Noname manuscript No.
(will be inserted by the editor)

Accuracy of Author Names in Bibliographic Data Sources: An Italian Case Study

Camil Demetrescu · Andrea Ribichini · Marco Schaerf

Abstract We investigate the accuracy of how author names are reported in bibliographic records excerpted from four prominent sources: *WoS*, *Scopus*, *PubMed*, and *CrossRef*. We take as a case study 44,549 publications stored in the internal database of Sapienza University of Rome, one of the largest universities in Europe. While our results indicate generally good accuracy for all bibliographic data sources considered, we highlight a number of issues that undermine the accuracy for certain classes of author names, including compound names and names with diacritics, which are common features to Italian and other Western languages.

Keywords Author names, accuracy, Scopus, WoS, CrossRef, PubMed.

1 Introduction

Research evaluation has become increasingly important for universities, and other state-recognized research institutions, in the past few decades as governments in a growing number of countries have launched periodic national research assessment campaigns. The objective of these campaigns include boosting research productivity, allocating public fund based on merit, and assessing the national-level research infrastructure and the impact of policies on promoting research [Franceschini and Maisano, 2017]. Research assessment campaigns are conducted with a variety of methodologies [Abramo et al., 2011], which may also change over time, based on accumulated experience, available resources, theoretical innovations, and policy objectives. Contemporary evaluation efforts are often based on hybrid approaches, combining peer review with analyses based on bibliometric indicators (e.g., citation counts), especially for hard sciences, for which coverage by international bibliographic databases tends to be more comprehensive.

The popularity of bibliometric indicators stems from their reduced cost and assessment efficiency compared to peer review. Moreover, it is relatively objective and reproducible. However, it raises the question as to whether bibliographic data sources are accurate enough to support fair evaluation processes. Bibliographic records are complex objects, containing a wealth of information (article's title, author names, journal title, volume number, issue number, starting and ending pages, publication year, just to name a few). As pointed out by Olensky, errors in any, or a combination, of these fields may cause citations to be incorrectly matched, or missed altogether [Olensky, 2014]. This can impact on the results of the

C. Demetrescu
Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy.
E-mail: demetres@diag.uniroma1.it

A. Ribichini
Department of Physics, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy.
E-mail: ribichini@diag.uniroma1.it

M. Schaerf
Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy
and
Institute of Information Technologies and Telecommunications
North Caucasus Federal University, Stavropol, Russian Federation
E-mail: marco.schaerf@uniroma1.it

evaluations they inform, which may not only be national research assessment exercises, but also global university rankings such as the Shanghai Ranking (`http://www.shanghairanking.com/`) or the Leiden Ranking (`http://www.leidenranking.com/`), or even evaluations of individuals (e.g., for recruitment decisions, career advancements, allocation of research funds, and identification of experts in specific fields) [Olensky, 2015].

Contributions of the Article. In this article we study the accuracy, with regards to author names, of four prominent sources of bibliographic data (namely, *WoS*, *Scopus*, *PubMed*, and *CrossRef*). In this study we take 44,549 publications stored in the internal database of Sapienza University of Rome (one of the largest universities in Europe), with publication years ranging from 1960 to 2015. Our case study therefore contains mostly, but not exclusively, author names of an Italian origin. How these names are handled by the data sources we have considered has become a matter of rising interest in recent years, due to the mixed approach (peer review and bibliometric analysis based on data from international databases) employed by the two most recent national research assessment exercises conducted by the Italian Ministry of Education, University and Research, covering respectively the years 2004-2010 and 2011-2014. Our investigation reports generally accurate results for all the considered bibliographic sources, even though some specific areas of concern emerge, particularly in the handling of compound names and names with diacritics.

2 Overview

Errors in author names can occur due to several reasons, which can be generally traced back to the unstructured, ambiguous, or imprecise nature of the original data from which bibliographic entries are derived [Meho and Yang, 2007]. Prominent examples include: lack of a clear separation between first name and last name in the article, missing or incomplete information such as abbreviated first names, character encoding errors arising in the data processing pipeline, typos in the metadata inserted by the authors at submission time, and errors introduced by optical character recognition (OCR).

