applied to both.

4.2 Method’s outputs

Our approach identifies countries for approximately 71%–75% records in each data set (the percentages of different categories of the records processed are shown in Table 2), with the remaining records being either unidentified (~15–17%) or

unidentifiable (~8–14%) because of empty or placeholder affiliation values (e.g. “NA” or “None”). Specifically, while the substring matching approach identifies countries for nearly 60–70% of records, the Wikidata querying approach adds an additional roughly 10% in ACM DL and 6% in MAG. In other words, the method provided identifies approximately 82% of the possible records. These results suggest that our approach succeeds in matching the majority of bibliographic records, in both general and domain-specific data sets, to the relevant countries.

Table 2. Results of country identification.

Code	Results	ACM DL	MAG
C	Affiliations in co-authored works	384,672	853,482
C1	“NA”, “None”, etc values	52,454 (13.64%)	66,924 (7.84%)
C2	Identified	273,245 (71.02%)	643,678 (75.42%)
C2.1	Identified by string matching	236,100 (61.38%)	594,911 (69.70%)
C2.2	Identified by Wikidata	37,106 (9.65%)	48,767 (5.71%)
C3	Not identified (Other values)	59,012 (15.34%)	142,888 (16.74%)

4.3 Method's effectiveness assessment

We describe here the rationale and procedures for evaluating our method, using the three measurement criteria of utilization, quantitative accuracy, and fairness/justice (Jacobs, Kafry, & Zedeck, 1980), implemented as magnitude, accuracy, and consistency, respectively:

- To assess the magnitude of additional information, we examined how the new method improves upon the presumed baseline process for country identification.
- To assess the method's accuracy we investigated the extent to which the outputs of the proposed approach (on the C2.2 and C3 collections as described in Table 2) are correct and free of error. This is because the reliability of the new method could not be ensured if there are many cases identified are false.
- To assess the method's consistency we checked whether the improvement of the combined method is consistent for all countries, or if there are any countries that benefit more or less than others. The validity of the improvement will be threatened if the method's outputs are strongly disproportionate in favour of or against certain (groups of) countries.

4.3.1 Assessment of magnitude of additional information

In this assessment, we had a look at the patterns of the improvement overtime (comparison by years). The comparison was made for the period 1951–2016, in which both data sets tested have whole year data. The charts (a) and (b) in Figure 2 show the ratios by year of improvement made by using Wikidata query in ACM DL data set and MAG data set, correspondingly. In details, these charts compare (1) the numbers of affiliation data unidentified after using string matching (and excluding unidentifiable cases, e.g. “NA” or “None”), and (2) those received after the above step by applying Wikidata query.

The magnitude at which some method of identifying countries improves upon the baseline is an indication of the improvement in the resulting IRC measurement. As presented in the method’s outputs section above, the rates of improvement for the total affiliation data are 9.65% and 5.71% for ACM DL and MAG data sets. The rates of identification roughly follow a logarithmic trendline ($R^2=0.33$ for ACM DL and 0.63 for MAG), as contrasted with a linear improvement ($R^2=0.31$ and 0.48). In other words, the rates of improvement by year initially increase quickly, but then stabilize around 40% and 30%. It is thus predictable that rates of improvement will not change considerably as more data (i.e. after 2016) are available and considered.

4.3.2 Assessment of the accuracy of additional information

We checked the accuracy of applying Wikidata query in this study on 100 random pieces of affiliation data identified by using this query (C2.2 collection) and 100 random pieces of affiliation data unidentified in each data set (C3 collection), respectively. We used two metrics in this assessment. First, false match rate (FMR) is the rate at which the Wikidata query incorrectly matches affiliation information to relevant countries. Second, false non-match rate (FNMR) is the rate at which this method fails to link affiliation information with corresponding countries in cases that could otherwise be linked (i.e. valid inputs that a human could disambiguate but the method failed to). The results of the assessment are shown in Table 3.

Table 3. The accuracy of the method using Wikidata query.

