Scientometrics
https://doi.org/10.1007/s11192-025-05368-1

3SA: an entity-linking algorithm for the Institution Name Disambiguation problem in affiliations using edit distance

David Muñoz-Jordán[1] · Gonzalo Ruiz[1,2] · Pablo Cabriada[2] · Juan Luis Durán[2] · David Iñiguez[1,2,3] · Alejandro Rivero[1,2]

Received: 6 February 2025 / Accepted: 13 June 2025

Abstract
When researchers sign an article, they reference all the institutions they belong to, writing one or more affiliations containing them. Researchers sign in many different ways, and different journals also have varying standards in this regard. In this article we will focus on the Institution Name Disambiguation (IND) problem, also known as Organization Name Disambiguation (OND). Common issues associated to IND problem arise because researchers may write the name of the institution differently in various publications, and different researchers from the same institution will certainly write it differently as well. On the other hand, a researcher may be affiliated with several centers simultaneously or at different stages of their professional life, which introduces the factor of time as an additional variable to consider. As a result, analyzing and linking scientific work from different areas for various institutions is challenging. Databases like Web of Science collect articles from various journals across different fields. In this article, we will propose a method named 3 Steps Affiliation (3SA) based on, firstly, preprocessing the information, secondly, candidate extraction via localization and classification type of the institutions and, thirdly, on entity linking to extract the institutions from affiliations downloaded from Web of Science articles using an edit distance. We use a world-wide open source database with more than 100k institutions to solve the Institution Name Disambiguation problem. We show that the proposed method has a state-of-art performance by comparing it with other methods. Additionally, we evaluate the impact of different edit distance metrics within our method to identify which yields the best results.

Keywords Institution Name Disambiguation · Organization Name Disambiguation · Affiliations Disambiguation · Infometrics · Entity linking · Edit distance

Muñoz-Jordán, David and Ruiz, Gonzalo are the main and equal contributors to the article, having led its conceptualization, methodology, analysis and writing. The rest of the authors contributed with material preparation, literature review, methodology, analysis and reviewed the manuscript. All authors read and approved the final manuscript.

Extended author information available on the last page of the article

Introduction

Scientific production continues to grow globally every year. The volume of data is immense, and processing it for analysis becomes an increasingly challenging task. This scientific production is published in various outlets, using different formats and standards. Databases or repositories exist to compile all this scientific production, enabling searches by titles, keywords, researchers, institutions, areas, and various criteria. Within the scope of this work, we will focus on the bibliographic repository Web Of Science[1] because it is widely used and has specific characteristics, although alternatives such as Scopus also exist, and new open ones are appearing and rapidly growing such as OpenAlex. Large-scale analysis of scientific production is particularly valuable for studying collaboration patterns, identifying the most influential works, determining leading institutions in different fields, computing metrics, and more. The foundation for achieving this lies in accurately linking the scientific production to the researchers and participating institutions, which should also be correctly identified. Due to the great variety of formats in which this information is presented and the specific way each researcher provides it when publishing a new work, this becomes a highly complex task. In this work, we focus on addressing the problem of identifying institutions mentioned in scientific publications indexed in Web Of Science through affiliations. This problem is known as Institution Name Disambiguation (IND), although some authors have also referred to it as Organization Name Disambiguation (OND). For consistency, we will refer to it as IND throughout this article. The objective is to accurately detect all institutions mentioned in a given affiliation. This problem was first introduced in the late 1980 s (Bruin & Moed, 1990).

The solution to this problem is highly beneficial for addressing another closely related challenge, known as Author Name Disambiguation (AND). Similarly to how the IND problem aims to identify institutions from an affiliation provided by a researcher, AND involves, given a set of scientific publications, determining for each researcher the list of publications in which he or she has participated. This problem is complex because, as with affiliations, the way a researcher writes his or her name when publishing a work may vary throughout their career for various reasons, such as the publishing journal or the institution where they are based at a specific time. Additionally, in some cases, only partial information about a researcher's name is provided, such as using initials for compound names, omitting parts of the name, making slight variations to avoid confusion with other researchers sharing the same name, and other similar cases. Due to this variability in how a person's name is represented, as well as the potential coincidence of that name with others, it is necessary to complement a researcher's information beyond just their name. To achieve this, solving the AND problem relies, among other factors, on the institutions to which the researcher is affiliated, making it essential to first address the IND problem.

The typical challenges encountered when solving the IND problem are as follows:

1. **Different languages.** It is common in some countries to report affiliations in the local language. This issue can be exacerbated in countries with co-official languages, such as Spain, where affiliations might appear in Catalan, Galician, or Basque, for example.
2. **Lack of delimiters.** Authors, when providing their affiliation, may include several institutions within it. However, these are not always separated by commas or any other

[1] https://www.webofscience.com/

characters that would allow for a clear and immediate division, making it difficult to identify each institution within the same affiliation.
3. **Acronyms.** Sometimes, authors use an acronym instead of the full name of an institution to reference it, requiring knowledge of the acronyms associated with various institutions.
4. **Irrelevant information.** Affiliations often include references to non-essential information, such as the research group the author belongs to, or postal address elements of the institution's location.
5. **Institutional transition.** Over time, an institution may change its name, merge with others, or split into separate entities.
6. **Spelling errors.** As humans, researchers may occasionally make mistakes when writing their affiliation.
7. **Multiple locations.** A single institution may have different branches or operate in multiple locations.
8. **Synonyms.** The same institution might be referred to in multiple ways, using different variations of its name.
9. **Homonyms.** Sometimes, institutions with the same name are entirely distinct. This situation often occurs in organizations based in different regions, typically in different countries.
10. **Abbreviations.** Authors may use abbreviations, such as writing "ctr" instead of "center," for example.