There are two methodological aspects that need to be considered in assessing name accuracy. As a first aspect, it requires a "ground truth" that provides a canonical, unambiguous form to precisely tell whether a name is correctly reported in an author list. Sometimes there may even be multiple canonical forms for the same author as in the case of pseudonyms or different names used throughout their career [Bennett and Williams, 2006].

As a second aspect, the notion of author name correctness may have some grey areas that need to be taken into account: while certain differences between how an author name is reported and its canonical version should always be considered as mistakes, such as typos, others should be tolerated. For instance, cases where: (i) first names are reported with initials rather than in full, (ii) multiple initials are concatenated with no punctuation, or (iii) different, but compatible diacritics are used, could all be considered correct mutations. Ideally, we would like to have a clearly defined notion of compatibility between names that allows us to draw a line between legit variants of the same canonical name (or pseudonym) and genuinely incorrect variants.

As a concrete example, author name variants "De Martino, P. L.", "De Martino, P.", "De Martino, Pier L.", "De Martino, Pierluigi", should all be considered compatible with the canonical version "De Martino, Pier Luigi" (or its pseudonym "De Martino, Pierluigi"), where "De Martino" is an author's last name and "Pier Luigi" is his given name. Conversely, "Martino, P.", "Martino, P. D.", and "Di Martino, P. L." should all be regarded as incorrect, each for a different specific reason.

In the remainder of this article, we discuss related work (Section 3) and we elaborate on the two methodological aspects addressed above, discussing: (i) the data set and ground truth for author names we considered (Section 4) and (ii) the compatibility notion we used to compare author names and our error classification (Section 5). We then present the results of our investigation, showing how different bibliometric sources are affected by the different error categories we considered (Section 6). We conclude the article with some final remarks (Section 8).

3 Related Work

Issues with author names are often identified in the literature as a cause for missed citations in bibliographic databases. Harzing and Wal point out that names containing diacritics, apostrophes or ligatures are problematic for both *WoS* and *Google Scholar* [Harzing and Wal, 2008]. Meho and Yang report that critics of *WoS* note inconsistencies in the use of initials and in the spelling of non–English names, though these seem to come from the original documents, rather than being the results of faulty data acquisition processes on ISI's part [Meho and Yang, 2007]. Tunger *et al.* claim the most frequent errors in the bibliographic data they analyzed from *Scopus* to be related to Chinese given names, which are often abbreviated with two initials rather than one (e.g., "Hongbao" abbreviated "HB" rather than "H") [Tunger et al., 2010]. They further report that other common errors are authors' names and affiliations missing or misspelled. The nature of the misspellings (e.g., "rn" instead of "m"), they claim, leads to the conclusion that these are due to OCR errors.

As pointed out in [Garfield, 1981], "People are generally sensitive both about the way their names are pronounced and how they are presented in print. Some authors may consider it a dishonor to their heritage when we abbreviate their names for our convenience." Garfield further remarks that compound last names indicate family roots, and they are usualy prefixed by foreign articles and/or prepositions, and that hyphenated names usually signify the combination of two distinguished family lines. It is also observed that some authors may publish under various versions of their own last names, and they may change the spelling of their last names if, for example, they include accents or other diacritical markers which may not be printed correctly.

Ruiz-Pérez and his colleagues [Ruiz-Pérez et al., 2002] discuss the mishandling of Spanish names in English-language databases, where compound last names are common practice, showing that about half of all Spanish authors seem to have lost their second last name in bibliographic records. Interestingly, they speculate that many Spanish researchers might have taken active precautions to simplify their names as they appear in their publications to make indexing in international databases more reliable, by deleting the second last name or by hyphenating compound last names, which are likely to be mishandled in the process.

Hood and Wilson remark that the use of databases in informetric studies presents both opportunities and challenges [Hood and Wilson, 2003]. Clearly, databases may act as data sources for informetric studies, and their delivery mechanisms may provide sophisticated analytical tools. However, electronic data may also contain errors or lack consistency. In particular, as far as author names are concerned, they report that abbreviated forms of author names may result in ambiguities, that authors may, from publication to publication, change the preferred form for their names, and that different journals may have different policies regarding the representation of author names. Regarding errors, they refer to Pao's categorization [Pao, 1989] into the following nine headings: additions, omissions, transpositions, misspellings, spacings, punctuations, capitalizations, compound names, and combinations of the above.

