Backes, Tobias; Hienert, Daniel; Dietze, Stefan

Towards hierarchical affiliation resolution: framework, baselines, dataset

International Journal on Digital Libraries

International Journal on Digital Libraries (2022) 23:267–288
https://doi.org/10.1007/s00799-022-00326-1

Towards hierarchical affiliation resolution: framework, baselines, dataset

Tobias Backes[1] · **Daniel Hienert**[1] · **Stefan Dietze**[1,2]

Received: 12 July 2021 / Revised: 25 April 2022 / Accepted: 27 April 2022 / Published online: 28 May 2022

Abstract
Author affiliations provide key information when attributing academic performance like publication counts. So far, such measures have been aggregated either manually or only to top-level institutions, such as universities. Supervised affiliation resolution requires a large number of annotated alignments between affiliation strings and known institutions, which are not readily available. We introduce the task of *unsupervised hierarchical affiliation resolution*, which assigns affiliations to institutions on all hierarchy levels (e.g. departments), discovering the institutions as well as their hierarchical ordering on the fly. From the corresponding requirements, we derive a simple conceptual framework based on the subset partial order that can be extended to account for the discrepancies evident in realistic affiliations from the *Web of Science*. We implement initial baselines and provide datasets and evaluation metrics for experimentation. Results show that mapping affiliations to known institutions and discovering lower-level institutions works well with simple baselines, whereas unsupervised top-level- and hierarchical resolution is more challenging. Our work provides structured guidance for further in-depth studies and improved methodology by identifying and discussing a number of observed difficulties and important challenges that future work needs to address.

Keywords Entity resolution · Affiliation resolution · Formal concept analysis · Association rule learning · Taxonomy induction

1 Introduction

In the light of an ever-growing body of scholarly documents, it is crucial for entity-centric analysis to disambiguate publications, authors and institutions by specialized *Entity Resolution* (ER) methods such as *deduplication*, *author disambiguation* and *affiliation resolution*. One application is to automatically quantify scientific performance, e.g. to attribute unique publications or their citations to individual persons or institutions.

✉ Tobias Backes
tobias.backes@gesis.org

Daniel Hienert
daniel.hienert@gesis.org

Stefan Dietze
stefan.dietze@gesis.org

[1] GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany

[2] Heinrich Heine University Düsseldorf, Düsseldorf, Germany

This work is concerned with the distinction of higher education institutions, such as ranked, for example, by *Times Higher Education* [8]. In addition to prestige ("informed prejudice" [6]), these are important for making funding decisions or finding a place to study or work [7,26,31,46]. Previous work on such performance indicators has either used manually compiled lists of decision-making units mapped to the corresponding researchers or publications [2,12,14,27,29,34,36,47] or has automatically resolved affiliations to *top-level* institutions [5,13,15,22–25,30,37,38,44]. In contrast, impact measures for lower-level institutions could be aggregated automatically if affiliation records collected by providers like the Web of Science (WoS) [3,4] were properly resolved on the respective level.

This could be achieved by new methods addressing the task that we introduce as *'hierarchical affiliation resolution'*: affiliation resolution maps affiliation strings to real-world institutions. All prior resolution methods merge affiliations under top-level institutions. In reality though, institutions are hierarchically organized into substructures, e.g. a *university* might have *faculties* as branches, which again branch into

areas and then into *chairs*. Hierarchical resolution allows to arrange institutional entities on different levels to allow for a variable choice of granularity.

The focus of this work is to analyse this new task, providing a framework, a first baseline, data and evaluation methods to start experimentation and perform a thorough error analysis. We obtain a comprehensive first understanding of the task and lay foundations for future work by offering the following contributions:

C1 Requirements for ordering affiliations hierarchically
C2 Systematic framework meeting the requirements
C3 Implementation with first baseline components
C4 Datasets and measures for evaluation
C5 List of challenging aspects for future work

In Sect. 2, we cover related literature, discussing both application scenarios and existing top-level institution resolution methods. In Sect. 3, we discuss preliminaries such as bottom-up and top-down considerations concerning the structure of institutional hierarchies and their potential induction from data. Here, we arrive at a number of tasks inherent to the problem and derive a conceptual framework of modular components that can be specified to solve the tasks. Section 4 relates to the framework concrete observations made in the WoS affiliation data, suggesting first baseline methods to approach each of them. In Sect. 5, we discuss evaluation techniques, the datasets used and created by us and our experimental setup, whereas Sect. 6 introduces the results of these experiments and provides a thorough error analysis. In Sect. 7, we discuss those results and we conclude in Sect. 8.