There are two main reasons why we have chosen WoS. Firstly, it has been observed in it that since 2014, 90% of author signatures have an affiliation associated (although the institutions referenced in the affiliation are not identified, which is the problem we want to solve), and by 2016, this percentage increased to 98% (Maddi & Baudoin, 2022). Therefore, WoS is a good choice to address the AND problem due to its completeness. Secondly, as we will show later, it is also a clear example of the last problem mentioned above, as it contains many abbreviations in its affiliations. We will compare affiliations extracted from WoS and OpenAlex to see the differences in how the same affiliation is written and stored differently and to demonstrate how this may affect the performance of methods discussed in the literature.

In Section 2, we will survey the literature, presenting the first steps that were taken to address this problem, how different strategies have been emerging thanks to the appearance of institution databases, and reviewing the state of the art. Subsequently, in Section 3, we will describe the strategy proposed in this work to solve the problem. Next, in Section 4, we will present the data used and the experiments conducted to evaluate our method against others. In Section 5, we will analyze the results and compare them with those obtained using alternative methodologies found in the literature. Finally, in Section 6, we will summarize the work carried out and the results achieved.

Literature review

During the past decades, various methods have been developed to address the problem, evolving significantly over time. In their initial approaches, there was no starting database that collected institutions at an international level, so methods based on unsupervised learning models, such as clustering, were developed. For instance, in Jiang et al. (2011),

an agglomerative algorithm is proposed, where the objective is not to identify institutions per se, but to unify all the strings within the same affiliation without identifying the distinct institutions within it. The starting point of this method involves having as many clusters as different affiliations we have, which are iteratively grouped into clusters until the desired number, a parameter of the model, is reached. The distance between clusters is defined as the average distance of all pairs of elements within each cluster. The distance between two different affiliations is calculated using the Normalized Compressed Distance (NCD), which involves compressing the strings with the gzip[2] library. However, for the validation of the method, affiliations from a single university are used.

At the same time, other authors approached this problem under the name of Organization Name Disambiguation (OND), as we previously mentioned. For example, in Jonnalagadda and Topham (2011), the problem is addressed using a clustering algorithm based on word edit distance. This work also tackles the Named Entity Recognition (NER) task by using a word corpus to detect which part of an affiliation corresponds to an organization, a city, a country or state.

In 2013, the study (Cuxac et al., 2013) introduced both a supervised method employing Bayes' rule to calculate probabilities and a semi-supervised method with modifications to the k-means algorithm. The most widely adopted clustering-based approach is the method by (Huang et al., 2014), which consists of three steps. First, an author-affiliation table is created by unifying all authors with the same last name and first initial. Second, for each resulting author, pairs of institutions that could potentially be the same are identified using various rules and country information. The third and final step involves merging the pairs obtained in the previous step, based on the following condition: if an author has the pair of affiliations (N1, N2) with a frequency exceeding a threshold, and another author has the pair (N1, N3) with a frequency exceeding the threshold, then (N1, N2, N3) are considered the same affiliation. This method achieves high precision but low recall. Subsequently, further advancements have been made in this problem, leading to the development of new methods. For example, in Donner et al. (2020), a method called KB System is compared using data from WoS and Scopus, although it focuses exclusively on German institutions.

More recently, international databases of institutions have emerged, leading to a new methodology for addressing the IND problem, known as entity-linking. In this methodology, a database of institutions is available, and the goal is to identify these institutions within the affiliations of scientific production. The earliest models developed under this approach continued to use metrics based on edit distances. For instance, in Zhou et al. (2020), affiliations are divided into n-grams. However, with the availability of a database of institutions used to disambiguate affiliations in scientific production, the concept of the type of institution gained prominence, as it provides critical information for disambiguation. In Backes et al. (2022), a hierarchical method is proposed to identify institutions belonging to different levels of hierarchy, using a list of keywords to detect their type. In Ancona et al. (2023), the IND problem is addressed within the CORDIS data. To do so, they use a dataset of abbreviations from WoS to detect different forms of writing a word and normalize them. After this normalization step, affiliations are compared with institutions from a database using a set of rules based on character and word counts, along with a "cosine" edit distance. They also define a control variable, the country, which if it is different between the affiliation and the institution on the database, automatically leads to the entities being considered as distinct institutions.

[2] https://www.gzip.org/

The advantages and utility of this approach compared to clustering proved overwhelming, prompting efforts to construct an international database of institutions as comprehensive as possible (Huang et al., 2020; Lammey, 2020). Currently, with the advent of Pretrained Language Models (PLMs), research focuses on replacing metrics based on edit distances with those based on semantic distances (Duran-Silva et al., 2024; Jia et al., 2024).

The application of these techniques forms the foundation for solving other problems, such as Author Name Disambiguation, or conducting studies and analyses based on the scientific production of each institution. For example, in de Marcos et al. (2024), a study focused on journals and JCR impact is conducted, leveraging graph construction based on institutions, for which their accurate identification is essential.

Our method. 3SA

Our method follows the trend of recent works mentioned above, specifically the technique known as entity-linking. This technique involves, given an affiliation, attempting to infer which institutions are explicitly mentioned within it by comparing to a reference database of institutions, specifically the Research Organization Registry[3] (ROR). Note that we define an affiliation as a list of institutions, generally related to each other and of different hierarchical levels, that an author has referenced in a scientific work to indicate their relation with these institutions at the time of publication.

In some studies, methods have been proposed to infer additional institutions not explicitly mentioned in the affiliation by leveraging relationships between institutions recorded in the reference database itself (Backes et al., 2022). In our method, we restrict ourselves to institutions explicitly named in the affiliation, as while it is sometimes possible to infer institutions through such relationships, this task depends entirely on the quality of the institution database used as a reference.