Bennett and Williams point out that issues related to author names may be grouped into three categories: (1) name variations that signify the same author, (2) similar or homonymic names that belong to different authors, and (3) deliberate changes, generally due to changes in marital status or other religious or legal reasons [Bennett and Williams, 2006]. They emphasize that manually maintained authority files, whose objective is to determine when name variations belong to the same individual, may yield highly accurate results, but do not scale well when transitioning from library catalogues to larger databases. It is argued that, in these cases, automated disambiguation methods based on additional metadata beyond simple author names, and yielding probabilistic results, may be more suitable.

Aksnes studies the frequency of homonyms among the population of Norwegian researchers (slightly more than 30,000 individuals) [Aksnes, 2008]. Separate statistics are kept for full names and for an abbreviated form consisting of last name followed by a sequence of initials (akin to the way *WoS* indexes author names). Results indicate that 14% of authors share their name with one or more other researchers, when abbreviated names are considered, while this percentage decreases to 1.4% when full names are used. Moreover, it is noted that the distribution of last names is quite skewed, with 13% of last names (the most common) accounting for 50% of all authors.

Olensky compares the results of an automated accuracy assessment method on bibliographic data with those obtained by manual verification [Olensky, 2014]. The automated method is based on comparing, through Levenshtein distance, a number of text fields (e.g., article title, journal title, author names) as they are found in bibliometric records retrieved from *Scopus* and *WoS*, with the correct data obtained from the original publications, plus some numeric fields such as publication year, volume number, start and end pages. Olensky concludes that Levenshtein distance is a good means to determine whether two bibliographic records exhibit discrepancies. At the same time, however, obtained scores do not seem to provide a good measure of how accurate individual fields are (i.e., whether detected inaccuracies are major, medium or minor ones).

Source coverage				
WoS	*Scopus*	*PubMed*	*CrossRef*	CRIS publication records
✓	✓	✓	✓	9666 (21.70%)
✓	✓	✓	–	1597 (3.58%)
✓	✓	–	✓	6945 (15.59%)
✓	✓	–	–	1189 (2.67%)
✓	–	✓	✓	354 (0.79%)
✓	–	✓	–	322 (0.72%)
✓	–	–	✓	824 (1.85%)
✓	–	–	–	1707 (3.83%)
–	✓	✓	✓	2995 (6.72%)
–	✓	✓	–	1519 (3.41%)
–	✓	–	✓	3512 (7.88%)
–	✓	–	–	1628 (3.65%)
–	–	✓	✓	490 (1.10%)
–	–	✓	–	2153 (4.83%)
–	–	–	✓	2119 (4.76%)
–	–	–	–	7529 (16.90%)
22604 (50.73%)	29051 (65.20%)	19096 (42.85%)	26905 (60.39%)	44549 (100%)

Table 1 Coverage of publications in our CRIS by different bibliographic sources.

	WoS	*Scopus*	*PubMed*	*CrossRef*
Total author lists	22630	29322	19178	27425
Malformed author lists	26 (0.11%)	271 (0.92%)	82 (0.43%)	520 (1.90%)

Table 2 Malformed author lists in different bibliographic sources.

4 Data Set

Sapienza University of Rome keeps an extensive database of bibliographic records relative to publications coauthored by its faculty members, Ph.D. students and postdocs, in a dedicated Current Research Information System (CRIS). Thanks to automated tools and manual intervention by dozens of librarians, bibliographic records are kept linked to the corresponding records extracted from the *WoS*, *Scopus*, *PubMed*, and *CrossRef* sources considered in this study. The CRIS snapshot we considered includes 44,549 publications, from years ranging from 1960 to 2015, 83% of which are linked to at least one external bibliographic source and 21.7% to all four. Table 1 reports detailed coverage figures.

The CRIS is also linked to a master authority list of the canonical names of all authors affiliated with Sapienza University of Rome, which we used as ground truth. The list clearly separates first and last names, which are reported in their entirety without initials or abbreviations.