2 Related work

A review of the literature suggests that the use of performance indicators in higher education for the purpose of evaluating individual institutions had become a major topic after the 1970's—and by the early 1990's had attracted considerable research interest.

Higher education performance indicators As early as 1977, Birch and Calvert [10] have discussed the use of performance indicators in higher education, but have focused on efficiency and effectiveness of teaching after concluding that the overall benefits of higher education remain elusive and thus cannot be measured. Ten years later, Ball and Halwachi [6] stress again the difficulty of defining quantifiable institutional goals. Citing the "Jarratt Report," they list a large number of potential performance indicators, among them various forms of staff publication output. Johnes [26] continues this discussion with a focus on publication and citation count as performance measures, both of which seem to be particularly popular today. Here, he lists a number of problems associated with both of them. While he acknowledges the now well-recognized argument that "some publications may be cited often (and so have considerable impact) not because they are of high quality, but because they are wrong," he does not mention the difference between quantity and quality when it comes to number of publications. Other related works from this time include Kells [31], Sizer et al. [46] as well as Ball and Wilkinson [7]. More recently, Johnes [28] summarizes public and private benefits of higher education, listing three major aspects of the measurements: (a) the entities for which measurements are aggregated, (b) the measures themselves and (c) their weighting. In 2019, Aksnes et al. [1] have contributed a more comprehensive overview of the relationship between citations and research quality.

Departmental comparisons In the context of our work, we are particularly interested in the above point (a) the institutional entities for which measurements are taken. Assuming that the task tackled in our work is particularly important for assessment of lower-level institutions, in the following, we list a number of performance comparison studies conducted on the departmental level. Davis and Papanek [14] compare 122 major economics departments by citation count. The departments where hand selected. Liebowitz and Palmer [34] tabulate the *development* of SSCI citations aggregated by sorting over 3000 researchers manually into more than 100 economics research groups (in the US, Canada and the UK) and compare the corresponding scores based on different weighting schemes. They stress various advantages of this methodology over previous survey-based approaches. Johnes and Johnes [27] analyse publication lists compiled manually by UK economics departments upon request from the Royal Economic Society. Su [47] compared survey-based assessments of university departments in Taiwan based on academics' opinions to those conducted by the government. Johnston et al. [29] compare peer reviews of all UK university departments' research quality collected by Universities Funding Council to the departments' sizes. Altanopoulou et al. [2] compare the h-index of 93 Greek university departments, which was collected manually by using Google Scholar to count the citations of the departments' research staff listed on department websites. Miroiu et al. [36] conduct a similar study for Romania and use the g-index instead. Recently, Chen and Chang [12] compare the performance of 33 departments of Chung Chen University in Taiwan, among others based on publication counts. They briefly discuss the manual selection process used to obtain these 'decision-making units.'

Interestingly, with the slight exception of the last, none of the above studies reports any problem in defining what counts as 'department' in an institutional hierarchy comprised of a multitude of levels. All of the studies use manually

created resources to compile the list of departments and to assign researchers or publications to them (so that, for example, publication counts can be aggregated by department). In our work, we show the difficulty of separating individual decision-making units in the institutional hierarchies (that also change over time) and explore opportunities to automatically assign publications to such units (independent of their level in the hierarchy). Thereby, we address the lack of scalable attribution methods in the above mentioned literature as well as obvious issues with limiting analysis to top-level institutions noted among others by Borgen and Mastekaasa [11] ("the college quality literature has generally treated educational institutions as homogeneous entities and has largely neglected the possibility of substantial within-institution quality variation, specifically across departments") and by Dillon and Smith [18] ("we know that individual students at larger colleges experience very different parts of what their institutions have to offer–for example, faculty research and teaching quality may differ across departments."). While there is no baseline for *hierarchical* affiliation resolution, a number of methods for top-level resolution have been proposed. In the following, we distinguish *linking*- (supervised), *clustering*-based (unsupervised) and mixed (semi-supervised) methods.