Our method, called 3 Steps Affiliation (3SA), is based on three steps: pre-processing the affiliation, generating candidates, and entity-linking.

Step 1. Pre-process affiliation

As explained in Section 1, the challenges of the problem we aim to solve stem from the numerous variations in how the same institution is referenced in scientific production. These variations include different abbreviations of the same word, the language in which the affiliation is written, transcription errors, and the specific manner in which the institution is referenced within the affiliation. To address these challenges, we propose a preprocessing method applied both to the institutions recorded in the reference database and to the affiliations analyzed from scientific production. The objective of this pre-processing step is to extract as much local information as possible from the institution and to process the text string of the institution's name to allow better comparisons. Given an affiliation extracted from WoS, we will split it by commas and search for institutions in the reference institution database.

As mentioned before, we define an affiliation as a text string that may reference one or several institutions, generally at different hierarchical levels and related to each other. A typical example of an affiliation extracted from WoS would be "Royal North Shore Hosp,

[3] Link to the database: https://ror.org/

Northern Sydney Local Hlth Dist, St Leonards, NSW 2065, Australia" Within each affiliation, we can find different types of information about the referenced institutions, such as the country, city, address, postal code, or names of institutions at various hierarchical levels. Some works focus on detecting and classifying this information (Duran-Silva et al., 2024), as the degree of differentiation and classification depends on the source. In the case of WoS, the focus of this study, this step is not necessary, as the information is well-structured and standardized. It uses commas to separate the different institutions, city, address, postal code, or country, and follows an order from lower to higher hierarchical levels. In the example above, by splitting the affiliation string by commas, the first element corresponds to the institution of Royal North Shore Hospital, the second refers to the institution of Northern Sydney Local Health District, the third is the city, in the fourth there is a postal code, and the last is the country. As mentioned above, the last elements always refer to geographic locations, while the first elements contain the institutions. Thus, in this preprocessing step, the first task is to normalize the location of the affiliation to leverage this information for entity identification.

To normalize location information, that is, to standardize place names in affiliation data by matching them to a consistent reference format, we adopt a strategy similar to the one used for institutions, starting from a reference database of places for comparison. This database contains four hierarchical levels in a tree structure: country, level 1 administrative division, level 2 administrative division, and city. Levels 1 and 2 may be present or not depending on the country, as administrative divisions vary. For example, in the United States, level 1 corresponds to states, while in Spain, level 1 corresponds to autonomous communities, and level 2 to provinces. Each entry in this database includes the name in UTF-8, its ASCII and lowercase representation, and, in some cases, alternative accepted spellings. Additionally, there is supplementary information for each record, such as ISO2 and ISO3 codes for countries. Finally, parent-child relationships are included. We use the Levenshtein distance, also known as edit distance, to compare geographic reference database entries with those extracted from scientific production, aiming to normalize the locations associated with affiliations. The comparison uses a similarity threshold of 0.1; that is, when the edit distance between the reference place string from the affiliation and the name of the place in the database is below this value, they are considered to represent the same location. This process allows for disambiguating and standardizing places that may exhibit different spellings or typographical errors. Before performing comparisons, text strings are pre-processed by converting them to ASCII and lowercase format. The set of reference locations was generated from data extracted from the GeoNames database[4].

Another relevant aspect of institutions is that they can be classified. This classification can vary in granularity and allows for hierarchical structuring. For example, in the research domain, an institution can be a university, corresponding to a specific type of entity. Within a university, we can identify several entities such as research institutes, centers, schools, faculties, laboratories, or departments. Similarly, university hospitals exhibit comparable structures, as they also contain departments. Defining a hierarchy of research institutions is highly complex due to the broad variety of cases. For instance, research institutes may belong to a single university, but there are also those affiliated with multiple universities, independent institutes that do not depend on any university, and even mixed institutes, which are affiliated with both a university and another organization.

[4] Link to GeoNames: https://www.geonames.org/

In our method, by splitting each affiliation into blocks separated by commas, we attempt to classify the type of institution represented in each block. To this end, we have defined a classification with the most common types of institutions, as shown in Table 1, which are critical for our method to work properly. For each classification, we have compiled various forms of writing or synonyms, including typical abbreviations and languages for that classification. Each extracted block is converted to lowercase ASCII, and certain stop-words, such as articles or prepositions, are removed. If a word in the block matches one of the terms defined for a classification, we assign that classification to the block. If no match is found, we classify it as "Other". In cases where multiple words in the same block suggest different classifications, we prioritize them according to the following order:

1. Department
2. School
3. Faculty
4. Center
5. Laboratory
6. Institute
7. Foundation
8. Hospital
9. University

This pre-processing step is also applied to the reference institution database to normalize the locations of institutions and classify them into different categories.

Step 2. Candidate extraction

After the pre-processing step described above, we leverage the extracted location and classification information. The location of an institution, along with its classification, is a key factor in recognizing it. Location can help differentiate between two institutions with very similar or even identical names. For example, there is a "Universidad de Córdoba" in both Spain and Colombia. Similarly, in the United States, we can find "Miami University" in Oxford, Ohio, and the "University of Miami" in Coral Gables, Florida. As we delve deeper into hierarchical levels involving institutes, centers, and other types of institutions, the number of homonyms is likely to increase, making location even

Table 1 Institution classifications and abbreviations defined with alternative names

Classification	Abbreviation	Alternative names
University	univ	universidad, université, università, universität
Hospital	hosp	hôpital, ospedale, krankenhaus
Foundation	found	fundación, fondation, fondazione, stiftung
Center	ctr	centro, zentrum
Institute	inst	instituto, institut, istituto
Laboratory	lab	laboratorio, laboratorie, labor
Department	dept	departamento, département, dipartimento, abteilung
School	sch	escuela, école, scuola, schule
Faculty	fac	facultad, faculté, facoltà, fakultät

more valuable for disambiguation. The other crucial aspect obtained from the previous step is the classification of the institution. While it is not always possible to infer this from the institution's name, it provides highly valuable information when available. For example, in the ROR database, both the "University of Navarra" and the "University of Navarra Foundation" exist. By applying the pre-processing step in our method, we classify the former as a university and the latter as a foundation, significantly reducing the likelihood of considering them as the same institution.