We found that author names in the bibliographic records extracted from *WoS*, *PubMed*, and *CrossRef* are generally listed in the *comma-semicolon* format (e.g., "Smith, J.; Brown, D."). Conversely, *Scopus* author names are listed in the *comma* format (e.g., "Smith J., Brown D."). As shown in Table 2, almost all author lists in the considered bibliographic sources obey the formats described above. Malformed author lists (i.e., author lists that do not conform to the above standards) are skipped for the purpose of our analysis.

We parsed the author lists appearing in the considered sources, obtaining a *signature* for each extracted individual author name. A signature is formed by a pair of strings that separately represent the last name and the first name as they appear in the bibliographic source, respectively. Last and first name strings may be internally formed by multiple tokens. The signature also keeps track of the author ID in the authority list (for Sapienza authors). A signature example may be: ("De Rossi", "M. G.", 4112). Notice that the same signature may appear in different publications by the same author.

In our study, we focused on all signatures corresponding to authors affiliated with Sapienza University of Rome. This led to a list of 18,121 distinct signatures for 8,951 authors, which we compared to the canonical name versions obtained from Sapienza's authority list as we describe in Section 5. All authors in our repository but 306 (3.41%) have last names of Italian origin. Non-Italian last names account for 655 distinct signatures (3.61% of all signatures). In Section 6.2, we separately analyze the accuracy of bibliographic databases on the smaller subset of non-Italian last names.

5 Methodology

This section provides details about our assessment methodology. We first address the author name compatibility notion we used in our study and then discuss the different error categories we considered.

Author Name Compatibility. To compare author names, we first *normalize* them to remove certain irrelevant details that pertain to character encoding, for instance.

In particular, we apply the following transformations to first and last name strings:

- string is forced to upper case;
- dashes are replaced with standard white spaces;
- all unicode spaces are replaced with a standard white space character;
- all sequences of separators (e.g., tabs, spaces, line breakers) are replaced with a single standard white space;
- string is trimmed (i.e., leading and trailing spaces are removed);
- unicode quotes and apostrophes are replaced with standard versions.
- trailing accents are replaced with apostrophes (e.g., “Laganà” becomes “Lagana’”).

The last operation is performed because, as a common practice mainly due to the widespread usage of keyboards without accented letters, a trailing accented letter may be typed as the combination of an unaccented letter followed by an apostrophe (e.g., “ -à ” may be typed as “ -a’ ”).

Two names are *compatible* if and only if, after normalization, last names are identical and the corresponding tokens in given names are identical or one is a prefix of the other. For instance, (“LAGANA’”, “M GIULIA”) matches with (“LAGANA’”, “MARIA GIULIA”), but not with (“LAGANA”, “M GIULIA”).

For each signature in our data set, we checked whether the signature’s name is compatible with the corresponding canonical Sapienza name. This partitioned our list of signatures into a list of “compatible” and a list of “incompatible” signatures. Since Sapienza’s master list contains a unique canonical name version for each author, the handling of pseudonyms required a separate treatment that involved a substantial amount of manual analysis to rule out false positives in the incompatible signatures list.

Error Classification. The last step of our investigation consisted in a classification of the different types of errors arising in the incompatible list. Our error classification, presented in Table 3, was designed to account for the most common classes of errors that we observed in our data set. The eight classes cover typos in the names, errors in diacritics and apostrophes, and incorrect handling of compound names. To rule out cases where a name is too different from its expected canonical version to be considered a typo, i.e., due to an occasionally incorrect record linkage in our data set, we classified as typos only differences within relatively small values of the Levenshtein distance [Levenshtein, 1966]. Classes are not mutually exclusive, i.e., an incompatible signature may contain more than one error type.

6 Results

In this section, we report our findings obtained by applying the error classification of Section 5 to incompatible signatures, considering first the general case (Section 6.1), and then the specific case of authors of non-Italian origin (Section 6.2).

6.1 Overall Accuracy

We first observe that author names included in the considered bibliographic databases are generally accurate: as reported in Table 4, percentages of signatures compatible with the official signatures stored in the Sapienza database range from 96.99% to 99.25%, depending on the database. Note that each of the 18,121 distinct signatures in our data set is counted for each bibliographic record in which it appears. For instance, the total number of signatures in *WoS* is 57,204, so each distinct signature appears on average 3.15 times. Detailed figures on the errors we found in our data set appear in Table 5 and are discussed below.