Linking-based resolution Linking is relatively simple as target institutions are known and affiliations only need to be assigned to the most similar institution, simplifying the task to finding a good similarity measure. Jacob et al. [24] extract affiliations from 48M resumes and link them to institutions extracted from Wikipedia applying retrieval-based candidate selection followed by similarity-based filtering. They report an accuracy of over 90%. Orduna-Malea et al. [38] use *Google Scholar*'s built-in affiliation suggestion method and evaluate it manually for Spanish universities, noting a lack of recall caused by missing institutions and weak typo-handling as well as some precision problems where the system incorrectly groups distinct universities. Shao et al. [44] deploy a method very similar to [24], linking to Chinese knowledge graphs and achieving 75% accuracy in their evaluation.

Normalization-based unsupervised resolution Bruin et al. [16] discuss normalization techniques, mentioning among others that due to address variations, alphabetical sorting is not sufficient for grouping the same or similar institutions. They propose a rule-based approach with considerable manual work to overcome these problems and conclude their method was *"highly successful"*. Galvez et al. [21] give many examples of ambiguous affiliations in the WoS. They use finite state transducers implementing mappings like *Computac* → *Comp* to normalize them. Using only a small set of related affiliations, they certainly underestimate the extend of variation in affiliation strings. The *Swedish Research Council (SRC)* [32] uses a pipeline of various match queries that complete the insufficient organizational information provided by *Thomson Reuters* and resolve affiliations to top-level institutions. Morillo et al. [37] perform position dependent keyword extraction to label affiliations by top-level sectors like university, company, health, NPO, other, etc. The difficulty of the task varies for different sectors (university: 0%–other: 50+%). Although this disambiguates to the sector-level and not the top-level institutions, knowing the sector can be useful for the latter task. Keyword extraction certainly is essential in our work. *Clarivate Analytics*, the current owner of the WoS, undertakes a procedure to normalize affiliation strings and enable top-level aggregation by automatic normalization, manual assignment and customer feedback [3,4]. All following works (like ours) are based on their normalized affiliations.

Comparison-based unsupervised resolution Aumueller and Rahm [5] estimate pairwise matching likelihood in terms of top-k search result overlap when entering normalized affiliations as queries into large-scale search engines. Like us, they first remove the address component of the affiliation string and then proceed with the institutional functions like university, faculty, chair—of which they extract the most general. They achieve 83% F1 with Google's top-8 results. Jiang et al. [25] use *normalize compression distance* to group WoS affiliations in a fixed number of clusters, which achieves better and more predictable results than a k-means baseline on a set of extracted features. Huang et al. [22] derive rules to recognize synonymous affiliations from patterns learned from matched affiliation strings. They align similar affiliation strings for the same author name and learn a rule-based mapping between them. While this is an interesting approach, it suffers among others from author name homonymy. Huang et al. [23] use country-base blocking followed by key collision and nearest neighbour search to achieve 91-98% F1. However, in the scenario evaluated in [44], their method only achieves 64% accuracy.

Semi-supervised resolution In line with our approach, Jonnalagadda and Topham [30] first identify address components like country or street in affiliations and continue to parse the rest using manually curated keyword lists. Cases of synonymy—in particular missing information—are tackled by agglomerative clustering with edit distance and retrieval towards the resulting *"organization clusters"*. Similarly, Jiang et al. [25] use agglomerative clustering on a *"normal compressor distance"* to disambiguate affiliations and interlink entities in a semantic web library. Cuxac et al. [13] propose a supervised and a semi-supervised approach for resolving affiliations: a naive Bayes classifier to assign affiliations to existing sets of institutional references as well as a mix of soft clustering and Bayesian learning for grouping similar affiliations. Rimmert et al. [43] assign all German affiliations to their top-level institutions and discuss most

problems that we experience, such as overly general representations, affiliated institutions with uncredited top-level and unobserved top-level organizations. Essentially, they reconstruct the German institution hierarchy and define a pipeline of normalization and assignment rules to map affiliations to top-level nodes. This requires considerable manual effort, but enables maximum performance: The result of their method was thoroughly evaluated in the follow-up work by Donner et al. [19] and can be considered close to perfect.