In earlier methods that used clustering-based algorithms, some authors differentiated affiliations by country (Huang et al., 2014). In more recent methods that employ PLMs, the city and/or country are also included in the institution's name during comparisons (Duran-Silva et al., 2024; Barret & Priem, 2024). The work by Backes et al. (2022) also classifies the types of institutions found in an affiliation, defining specific types and levels.

Our proposed method relies on the extracted location and classification information to search for candidates from the reference institution database, applying a progressively less restrictive approach. This candidate search method, followed by selecting the best match, is also employed in other works such as Zhou et al. (2020).

After pre-processing the affiliation extracted from scientific production in the first step, we obtain a list of normalized and cleaned blocks as follows:

- Converted to lowercase.
- Converted to ASCII.
- Stop-words such as prepositions or articles removed.
- Keywords referring to types normalized to their abbreviations.

Additionally, we extract normalized and cleaned location and entity-type information, the former using GeoNames' location dataset and the latter using the classifications defined earlier. For each block, we prepare sufficient information to compare with the reference institution database, which has undergone a similar process. Using this curated information, we perform a search in the ROR database for institutions likely to be included in the block, which will form the candidate set. A scoring system is then applied to identify the closest match among the candidates. This matching process is designed to be sequential: we apply a series of increasingly less restrictive criteria, and as soon as a suitable match is found, the process stops. This allows us to achieve the best results while minimizing comparisons. The sequence is defined as follows.

1. The first set of candidate institutions is obtained by searching for acronyms using regular expressions. Acronyms extracted from ROR are searched within the unprocessed affiliation string, leveraging the extracted location information to refine the match. Acronyms can sometimes overlap across countries or even within the same country. To avoid selecting the wrong candidate or disambiguating acronyms that correspond to multiple institutions in the same country, location information is used. For example, in Spain, "UPV" refers to both the Basque Country University and the Polytechnic University of Valencia; however, one is located in the Basque Country, and the other in the Valencia Autonomous Community. Acronyms receive special treatment in our approach. When an acronym is detected, the corresponding ROR institution is assigned to the affiliation; however, the matching sequence does not terminate at this point. Instead, the acronym is

removed from the affiliation string, and the process continues with the remaining text, if any, to identify additional institutions.
2. The second set of candidates is obtained by matching both location and classification, verifying institutions that match on these criteria.
3. The next criterion generates candidates based solely on classification.
4. The final set of candidates is generated based only on the location of each block, following a progressively less restrictive strategy.

Candidate selection by location also follows a progressively less restrictive strategy. The most restrictive case requires matching the city, although this can sometimes be limiting since a single institution may have branches in different cities, with only the main location recorded in ROR. Therefore, we also allow matches based on level 1 and level 2 administrative divisions, as explained in Section 3.1. The least restrictive case involves matching only by country. In the third step described above, where candidates are selected based solely on classification, we only consider institutions where the country also matches. Thus, normalization of location information in both the affiliation under analysis and the reference database is critically important.

Step 3. Entity linking

The first set of candidates obtained from acronyms is directly considered as correctly identified institutions. For the other sets of extracted candidates, a distance is calculated between the corresponding block of the affiliation to be identified and each candidate institution from ROR. This distance is based on a variant of the Levenshtein distance[5] called "setratio", which takes two phrases as input, splits them into words using spaces, and seeks the best match between both sets of words, regardless of their order. The distance between two words is still calculated using the Levenshtein distance, and the total distance between the two phrases takes a value between 0 and 1. When an institution has multiple names or synonyms, the name with the minimum distance is used. For each block, the institution with the smallest distance is selected, and if the distance is less than 0.18, we consider that the block corresponds to the candidate ROR institution. To support the choice of edit distance, we compare different edit distance options and demonstrate that "setratio" yields the best performance.

To enhance the efficiency of this step in our method, the candidate extractions for each block corresponding to an affiliation are progressively obtained. That is, if in a candidate extraction we find an institution with a distance below 0.18, subsequent candidate extractions for that block are skipped. As a result, candidates with matching locations or/and classifications have priority to be linked.

Comparison with existing techniques

In our literature review, we have observed the evolution of various approaches to the IND problem. With the emergence of institutional databases such as ROR, the prevailing approach is to extract institutions from these databases as they appear in affiliation strings.

With recent advances in PLMs, some authors have begun using these models to compute similarity metrics. For example, Duran-Silva et al. (2024) use PLMs to compute string

[5] Link to the Python library: https://rapidfuzz.github.io/Levenshtein/levenshtein.html#setratio

similarity by embedding geographical information directly into the text. Similarly, Barret and Priem (2024) fine-tune a DistilBERT model using a custom dataset, also embedding geographical data in the text. In contrast, our method relies on edit distances, avoiding the need for specialized hardware required to run PLMs. Moreover, our approach includes the normalization of geographical information through the use of a dedicated corpus, which helps correct errors or identify different spellings of the same location (synonyms). This enables a sequential extraction of candidate institutions from the reference database, effectively reducing unnecessary comparisons that could lead to errors.