Bad Split over Last Name. As shown in Table 4, about 8% of all signatures have compound last names. A non–negligible fraction of these signatures suffer from bad splits (i.e., part of the compound last name is mistakenly assigned to the first name). More than 10% of signatures with compound last name are affected by this issue, in the case of *WoS*, *Scopus*, and *CrossRef*. *PubMed*, with a percentage of 5.24%, seems more accurate in this respect.

Designation	Description	Example
Bad Split over Last Name	Part of a compound last name is mistakenly attributed to an author's first name.	"De Rossi, Giuseppe" becomes "Rossi, G.D."
Incomplete Last Name	One or more tokens are dropped from a compound last name.	"La Torre, V." becomes "Torre, V."
Last Name with Omitted Diacritics	Diacritics are omitted from a last name.	"Trifirò, S." becomes "Trifiro, S."
Last Name with Incorrecly Imported Diacritics	Diacritics in a last name are not correctly imported.	"Spanò, A." becomes "Span□, A."
Last Name with Omitted Apostrophes	Apostrophes are omitted from a last name.	"D'Innocenzo F." becomes "Dinnocenzo F."
Last Name with Typos	We assume a last name to contain typos if it is not correct, but its Levenshtein distance from the correct version is comparatively small.	"Accornero, F." becomes "Accomero, F."
Bad Split over First Name	Part of a compound first name is mistakenly attributed to an author's last name.	"Verdi, Carlo Maria" becomes "Maria Verdi, C."
First Name with Typos	We assume a first name to contain typos if it is not correct, but its Levenshtein distance from the correct version is comparatively small.	"Bianchi, Erica" becomes "Bianchi, Enrica"

Table 3 Error classification. Examples show how author names are misrepresented in bibliographic databases.

	WoS	*Scopus*	*PubMed*	*CrossRef*
Total signatures	57204	72169	49982	61176
Compatible signatures (% of total signatures)	55484 (96.99%)	70498 (97.68%)	49605 (99.25%)	59812 (97.77%)
Signatures with compound last name (% of total signatures)	4393 (7.67%)	5857 (8.11%)	3929 (7.86%)	5012 (8.19%)
Signatures with compound first name (% of total signatures)	5143 (8.99%)	6264 (8.67%)	4388 (8.77%)	5387 (8.80%)
Signatures with diacritics in last name (% of total signatures)	649 (1.13%)	778 (1.07%)	667 (1.33%)	673 (1.10%)
Signatures with apostrophes in last name (% of total signatures)	803 (1.40%)	1040 (1.44%)	706 (1.41%)	844 (1.37%)

Table 4 Statistics on the author signatures in our data set.

Incomplete Last Name. Among compound last names, incomplete last names (i.e., last names in which one or more tokens have been omitted) are rare in both *PubMed* (0.10%) and *Scopus* (0.60%), while they are more frequent in *WoS* (1.25%), and in *CrossRef* (5.85%).

Diacritics–Related Issues in Last Name. Only slightly more than 1% of all signatures contains diacritical marks in last names. However, a vast majority of these are dropped by both *WoS* (97.07%) and *Scopus* (94.99%). *PubMed* and *CrossRef* omit a far smaller, but still significant percentage of diacritical marks (8.70% and 19.32%, respectively). *CrossRef* is the only source that exhibits encoding–related problems while importing names with diacritics, at least for the data set we considered. The extra 4.01% brings *CrossRef*'s total for diacritics–related issues to 23.33%, still considerably lower than the percentages of either *WoS* or *Scopus*.

Last Name with Omitted Apostrophes. Less than 1.5% of all signatures contain apostrophes. These are rarely omitted in the case of *Scopus* (0.67%) and *PubMed* (0.57%). The percentage is slightly higher for *CrossRef* (1.90%), and it increases significantly for*WoS* (9.59%).

Last Name with Typos. Typos in last names are well below 1% of total signatures, for all sources.