3 Preliminaries

An institution is a real-world organizational entity dedicated to fulfilling some more or less steady and well-defined function, in our case the purpose of higher education. An affiliation is a reference to an institution, commonly associated with an author on a scientific publication. An institutional hierarchy encodes the organizational reporting structure and/or part of relation. For example, a department and its employees (including the professor) might be part of a faculty (and report to its dean), which in turn is part of a university (managed by a rector). Lower-level institutions are subordinate to higher levels. Normally, a hierarchy is considered a tree, i.e. any lower-level institution is directly subordinate to exactly one higher-level institution. However, constructs such as interdisciplinary research groups might in fact be subordinate to multiple higher-level institutions. For the set of all higher education institutions, there is no such thing as one overarching top-level entity. Instead, they are partitioned into independent top-level institutions (mostly universities) that each have their own hierarchical structure, which depends on the institution's focus, its country's regulations, cultural conventions, etc.

Hierarchical affiliation resolution The task of hierarchical affiliation resolution entails ordering two affiliations a, b such that if a, b refer to institutions A, B, respectively, and $A < B$ in the hierarchy, then $a < b$ in the solution. Likewise, if $A = B$, then a and b should be assigned the same hierarchical node. This means that every institution referred to by at least one affiliation has exactly one node in the solution hierarchy. In addition, it is possible to interpolate unobserved nodes. For example, a faculty might never be referenced directly, but is implicit in the reference to a chair subordinate to it. Irrespective the differences in the target hierarchies and independent of a particular approach for finding them, we can determine the following aspects inherent to the solution of hierarchical affiliation resolution: (a) the top-level institutions, (b) the mapping between affiliations and their top-level institution(s), more specifically (c) the individual lower-level institutions referred to by the affiliations, (d) the mapping between affiliations and the latter and (e) the hierarchical ordering of lower-level institutions. This corresponds to the five subtasks from Fig. 1:

T1 Discover top-level institutions.
T2 Assign affiliations to a given top-level institution.
T3 Discover institutions at all levels.
T4 Assign to any institution in the hierarchy.
T5 Order institutions hierarchically.

In our work, we assume affiliations as input. Other works concerning top-level resolution sometimes also require the institutions to be known and confine themselves to solving the task T2 of assigning affiliations to top-level institutions as illustrated in Fig. 1c. Thus, they exclude tasks T1, T3, T4 and T5 (Fig. 1b, d, e, f). The steps illustrated in Fig. 1b, c can be reversed in order by clustering affiliations first and then distilling from each cluster the top-level institutional identity. This is done by some other works not requiring knowledge about the institutions. Should the top-level institutions be known due to the existence of some curated list or knowledge base, T1 should be skipped and the known institutions should be used. Although the *GERiT hierarchy* (as described in the next paragraph) is an exception, curated institution hierarchies are (a) rare and only available for certain countries and (b) their existence does not imply that a lower-level linking approach can do without regarding the relationships between the affiliations to be linked as does our data-based approach. Probably, in such a case the optimal solution would consider both the similarity of a known institution to some affiliations and their equivalence or hierarchical order.

Affiliations as paths of an institutional hierarchy GERiT [17] encodes the majority of the German higher education institutional hierarchy. Here, we find all hierarchy nodes labelled with an official name. These names do not make for good affiliations, since they could be used multiple times under the same or another top-level institution. As shown in Fig. 2, computational linguistics departments exists at Bielefeld as well as Saarland University (and others). However, without knowledge of other institutions, any lower-level institution can be referenced in an unambiguous way by listing all its higher-level institutions (*ancestors* in the tree), e.g. *"Computational Linguistics Group, Degree Course in Linguistics, Faculty of Linguistics and Literature, Bielefeld University"*, which uses a complete institutional 'path' to locate the referenced institution in the true hierarchy.

Hierarchies as the union of their affiliations Affiliations are references to nodes in an institutional hierarchy and allow to draw inferences regarding the latter, in particular if the institutional paths that they specify share the following properties.