Other methods also employ edit distance strategies. For instance, Ancona et al. (2023) use a cosine-based edit distance built on q-grams. They also address the problem of abbreviations by using a corpus from WoS to detect and normalize them, which aids in the comparison process. Additionally, they use country information as a "control variable", preventing comparisons when countries differ. What distinguishes our method is, first, the use of a geographical corpus, allowing us to recognize that "España" and "Spain" refer to the same country. Second, beyond handling abbreviations, we introduce a classification of abbreviations for later use. Third, thanks to the normalized location data, capable of identifying even the city when available, and the extracted classification, we apply a sequential candidate extraction process. This helps prevent unnecessary or misleading comparisons.

In summary, the novelty of our method lies in the combined use of a geographical corpus for location normalization, the detection and classification of abbreviations, and a sequential candidate extraction process. Table 2 highlights the main differences between our approach and existing techniques.

Data and experiments

To apply our methodology, it is necessary to have a dataset containing the institutions that we aim to identify within affiliations extracted from the scientific production to be analyzed. In our case, as previously mentioned, we will use the ROR database. Currently, this dataset contains 110,723 institutions worldwide[6], approximately classified as follows: 30,000 as "company", 22,000 as "education", 15,000 as "nonprofit", 13,000 as "healthcare", 11,000 as "facility", 7,000 as "government", 3,000 as "archive", and 9,000 as "other". Additionally, 17,000 parent-child relationships exist between institutions.

On the other hand, to test our method, we downloaded scientific articles from Web of Science and analyzed their affiliations. Specifically, scientific articles from 2019 to 2023 were selected, covering various topics shown in Table 3. A total of 53,358 articles were downloaded, from which 210,437 affiliations were extracted. Out of these, we randomly selected 1,000 affiliations and manually labeled with the ROR institution identifiers that the algorithm should detect. The labeling criterion required that institutions be explicitly named in the affiliation, either by their name or acronym. Thus, although some institutions could be inferred through relationships with others explicitly mentioned in the affiliation, they were not labeled, as this inference depends solely on the relationships in the reference database used (ROR in our case) rather than on the algorithm.

To evaluate the effectiveness of the proposed method and compare it with others in the literature, we generated two datasets. The first one, which we will reference as "WoS dataset", consists of ROR institutions manually labeled from 1,000 affiliations

[6] We downloaded version 1.52

Table 2 Comparison of features across affiliation disambiguation methods

Feature	3SA (This work)	(Ancona et al., 2023)	(Huang et al., 2014)	AffilGood (Duran-Silva et al., 2024)	OpenAlex (Barret & Priem, 2024)
Abbreviation handling	Structured, normalized and cleaned; common patterns and variants are standardized	Partial, heuristics plus some manual replacement	Not structured; based on keyword replacement	Often delegated to PLMs without explicit control	Often delegated to PLMs without explicit control
Institution classification	Explicitly extracted and used to filter and score candidates	Not used	Used only as loose rules	Typically ignored	Not used
Geographic information	Multi-level (country + admin levels 1 & 2 + city); normalized via corpus and used in filtering and disambiguation	Only country used as a control variable; not corrected	Country-level only	Country-level embedding-based context	City, state/region and country embedded in the string
Candidate selection	Sequential, rule-based filtering (acronyms, classification, geography); limits search space efficiently	All-to-all comparison with thresholds; no hierarchical filtering	Block-level clustering rules	PLM returns top-k matches with semantic score	Fine-tuned PLM compared against full database
Error correction	Typos and variants corrected via curated geographic and institutional corpora	No correction for misspellings; mismatches lead to split entities	Limited normalization (e.g., via dictionaries)	Rely on model generalization (opaque)	Rely on model generalization (opaque)
Manual validation	Less than 0.005% of comparisons flagged for review thanks to layered filters	All equivalences in the gray zone (between thresholds) require checking	None (fully automated)	Not transparent or easily verifiable	Not transparent or easily verifiable
Scalability/Efficiency	Lightweight; suitable for large datasets without PLMs	Hybrid but requires parameter tuning and manual checking	Efficient but rigid; prone to over- or under-merge	Computationally costly; depends on PLM infra	Computationally costly; depends on PLM infra; GPU needed for fine-tune
Transparency/Interpretability	Rule-based, explainable, and reproducible	Rule-based but threshold-dependent and heuristic	Some rules, but logic embedded in stages	PLMs = black-box behavior	PLMs = black-box behavior

downloaded from WoS as mentioned before. The second dataset, called "OpenAlex dataset", contains 564 of the affiliations in WoS dataset that we were able to extract from the same publications but found in OpenAlex database. To match these publications, we have done DOI and/or title searches. After this process, 564 affiliations from the manually labeled WoS dataset were matched with those extracted from OpenAlex. The primary reason for this limited overlap is that affiliations in OpenAlex publications often have fewer abbreviations in their names, making it easier to recognize institutions, as shown in the examples in Table 4. A manual review was conducted to ensure that the matched affiliations were indeed the same, considering them identical only if the institutions in each matched affiliation were consistent. Some cases were discarded because OpenAlex affiliations, being more complete, also included additional recognizable institutions not found in WoS affiliations.

Prior to the main evaluation, we demonstrate the improvement provided by our method compared to an approach without candidate extraction using the filters described earlier. This comparison is performed on the WoS dataset.

As a metric for evaluating model performance, we use the F1 score, calculated as follows:

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{ Recall}}{\text{Precision} + \text{ Recall}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FN} + \text{FP}}. \tag{1}$$

Specifically, we calculate the F1 score for each affiliation. Let A_i be the set of labeled institutions for affiliation i, and let B_i represent the institutions predicted by a model for the same affiliation i. There are two scenarios: in the first one, both sets are empty, $A_i = B_i = \emptyset$, and in this case, the F1 score for affiliation i is 1. In the second scenario, the number of true positives for an affiliation is the cardinality of the intersection $\text{TP} = |A_i \cap B_i|$, the number of false positives is the cardinality of the predicted elements not in the labeled set $\text{FP} = |B_i \setminus A_i|$, and the number of false negatives is the cardinality of the labeled elements not predicted $\text{FN} = |A_i \setminus B_i|$. In this case, the F1 score for affiliation i is calculated using Equation 1. To estimate the F1 score with its standard deviation, the dataset of 1,000 affiliations is divided randomly into 10 equal blocks. For each block, the F1 score is the average of the F1 scores calculated for each affiliation. The total F1 score is estimated as the mean and standard deviation of the F1 scores across all blocks.