	ACM DL	MAG
False match rate (FMR)	0 %	0 %
False non-match rate (FNMR)	73 %	75 %

In the above table, the values in the first row of Table 3 show that our algorithm using Wikidata query has 0% errors in identifying matched cases in the test sample. These values received from both ACM DL and MAG illustrate the accurateness of our method, which is in part determined by the accuracy of Wikidata query. This result is aligned with the well-known quality of Wikidata.

Figure 2. Comparing numbers of affiliation data unidentified using string matching with numbers of those identified using Wikidata query by year.

Research Paper

Meanwhile, the second row of Table 3 shows the proportions of presumed valid inputs rejected incorrectly although these inputs can be recognized by a human being using tools such as Google Search (e.g. “University of Mississippi, Oxford” or “Beijing Inst. of Technol”). These false non-match rates are 73% and 75% (from 15.34% and 16.74% of records not being matched at ACM DL and MAG dataset, respectively). In terms of the overall data set, these rates are 11.2% and 12.56% for the two data sets mentioned. These values indicate that the effectiveness of Wikidata query could improve over time (i.e. more matches) as the underlying data set continues to grow. It is important to note that there is a high proportion of FNMR cases are affiliation organizations in the US (65% at ACM DL dataset and 13% at MAG dataset, i.e. “University of Nebraska, Lincoln, NB” or “Virginia Tech, Blacksburg”). This suggests that the US is the country that will benefit the most (i.e. get the highest number of additional collaborations identified) if the effectiveness of Wikidata query will be improved in the future.

4.3.3 Assessment of the consistency of additional information

Firstly, we surveyed the difference in the percentages of improvement for different countries (comparison by countries). This was done by calculating the value of (number of countries identified by Wikidata query/number of countries identified by string matching). Figure 3 provides a visual representation of the comparison. The outliers represent countries or territories with relatively few publications (e.g. State of Palestine and Serbia in ACM DL; Cape Verde, Uganda, and Nepal in the MAG), such that just a few additional collaborations resulted in large percentage increases. One such extreme outlier (Cape Verde) was excluded in Figure 3(b) to facilitate visualization.

Figure 3. Average improvement (and outliers) of affiliations identified per country using the ACM Digital Library (left) and Microsoft Academic Graph (right) data sets.

As the boxplots featured in the above figure reveal, the distributions of data for percentages of improvement are skewed and so the appropriate measures of centrality

and dispersion are the median and interquartile range respectively. The summary statistics provided in Table 4 confirm that, for countries having scientific publications, the normal percentages of improvement applying new method of country identification are nearly the same for both data set (the median values are 1.20% and 0.91% for ACM DL and MAG, respectively). The interquartile ranges are 10.18% and 6.30%, respectively, meaning that the improvement does not vary much among the countries, either using ACM DL data set or MAG data set. In other words, our method provides a fairly stable benefit among countries with only a few exceptions (e.g. places producing relatively few research outputs).

Table 4. Summary statistics of the method's results.

	ACM DL	MAG
Mean	5.70%	5.42%
Standard error	0.69%	1.29%
Median	1.20%	0.91%
Mode	0%	0%
Standard deviation	8.10%	17.88%
Interquartile range	6.30%	10.18%
Count	137	192

Secondly, we looked at how the improvement makes a difference to countries. Figure 4 shows the comparison in terms of the percentages of improvement received by our method and the numbers of affiliation records associated with each country (without Wikidata use) in the ACM DL and MAG data sets. As shown in Figure 3, an outlier in the improvement dimension (Cape Verde) was excluded in Figure 4(b) for easier comparison. Overall, we find no significant relationships between the percentages of improvement and the countries' numbers of affiliation data identified.