1. All paths are given in the correct sequence.
2. All paths are complete in all affiliations.

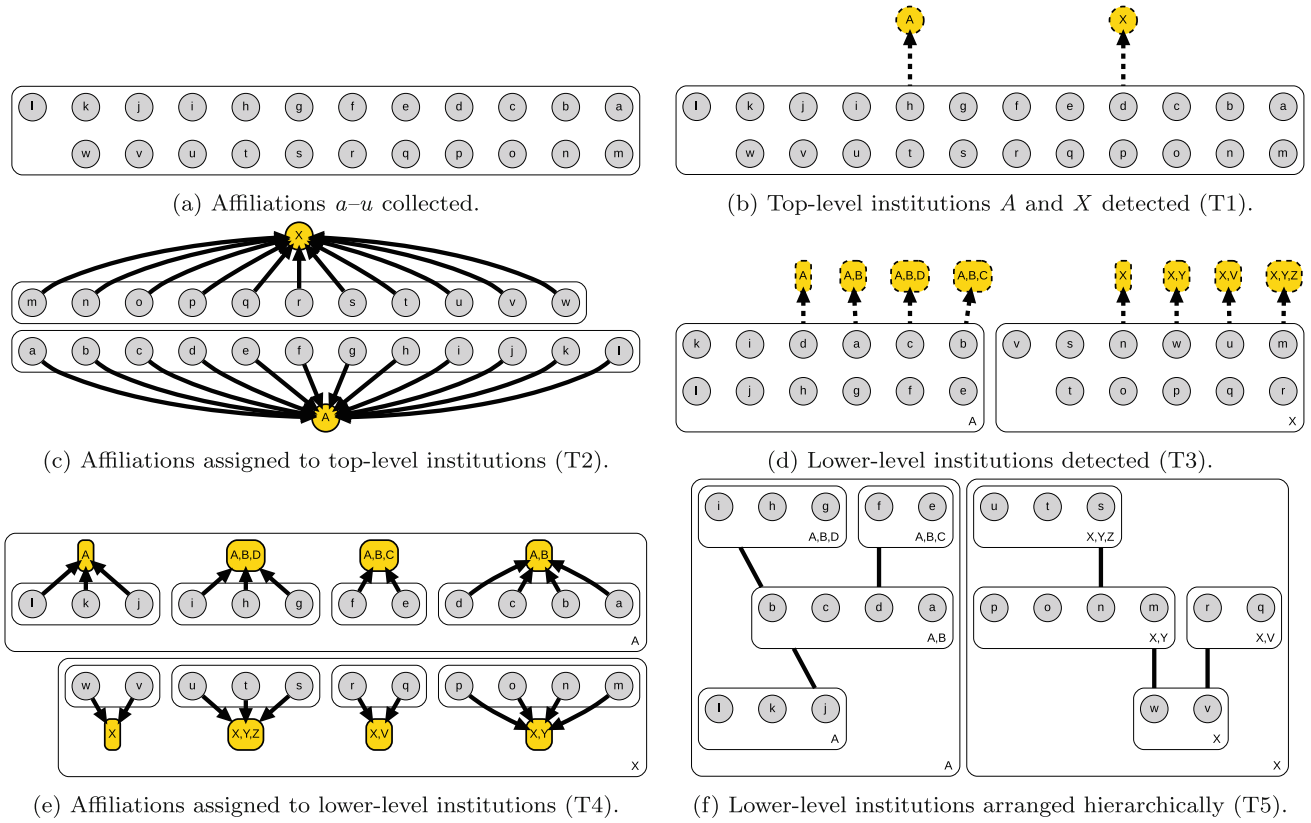
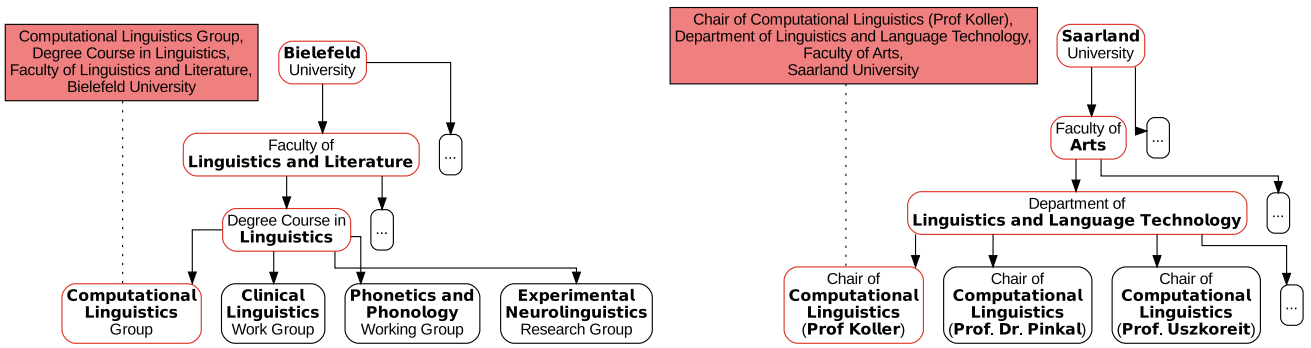


Fig. 1 Illustrating subtasks T1–T5 inherent to hierarchical affiliation resolution as iterative steps



(a) Computational Linguistics Group at Bielefeld University. (b) The three CoLi chairs at Saarland University.