In other works addressing this problem, authors calculate the macro F1 score, obtained by computing the F1 score for each institution and averaging these scores. However, we consider this metric less representative given the small labeled dataset available, as institutions with very few occurrences may lack sufficient data to estimate a statistically meaningful F1 score.

For the following experiment, we explore the performance of different edit distance metrics. We include the previously mentioned "setratio" distance, and compare it with Levenshtein, Jaccard, Jaro-Winkler, and the cosine distance as defined in Ancona et al. (2023). The cosine distance involves splitting each text string into q-grams and counting the frequency of each q-gram in the strings, followed by computing the cosine similarity. We experiment with $q = 1$, $q = 2$, and $q \in \{1, 2\}$, allowing for unigrams, bigrams, and both, respectively. We refer to these variants as "cosine1", "cosine2", and "cosine12". With the WoS dataset we adjust the optimal threshold and then we test it with the OpenAlex dataset.

Table 3 Web of Science categories selected to download papers

Web of Science category
Physics, Mathematical
Engineering, Mechanical
Computer Science, Artificial Intelligence
Radiology, Nuclear Medicine & Medical Imaging
Psychology, Psychoanalysis
Nutrition & Dietetics
Language & Linguistics
Clinical Neurology
Materials Science, Paper & Wood
Energy & Fuels

Table 4 Affiliation string comparison between Web of Science and OpenAlex for the same affiliation in the same paper

Affiliation from Web of Science	Affiliation from OpenAlex
Univ Sydney, Sch Phys, ISA, Sydney, NSW, Australia	ISA, School of Physics, University of Sydney, Sydney, Australia
Neurol Ctr San Antonio, San Antonio, TX USA	Neurology Center of San Antonio, San Antonio, TX, USA
Dokuz Eylul Univ, Fac Med, Izmir, Turkiye	Faculty of Medicine, Dokuz Eylul Universitesi, Izmir, Turkey
Hosp Univ St Joan Reus, Reus 43204, Spain	Hospital Universitari Sant Joan de Reus, 43204 Reus, Spain
Leibniz Inst Econ Res RWI, Essen, Germany	Leibniz Institute for Economic Research (RWI), Essen, Germany
Ocean Univ China, Sch Engn, Qingdao 266001, Peoples R China	School of Engineering, Ocean University of China, Qingdao, China
Univ Kiel, Chair Power Elect, D-24148 Kiel, Germany	Chair of Power Electronics, University of Kiel, 24148 Kiel, Germany
Neurofdn, Dept Neurol, Salem 636009, Tamil Nadu, India	Department of Neurology, Neurofoundation, Salem, Tamil Nadu 636009, India
Univ Lille, Fac Med, Lille, France	Faculty of Medicine, University Lille, Lille, France
Univ Pecs, Med Sch, Dept Pediat, Pecs, Hungary	Department of Pediatrics, University of Pécs, Medical School, Pécs, Hungary

In the last experiment, we compare our method with five others found in the literature. These include S2AFF (Feldman & Graham, 2023), the ROR API[7], AffRo (Myrto & Chatzopoulos, 2024), AffilGood (Duran-Silva et al., 2024), and the method used in OpenAlex (Barret & Priem, 2024).

[7] Link to documentation: https://ror.readme.io/v2/docs/rest-api

Results

Pre-process institution database

Using the location corpus constructed from GeoNames, we normalized the country for ROR institutions 99.98% of the time. Specifically, we were unable to normalize the country for only 18 institutions out of the 110,723 in the dataset. For 85% of ROR institutions (94,119 out of 110,723), we successfully normalized the city and/or administrative levels 1 and/or 2 in addition to the country. For 14.98% (16,586 out of 110,723), we were only able to normalize the country.

Regarding classifications, Table 5 shows the distribution of those extracted for ROR institutions using the definitions from Table 1. The classification "Company," derived from ROR's own categorization, has been added. Institutions classified as "Company" and "Other" represent over 60% of the total ROR institutions. Following these, the most common classifications are universities, institutes, centers, and hospitals. For our method, institutions classified as "Company" are treated as institutions classified as "Other".

Pre-process affiliation

While the WoS dataset includes 1,000 manually labeled affiliations, a total of 210,437 affiliations corresponding to the extracted data were preprocessed. Among these, the country was normalized in all but 4 cases. In 181,426 instances (86%), the city and/or administrative levels 1 and/or 2 were successfully normalized.

Regarding classifications, 557,085 blocks corresponding to institutions were extracted by splitting the 210,437 analyzed affiliations by commas. Table 6 shows the distribution of classifications obtained. It is notable that in 36.27% of blocks, no classification defined in Table 1 was found. Furthermore, the most common institution types are universities and departments.

Although the ROR database is highly comprehensive, it likely does not include all institutions worldwide. Estimating how many institutions are missing from ROR is a challenging task. To support this estimation, we also labeled which institutions in the WoS dataset do not appear in ROR. Table 7 presents the results. As shown, major institution types such as universities and hospitals are well covered by ROR. In contrast, for centers, companies, and institutes, the presence is more evenly split between institutions that are and are not included in ROR. Notably, faculties, schools, and departments, which are typically subunits of universities, are largely absent from the ROR database.