To test this more formally, we ran a linear regression analysis between the variables "countries' size in amount of related records found in the dataset" and "percentual improvement of country identification through the use of Wikidata", assuming there is a chance that affiliation information of organisations in countries of bigger scientific size may be more well-known than those in smaller ones. The analyses were performed once for all countries and once only for the top countries in terms of the amount of related records found in the ACM DL and MAG datasets. The reason for taking separately top countries into consideration is that IRC comparisons in absolute numbers are significantly affected by both country size (in terms of scientific papers) and collaborative activities (Luukkonen et al., 1993). Therefore, we wanted to compare the relationships for the whole list to those for the top countries only. The R squares observed (0.055, 0.22, 0.001, 0.15) indicate no significant relationship exists, neither in the list of all countries nor the list of top 30 countries in both data sets. In other words, only a small percentage of the

Figure 4. The number of affiliations identified for each country before and after using the Wikidata query, using ACM Digital Library and Microsoft Academic Graph data sets.

variability in the improvement can be explained by the relationship between the percentage of improvement and the country size (in terms of scientific papers), in both ACM DL and MAG data sets.

5 Discussion

We sought to address difficulties in identifying countries during IRC measurement, and to evaluate our proposed method’s magnitude, accuracy, and consistency. Barring existing comparable methods to compare ours to, we use the above evaluation to assess our contributed method. Evaluation of our method’s magnitude (i.e. extent of its improvement over a baseline in identifying countries) shows that

using knowledge graphs to complete and resolve affiliation data improves the rate at which countries are identified, improving upon text matching-only approaches by 5.71% and 9.65% in MAG and ACM DL, respectively. Overall our method of combining string matching with the Wikidata query resulted in correctly identifying 75% and 71% of affiliations' countries in the data sets used (MAG and ACM DL, respectively). Regarding our method's accuracy, we observed that, crucially, it made no false matches. While 75% of the randomly sampled non-matches from the Wikidata queries could indeed have been matched to particular countries, the vast majority of those were cases caused by incomplete or dirty data (e.g. "Lincoln, Nebraska, NB"), suggesting that improvements in data quality and/or Wikidata's query resolution algorithm could drastically reduce non-matches and thus further improve the method's performance (e.g. matching up to 88% of affiliations in MAG). This should be the subject of future work. Finally, the results of evaluating our method's consistency suggest it is consistent over time and its bias to particular countries is minimal.

In summary, the results of evaluating our method show that the approach of applying Wikidata querying to country identification is effective, accurate, and consistent (i.e. relatively free of bias). We next discuss the benefits of this method for bibliometric research and in practice.

5.1 Implications

The method proposed here promises to be useful in both bibliometric research and in practical applications of IRC measurement. To start, use of our method in future studies should lead to increased accuracy in measuring IRC. For example, a map depicting global IRC networks (e.g. Wagner & Leydesdorff, 2005) will be more accurate if generated using our method, as would characterizations of the patterns of research collaboration (e.g. Schubert & Braun, 1990) across groups of countries (e.g. Western Europe as compared with Eastern Europe). Given that our method had similar rates of affiliation identification (over 70%) for both data sets examined, regardless of the initial efficacy of string matching (e.g. low efficacy with ACM DL data), and because it requires only the affiliation data common in most bibliographic data sets, it is likely that uses of our method (like the examples here) will be effective regardless of the data set(s) used (e.g. Scopus or WoS); this is important as many studies are practically limited to one data set.

As our method works well and is shared online, it can also be collaboratively improved and easily adapted for similar uses in research beyond measuring IRC. For example, the improved country identification can be useful for other bibliometric analyses, such as measuring a country's total scientific production or the rates of its scientist emigration and immigration. Further, because our method uses Web-based

Research Paper

knowledge graphs, other missing bibliographic data can be completed using our method with minor changes, for example journals' languages or publishers' names. While the implementation of our method is publicly shared online, its performance may be further improved by the contributions of other researchers in the future. Given the logarithmic rate of improvement across years observed above, any improvement to the method is likely to improve IRC identification across all publication years.

Our method also may be useful for performing analyses to inform policy decisions. For example, the policy makers for joint research programs (e.g. European Union funding programs) may need to evaluate the technological, societal, and economic impacts of multinational research projects. The more accurately such impacts on each related country can be measured, the more confidently policy makers can be in making decisions for further investment (e.g. extending multi-national research projects). This extends further to strategic decisions about which countries to encourage collaboration with; for example, countries having advanced technical skills (e.g. Korea) may target funding collaborations with countries with complementary knowledge (e.g. theoretical expertise in the U.K.) and thus enable mutually beneficial IRC (e.g. the transmission of know-how in exchange for funding and research resources; Hwang, 2008). Identifying strategic collaborations and monitoring the efficacy of bi-lateral cooperation, however, will benefit from a more accurate and consistent method of measuring research outputs, such as the one provided here.