	WoS	*Scopus*	*PubMed*	*CrossRef*
Bad splits over last names (% relative to signatures with compound last names)	554 (12.61%)	597 (10.19%)	206 (5.24%)	610 (12.17%)
Incomplete last names (% relative to signatures with compound last names)	55 (1.25%)	35 (0.60%)	4 (0.10%)	293 (5.85%)
Last names with omitted diacritics (% relative to signatures with diacritics in last name)	630 (97.07%)	739 (94.99%)	58 (8.70%)	130 (19.32%)
Last names with incorrectly imported diacritics (% relative to signatures with diacritics in last name)	none	none	none	27 (4.01%)
Last names with omitted apostrophes (% relative to signatures with apostrophes in last name)	77 (9.59%)	7 (0.67%)	4 (0.57%)	16 (1.90%)
Last names with typos (% relative to total signatures)	395 (0.69%)	224 (0.31%)	88 (0.18%)	170 (0.27%)
Bad splits over first names (% relative to signatures with compound first names)	none	60 (0.96%)	16 (0.36%)	75 (1.39%)
First names with typos (% relative to total signatures)	5 (0.01%)	3 (<0.01%)	none	35 (0.06%)
Incompatible signatures with unclassified errors (% relative to total incompatible signatures)	4 (0.23%)	6 (0.35%)	1 (0.26%)	10 (0.73%)

Table 5 Statistics on how author names in our data set are misrepresented in different bibliographic databases.

Bad Split over First Name. Compound first names are below 9%, for all sources. Among these, bad splits of compound first names, in which part of a first name is mistakenly assigned to the last name, are rare occurrences, ranging from "none" in the case of *WoS* to 1.39% in the case of *CrossRef.*

First Name with Typos. First names with typos are extremely rare, the highest percentage over total signatures being *CrossRef*'s 0.06%. The reason is that most first names are abbreviated with initials.

Unclassified Errors. A handful of signatures, while being incompatible with the official Sapienza signatures they have been matched with, do not fit into any of the error categories described above. These signatures are so dissimilar from the ones stored in the Sapienza database to escape all the error detection rules we defined, while still being recognized by human reviewers (with various degrees of confidence) as referring to the same author. For example, "Pozza, L." may be loosely matched with Sapienza signature "Da Pozzo, Luisa", which will fit neither the typo class (the two last names are too different) nor the incomplete last name class (one last name cannot be obtained from the other by simply dropping a token). In other cases, unclassified errors are due to incorrect records in the Sapienza CRIS.

6.2 Accuracy for Non-Italian Names

In this section, we address the accuracy of bibliometric databases on the subset of non-Italian names in our repository. As shown in Table 6, the accuracy for all the considered sources is in line with the general case, with percentages of signatures compatible with the official signatures stored in the Sapienza database ranging from 96.67% to 99.60%, depending on the database. Note that compound names appear instead to be more frequent than in the general case, with percentages ranging from 13.73% to 17.16% for last names,

	WoS	*Scopus*	*PubMed*	*CrossRef*
Total signatures	1354	1750	1002	1591
Compatible signatures (% of total signatures)	1309 (96.67%)	1705 (97.43%)	998 (99.60%)	1565 (98.36%)
Signatures with compound last name (% of total signatures)	186 (13.73%)	247 (14.11%)	172 (17.16%)	213 (13.38%)
Signatures with compound first name (% of total signatures)	270 (19.94%)	321 (18.34%)	220 (21.95%)	263 (16.53%)
Signatures with diacritics in last name	none	none	none	none
Signatures with apostrophes in last name (% of total signatures)	23 (1.69%)	28 (1.60%)	29 (2.89%)	23 (1.44%)

Table 6 Statistics on the author signatures with non-Italian last names in our data set.

	WoS	*Scopus*	*PubMed*	*CrossRef*
Bad splits over last names (% relative to signatures with compound last names)	19 (10.21%)	16 (6.47%)	1 (0.58%)	12 (5.63%)
Incomplete last names (% relative to signatures with compound last names)	1 (0.53%)	1 (0.40%)	none	1 (0.46%)
Last names with typos (% relative to total signatures)	25 (1.84%)	27 (1.54%)	3 (0.29%)	11 (0.69%)
Bad splits over first names (% relative to signatures with compound first names)	none	none	none	2 (0.76%)
First names with typos (% relative to total signatures)	none	1 (0.05%)	none	1 (0.06%)

Table 7 Statistics on how signatures with non-Italian last names in our data set are misrepresented in different bibliographic databases.

and from 16.53% to 21.95% for first names. This is motivated by the fact that many foreign names in our dataset are of a Spanish origin, which tend to be naturally compound [Ruiz-Pérez et al., 2002].