Candidate extraction

The second step of our algorithm is the candidate extraction, so we conducted an experiment to test whether our method of candidate extraction represents a significant improvement in terms of F1 score compared to not applying it. For this purpose, we applied a "brute force" algorithm to the WoS dataset. This algorithm is analogous to ours except for the candidate extraction step. In this "brute force" algorithm candidates are generated solely based on acronyms, and the remaining comparisons are made against the entire ROR database, omitting the successive candidate extraction proposed in our algorithm. We observe that the "brute force" algorithm achieves an F1 score of 0.80 ± 0.04 while ours achieves an F1 score of 0.86 ± 0.03, representing an improvement of 2σ.

Table 5 Distribution of classifications extracted from pre-processing the ROR database

Classification	Count	Percentage
Company	30680	27.71
Foundation	928	0.84
Hospital	6677	6.03
Other	39273	35.47
University	11204	10.12
Center	7034	6.35
School	1753	1.58
Faculty	68	0.06
Institute	10594	9.57
Laboratory	1462	1.32
Department	1050	0.95

Table 6 Distribution of classifications extracted from pre-processing the downloaded affiliations from the WoS dataset

Classification	Count	Percentage
Foundation	344	0.06
Hospital	38660	6.94
Other	202058	36.27
University	122761	22.04
Center	28693	5.15
School	16603	2.98
Faculty	16935	3.04
Institute	40278	7.23
Laboratory	6979	1.25
Department	83774	15.04

This result demonstrates that candidate extraction is essential, not only for time performance but also for accuracy. The reason for this is that different institutions can have similar names worldwide. A clear example is the "University of Miami" in Ohio and "Miami University" in Florida. Depending on how the institution is written in the affiliation, the minimum distance may not match the intended institution. By generating candidates as in our method, since both universities are in different states in this example, one of them would not be included in the comparison set, thus avoiding this issue. Note that an error of this kind implies a double penalty, as it adds a false positive by identifying an incorrect institution and a false negative by failing to find the labeled institution.

Moreover, under the same hardware conditions, our method has been shown to be 5 times faster than the "brute force" algorithm. This is because thanks to candidate extraction the average number of comparisons per affiliation is significantly lower than in an all-to-all comparison scenario.

Entity-linking

This is the final step of our algorithm, so we have pre-processed the affiliation string normalizing it and extracting location and classification information, and we have

calculated the candidates for each affiliation, and now we have to look for the best match of the institution reference database.

Initially, we conducted an experiment to compare the performance of different edit distances. The results are presented in Table 8. From these results, we observe that "setratio" and "cosine2" show similar performance on the WoS dataset, with "cosine12" and Levenshtein performing just a bit behind. However, when tested on the OpenAlex dataset, which differs slightly due to the absence of abbreviations, as shown in Table 4, "setratio" performs even better, while "cosine2" shows a decrease in performance. Based on these findings, we conclude that "setratio" is the superior choice for our method.

Finally, we conducted an experiment to compare our full algorithm with those from the literature. The results are shown in Table 9. Firstly, we observe that our method represents a significant improvement over S2AFF, ROR, AffRo, and AffilGood, while it is statistically similar to OpenAlex (0.25σ better than it), demonstrating that our method is state-of-the-art.

Secondly, our method, due to the pre-processing step, is resilient to the abbreviation problem mentioned in Section 1. This can be seen by comparing the results using the WoS dataset with the obtained using the OpenAlex dataset. As we mentioned before, affiliations in the WoS dataset contain more abbreviations than in OpenAlex, as shown in Table 4. The methods S2AFF, ROR, and AffRo show a significant difference in the results obtained using the WoS and OpenAlex datasets, with worse results in the WoS case, indicating that they are not prepared to address the abbreviation problem. Our method, thanks to the pre-processing step, is able to resolve this issue.

Conclusion

In this article, we have discussed the main difficulties in solving the IND problem such as synonyms, homonyms, lack of delimiters or abbreviations. In particular, using Web of Science data, we observed the significant challenges posed by the use of abbreviations in institutional names when referencing affiliations. In many cases, these abbreviations

Table 7 Number of institutions inside and outside ROR by classification in the WoS dataset

Classification	In ROR (%)	Not in ROR (%)
University	588 (99.8)	1 (0.2)
Department	2 (0.5)	376 (99.5)
Other	266 (86.6)	41 (13.4)
Institute	145 (60.4)	95 (39.6)
Hospital	166 (94.3)	10 (5.7)
Center	74 (50.7)	72 (49.3)
School	13 (13.0)	87 (87.0)
Faculty	0 (0.0)	75 (100.0)
Laboratory	7 (25.0)	21 (75.0)
Company	12 (63.2)	7 (36.8)
Foundation	6 (100.0)	0 (0.0)
Total	**1279 (66.4)**	**785 (33.6)**

Table 8 Mean and standard deviation of F1 score for different edit distances using our method. Data has been split in 10 subsets, computed F1 score in each subset and computed the mean and standard deviation

Dataset	setratio	Jaro-Winkler	Levenshtein	Jaccard	cosine1	cosine2	cosine12
WoS	$\mathbf{0.86 \pm 0.03}$	0.75 ± 0.04	0.83 ± 0.04	0.74 ± 0.05	0.77 ± 0.05	0.86 ± 0.04	0.84 ± 0.04
OpenAlex	$\mathbf{0.88 \pm 0.04}$	0.76 ± 0.07	0.83 ± 0.05	0.72 ± 0.06	0.73 ± 0.05	0.84 ± 0.04	0.81 ± 0.06

Table 9 Mean and standard deviation of F1 score for different methods. Data has been split in 10 subsets, computed F1 score in each subset and computed the mean and standard deviation

Dataset	S2AFF[1]	ROR[2]	AffRo[3]	AffilGood[4]	OpenAlex[5]	3SA
WoS	0.57 ± 0.04	0.55 ± 0.06	0.54 ± 0.04	0.82 ± 0.04		$\mathbf{0.86 \pm 0.03}$
OpenAlex	0.61 ± 0.02	0.72 ± 0.06	0.78 ± 0.05	0.82 ± 0.04	0.87 ± 0.04	$\mathbf{0.88 \pm 0.04}$

[1] https://github.com/allenai/S2AFF

[2] https://ror.readme.io/docs/rest-api

[3] https://github.com/openaire/affro

[4] https://github.com/sirisacademic/affilgood

[5] https://docs.openalex.org/how-to-use-the-api/api-overview

are applied arbitrarily and without following standard acronyms, making them the most critical obstacle in disambiguation efforts.