5.2 Limitations

Our method has been designed and tested to perform well with usual, simple cases of IRC measurement (e.g. countries with singular identities and relatively frequent research outputs), and so it may not apply straightforwardly when measuring the IRC of countries that have, for example, changed identity (e.g. name or composition) or no longer exist as such (e.g. Yugoslavia). Bespoke approaches may be needed for such edge cases.

Although we tested our method with both general and domain-specific bibliographic databases, it has yet to be tested with other popular data sets (e.g. Scopus), and so may perform differently with them. Given the similar performance of our method on ACM DL and MAG, it is reasonable to expect similar performance with other bibliographic data sets, but this should be empirically validated.

Finally, although the evaluation of our method's reliability had promising results (i.e. no false matches), that evaluation was practically limited to manually inspecting the results of only 400 random Wikidata query results (i.e. <1% of the records in our test implementation). Though we have no particular reason to suspect further

random samples would indicate considerably different reliability, it is nonetheless true that manually inspecting a larger sample of results may have different implications about the reliability of our method; such inspection should be done in testing future revisions of the method.

6 Conclusion and future work

A current problem in IRC research is that it is difficult to identify countries by the affiliation information that bibliographic records provide. Previously, no method was available to overcome this, so methods used were *ad hoc*, impractical, and likely inconsistent with each other, thus potentially resulting in varying results across even studies using the same data sets, or worse, preventing IRC measurement altogether. Here we provided and evaluated a novel method for addressing the problem, using string matching and the SPARQL endpoint of the Wikidata knowledge graph. The method appears promising for use with other data sets as well, especially given that Wikidata will continue to grow and thus improve in matching affiliations to countries.

Several further works are warranted by our results. Firstly, to improve the string matching and knowledge graph querying (i.e. to treat false non-matches), natural language processing may be useful for country disambiguation. It is a fact that the current algorithms could not identify many pieces of dirty or incomplete affiliation data where the affiliation was otherwise clear (e.g. to a human), and so complementary approaches to traditional algorithmic processing (e.g. machine learning) are promising for addressing this problem. Secondly, tests of our method with data sets not examined here should be performed to further validate its benefit across bibliometric data more generally, taking that fact that the rate of absence of country identification is high in commonly used bibliographic datasets (e.g. Scopus and Thomson-Reuters/Web of Science) as mentioned in this research. Thirdly, exploring the underlying reasons for the differences of improvement between countries when applying our method should be undertaken. As presented in Figure 4, different countries have different rates of improvement, which cannot be explained by the country “size” (in terms of scientific papers). There are various reasons for these differences in matching using Wikidata query: some countries may have a higher rate of accurate affiliations stored in bibliographic data, or Wikidata may have a higher rate of predefined affiliations for certain countries. It is worthwhile then to investigate the reasons for this differentiation by country. Finally, the value of the method should be demonstrated in application, for example, by examining the difference it makes in studying the current landscape of global IRC networks and IRC patterns.

Author contributions

Background research: Ba Xuan Nguyen (ba.nguyen@vuw.ac.nz). Conceived of and designed the analysis: Ba Xuan Nguyen, Markus Luczak-Roesch (markus.luczak-roesch@vuw.ac.nz), Jesse David Dinneen (jesse.dinneen@vuw.ac.nz). Collected data: Ba Xuan Nguyen, Markus Luczak-Roesch. Contributed data or analysis tools: Ba Xuan Nguyen, Markus Luczak-Roesch. Performed analyses: Ba Xuan Nguyen, Markus Luczak-Roesch, Jesse David Dinneen. Wrote and revised the manuscript: Ba Xuan Nguyen, Jesse David Dinneen, Markus Luczak-Roesch.