As shown in Table 7, bad splits over last names range from 0.58% to 10.21%, while incomplete last names range from 0% to 0.53%. Compared to the general case, we noted a slightly higher number of typos in last names: the numbers are however too small to draw any statistically sound conclusions for this case. Also, bad splits over first names and first names with typos are rare occurrences in our sample. Error categories for which we found no instances in the sample are omitted from Table 7.

7 Summary

In this section we provide a summary of our findings, distilling the key messages. If you are a Sapienza author with a compound last name (roughly 8% of the total), then there is a non-negligible probability (between 5.24% and 12.61%, depending on the bibliographic source) that part of your last name will in fact be assigned to your first name. If your compound last name has non-Italian origin, then the maximum probability is slightly lower (10.21%). However, the probability of having a compound last name roughly doubles in this case. *CrossRef* and *WoS* also have non–negligible probabilities of dropping at least one token (5.85% and 1.25%, respectively). If your compound last name is not Italian, then the highest probability drops to 0.53% for this scenario. If your last name contains diacritics (admittedly a relatively rare occurrence, as approximately only 1% of last names exhibit this feature), then these are extremely likely to be omitted by both *WoS* and *Scopus* (probabilities are 97.07% and 94.99%, respectively). *PubMed* and *CrossRef* have much lower, but still significant probabilities to drop diacritics from your last name (8.70% and 19.32%, respectively). Moreover, *CrossRef* also shows incorrectly imported diacritics in approximately 4% of all occurrences. Apostrophes in last names are slightly more common than diacritics (roughly 1.4% of last names contain them). In this case, *WoS* has the highest probability of dropping them (9.59%), followed by *CrossRef* (1.90%), while the probabilities for *Scopus* and *PubMed* are much lower (0.67% and 0.57%, respectively). If your first name is compound, then both *CrossRef* and *Scopus* have low, but non–negligible probability of erroneously attributing at least one of them to your last name (1.39% and 0.96%, respectively).

The other error categories we have investigated (typos in both first and last names, as well as errors that escape our classification) seem to be quite rare occurrences, at least for Sapienza authors.

8 Conclusions

In this article we have focused on the accuracy of four prominent international bibliographic data sources (namely, *WoS*, *Scopus*, *PubMed*, and *CrossRef*) with respect to author names. Our case study consisted of 44,549 publications, stored in the internal database of Sapienza University of Rome, one of the largest universities in Europe (Section 4 provides for more details on our data set). Our results indicate generally good accuracy for all data sources, even though some specific areas of concern have emerged, particularly in the handling of last names composed of more than one token, and of diacritical marks (Section 6 provides a detailed exposition of our results). In some specific cases, errors can be severe and may impact further bibliographic analyses such as attribution of articles to specific authors. While this may be statistically irrelevant for large-scale research assessment exercises where a few errors may be negligible, errors that affect individuals such as name-related issues can play a significant role for identification of experts, recruitment decisions, and career advancements based on bibliometric indicators [Olensky, 2015].

The results we obtained for the case of Italian names are consistent with previous investigations for other Western languages such as Spanish [Ruiz-Pérez et al., 2002] (as discussed in Section 3), where compound last names and the use of diacritics are even more frequent than in Italian. Furthermore, most European languages, with the notable exception of English, use letters with a variety of diacritical marks or even new symbols that are not part of the basic Latin alphabet. We note that increasing the accuracy of bibliographic databases to support these special features calls for localized data processing procedures. We hope that the study presented in this article, and in particular the methodology used in Section 5 to preprocess and compare different name versions, can shed light on some of the critical aspects that emerge in the context of name processing in Western languages closely related to Italian.