In this work, we have developed a candidate extraction-based methodology to address the IND problem through entity-linking, which we have named 3 Steps Affiliation (3SA). Candidate extraction relies on leveraging geographic information and classifying institutions into different types. The methodology has been shown to improve both computation time and results compared to an approach that does not extract candidates and instead compares against a complete institution database.

Additionally, we employ an edit-distance-based metric to score the similarity between two text strings, demonstrating that this approach obtains comparable results to those achieved using PLMs like AffilGood or OpenAlex, without the resource-intensive requirements of such large models.

Acknowledgements This work was partly supported by Grant No. PID2022-136374NB-C22 funded by Ministerio de Ciencia, Innovación y Universidades and Agencia Estatal de Investigación (Spain), by the Aragon Government through the research group E30_23R and by the Universidad de Zaragoza under the temporary research contract program "Programa Investigo" (Programa Investigo-081-74), funded by the Servicio Público de Empleo Estatal and the European Union–NextGenerationEU.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Ancona, A., Cerqueti, R., & Vagnani, G. (2023). A novel methodology to disambiguate organization names: An application to eu framework programmes data. *Scientometrics, 128*(8), 4447–4474.

Backes, T., Hienert, D., & Dietze, S. (2022). Towards hierarchical affiliation resolution: Framework, baselines, dataset. *International Journal on Digital Libraries, 23*(3), 267–288.

Barret, J., & Priem, J. (2024). openalex-institution-parsing. https://github.com/ourresearch/openalex-institution-parsing. Version 2.0.

Bruin, R. D., & Moed, H. (1990). *The unification of addresses in scientific publications*. Elsevier.

Cuxac, P., Lamirel, J.-C., & Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics, 97*(1), 47–58.

de Marcos, L., Goyanes, M., & Domínguez-Díaz, A. (2024). Mapping science through editorial board interlocking: Connections and distance between fields of knowledge and institutional affiliations. *Scientometrics, 129*(6), 3385–3406.

Donner, P., Rimmert, C., & van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies, 1*(1), 150–170.

Duran-Silva, N., Accuosto, P., Przybyła, P., & Saggion, H. (2024). AffilGood: Building reliable institution name disambiguation tools to improve scientific literature analysis. In Ghosal, T., Singh, A., Waard, A., Mayr, P., Naik, A., Weller, O., Lee, Y., Shen, S., and Qin, Y., editors, *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 135–144, Bangkok, Thailand. Association for Computational Linguistics.

Feldman, S., & Graham, D. (2023). S2aff - semantic scholar affiliations linker. https://github.com/allenai/S2AFF.

Huang, S., Yang, B., Yan, S., & Rousseau, R. (2014). Institution name disambiguation for research assessment. *Scientometrics, 99*(3), 823–838.

Huang, Y., Li, J., Sun, T., & Xian, G. (2020). Institution information specification and correlation based on institutional pids and ind tool. *Scientometrics, 122*(1), 381–396.

Jia, Z., Fang, Z., & Zhang, H. (2024). Normalization of web of science institution names based on deep learning. *ALGORITHMS, 17*(7), 312.

Jiang, Y., Zheng, H.-T., Wang, X., Lu, B., & Wu, K. (2011). Affiliation disambiguation for constructing semantic digital libraries. *JASIST, 62*, 1029–1041.

Jonnalagadda, S., & Topham, P. (2011). Nemo: Extraction and normalization of organization names from pubmed affiliation strings. *CoRR*, abs/1107.5743.

Lammey, R. (2020). Solutions for identification problems: A look at the research organization registry. *Science Editing, 7*(1), 65–69.

Maddi, A., & Baudoin, L. (2022). The quality of the web of science data: A longitudinal study on the completeness of authors-addresses links. *Scientometrics, 127*(11), 6279–6292.

Myrto & Chatzopoulos, S. (2024). Affiliation-matching repository [aka affro]. https://github.com/openaire/affro.

Zhou, S., Cao, X., Yuan, S., & Wang, Y. (2020). Elad: An entity linking based affiliation disambiguation framework. *IEEE Access, 8*, 70519–70526.

Authors and Affiliations

David Muñoz-Jordán[1] · Gonzalo Ruiz[1,2] · Pablo Cabriada[2] · Juan Luis Durán[2] · David Iñiguez[1,2,3] · Alejandro Rivero[1,2]

✉ David Muñoz-Jordán
dmunoz@bifi.es

Gonzalo Ruiz
gruiz@bifi.es

Pablo Cabriada
pcabriada@kampal.com

Juan Luis Durán
jduran@kampal.com

David Iñiguez
david.iniguez@bifi.es

Alejandro Rivero
arivero@unizar.es

[1] Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, Calle de Mariano Esquillor Gómez, Zaragoza 50018, Spain

[2] Kampal Data Solutions, S.L., Calle María Zambrano 31, WTCZ, Torre Oeste, Planta 15, Zaragoza 50018, Spain

[3] Fundación ARAID, Gobierno de Aragón, Zaragoza 50018, Spain