References

- Beaver, D., & Rosen, R. (1978). Studies in scientific collaboration: Part I. The professional origins of scientific co-authorship. *Scientometrics*, 1(1), 65–84.
- Beaver, D., & Rosen, R. (1979). Studies in scientific collaboration: Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799–1830. *Scientometrics*, 1(2), 133–149.
- Chen, K., Zhang, Y., & Fu, X. (2018). International research collaboration: An emerging domain of innovation studies? *Research Policy*. doi:10.1016/j.respol.2018.08.005
- Enders, C.K. (2010). *Applied missing data analysis*. Guilford press.
- Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2018). Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, 9(1), 77–129.
- Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., . . . , & Vespignani, A. (2018). Science of science. *Science*, 359(6379), eaao0185.
- Glänzel, W., & Schubert, A. (2001). Double effort = double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, 50(2), 199–214.
- Guerrero Bote, V.P., Olmeda-Gómez, C., & de Moya-Anegón, F. (2013). Quantifying the benefits of international scientific collaboration. *Journal of the American Society for Information Science and Technology*, 64(2), 392–404.
- Guha, S., Steinhardt, S., Ahmed, S.I., & Lagoze, C. (2013). Following bibliometric footprints: The ACM digital library and the evolution of computer science. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* (pp. 139–142).
- Herrmannova, D., & Knoth, P. (2016). An analysis of the microsoft academic graph. *D-lib Magazine*, 22(9/10).
- Hinze, S. (1994). Analysis of country specialisation in bioelectronics with special focus on German activities. *Research Evaluation*, 4(2), 107–118.
- Hwang, K. (2008). International collaboration in multilayered center-periphery in the globalization of science and technology. *Science, Technology, & Human Values*, 33(1), 101–133.
- Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel psychology*, 33(3), 595–640.
- Jacsó, P. (2009). Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries. *Online information review*, 33(2), 376–385.
- Lemus-Rojas, M., & Odell, J.D. (2018). Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project using Wikidata and Scholia. *Journal of Librarianship and Scholarly Communication*, 6(1), p.eP2272. doi: 10.7710/2162-3309.2272

- Liu, W., Hu, G., & Tang, L. (2018). Missing author address information in Web of Science—An explorative study. *Journal of Informetrics*, 12(3), 985–997.
- Luukkonen, T., Tijssen, R., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics*, 28(1), 15–36.
- Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E.D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160–1177.
- Müller-Birn, C., Karran, B., Lehmann, J., & Luczak-Rösch, M. (2015). Peer-production system or collaborative ontology engineering effort: What is Wikidata? In *Proceedings of the 11th International Symposium on Open Collaboration* (p. 20). ACM.
- Nguyen, B.X., Luczak-Roesch, M., & Dinneen, J.D. (2019). Exploring the effects of data set choice on measuring international research collaboration: An example using the ACM digital library and microsoft academic graph. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics (ISSI 2019)*.
- Nguyen, B.X., Dinneen, J.D., & Luczak-Roesch, M. (2019). Enriching bibliographic data by combining string matching and the Wikidata knowledge graph to improve the measurement of international research collaboration. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics (ISSI 2019)*.
- Paszczka, B. (2016). Comparison of Microsoft academic (graph) with Web of Science, Scopus and Google Scholar (Doctoral dissertation, University of Southampton).
- Pellissier Tanon, T., Vrandečić, D., Schaffert, S., Steiner, T., & Pintscher, L. (2016). From freebase to wikidata: The great migration. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 1419–1428). International World Wide Web Conferences Steering Committee.
- Piscopo, A., & Simperl, E. (2019). What we talk about when we talk about wikidata quality: A literature survey. In *Proceedings of the 15th International Symposium on Open Collaboration* (p. 17). ACM.
- Schubert, A., & Braun, T. (1990). International collaboration in the sciences 1981–1985. *Scientometrics*, 19(1–2), 3–10.
- Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J.P., & Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 243–246). ACM.
- Wagner, C.S., & Leydesdorff, L. (2005). Mapping the network of global science: Comparing international co-authorships from 1990 to 2000. *International Journal of Technology and Globalisation*, 1(2), 185–208.

This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).