A few research directions remain open. First, with our processes and methodology in place (see Section 4 and Section 5, respectively), in a manner similar to what has been done in this article for author names, the accuracy of other fields in bibliographic records (e.g., article's title, journal title, volume number, publication year, etc.) could be assessed as well. Further analysis may help shed light on the causes of discovered inaccuracies in bibliographic data sources (e.g., errors in the original published version, OCR errors, format misinterpretation by some bibliographic source parsing algorithm). Finally, it would be interesting to assess how the detected inaccuracies (and their combinations) impact on the citation matching process of bibliographic sources (i.e., whether citations are missed or mismatched) and authorship attribution.

Acknowledgements

First of all, we wish to thank the anonymous reviewers for their useful suggestions. We are also indebted to Irene Bongioanni, Adriano Fazzone, Emanuele Fusco, and all members of Sapienza University of Rome who contributed to the analysis and cleaning of the bibliographic records stored in our local repository. Marco Schaerf was partially supported by H2020 Project Second Hands under grant agreement No 643950.

References

G. Abramo, C.A. D'Angelo, and F. Di Costa. National research assessment exercises: the effects of changing the rules of the game during the game. *Scientometrics*, 88(1):229–238, 2011. doi: 10.1007/s11192-011-0373-2. cited By 5.

Dag W. Aksnes. When different persons have an identical author name. how frequent are homonyms? *Journal of the American Society for Information Science and Technology*, 59(5):838–841, 2008. ISSN 1532-2890. doi: 10.1002/asi.20788. URL `http://dx.doi.org/10.1002/asi.20788`.

Denise Bennett and Priscilla Williams. Name authority challenges for indexing and abstracting databases. *Evidence Based Library and Information Practice*, 1(1):37–57, 2006. ISSN 1715-720X. doi: 10.18438/B81596. URL https://ejournals.library.ualberta.ca/index.php/EBLIP/article/view/7.

Fiorenzo Franceschini and Domenico Maisano. Critical remarks on the italian research assessment exercise vqr 2011-2014. *Journal of Informetrics*, 11(2):337 – 357, 2017. ISSN 1751-1577. doi: http://dx.doi.org/10.1016/j.joi.2017.02.005. URL http://www.sciencedirect.com/science/article/pii/S1751157716303005.

E. Garfield. What's in a surname? *Naturwissenschaften*, 68(10):519–520, 1981. ISSN 0028-1042. doi: 10.1007/BF00365376. URL http://dx.doi.org/10.1007/BF00365376.

A.W. Harzing and R. van der Wal. Google scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, 8(1):61–73, 2008.

William W. Hood and Concepción S. Wilson. Informetric studies using databases: Opportunities and challenges. *Scientometrics*, 58(3):587–608, 2003. ISSN 1588-2861. doi: 10.1023/B:SCIE.0000006882.47115.c6. URL http://dx.doi.org/10.1023/B:SCIE.0000006882.47115.c6.

V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Lokman I. Meho and Kiduk Yang. Impact of data sources on citation counts and rankings of lis faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, 58(13):2105–2125, 2007. ISSN 1532-2890. doi: 10.1002/asi.20677. URL http://dx.doi.org/10.1002/asi.20677.

Marlies Olensky. Testing an automated accuracy assessment method on bibliographic data. *Journal of Library and Information Studies*, 12(2):19–38, 2014.

Marlies Olensky. *Data accuracy in bibliometric data sources and its impact on citation matching.* PhD thesis, 2015. URL \url{http://edoc.hu-berlin.de/docviews/abstract.php?id=41398}.

M.L. Pao. Importance of quality data for bibliometric research. In *Proceedings of the 10th national online meeting. Medford, NJ, Learned Information*, pages 321–327, 1989.

R. Ruiz-Pérez, E.D. López-Cózar, and E. Jiménez-Contreras. Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies. *Journal of the Medical Library Association*, 90(4):411–430, 2002.

Dirk Tunger, Stefanie Haustein, Lena Ruppert, Gaetano Luca, and Simon Unterhalt. "The Delphic oracle" An analysis of potential error sources in bibliographic databases. 11th International Conference on Science and Technology Indicators, Leiden (The Netherlands), 9 Sep 2010 - 11 Sep 2010, Sep 2010. URL http://juser.fz-juelich.de/record/138630.