Fig. 2 Two examples from the 2021 GERIT hierarchy, both referring to computational linguistics departments, but at different universities. Above in red are hypothetical perfect affiliations that list all higher-level institutions (color figure online)

3. Path elements are homogenized across affiliations.
4. Each true node is referenced by some affiliation.

In this case, the hierarchy is simply the union of all affiliations. We can relax the above requirements by assuming that the path elements need not be ordered to recover the true hierarchy if all paths are complete and there is an affiliation for each lower-level institution, because then the order of longer paths is determined by the existence of the corresponding shorter paths that lead up to them. For example, when looking at an affiliation with path nodes *A,B,C*, we might not know if $A \rightarrow B \rightarrow C$ or $A \rightarrow C \rightarrow B$, etc., but if *A* and *A,B* exists and *A,C* does not, then it is clear that the first case is correct. Assuming complete paths, *C* cannot be referenced without reporting *B*. Under these circumstances, the *subset partial order* reproduces the true hierarchy from affiliations represented as sets of path nodes. Each subset relation $\subseteq$ over a set of sets (*representations* such as A, B, C) is a partial order as it is *transitive* ($A \subseteq B \wedge B \subseteq C \Rightarrow A \subseteq C$), *reflexive* ($A \subseteq A$) and *anti-symmetric* ($A \subseteq B \Rightarrow B \nsubseteq A$).

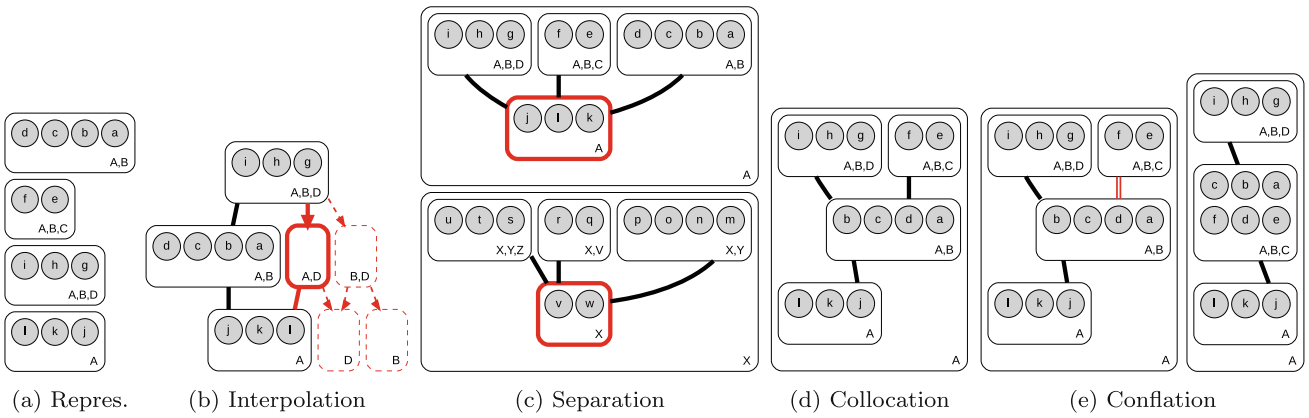


Fig. 3 Five components of our proposed framework. Grey circles a–l and m–w are affiliations, clusters A, B, etc. are representations/nodes. **a Representation** creates one feature set per affiliation. **b Interpolation** creates unobserved representations by means of generalization. Drop non-intermediate ones. In this example, the resulting nodes are all ignored later. **c Separation** finds minimal elements (in red) and/or connected components. **d Collocation** builds the Hasse diagram for each minimal element or connected component. **e Conflation** determines equivalent adjacent nodes (red edge) and merge them (right) (color figure online)

Table 1 True hierarchy paths compared to those given by affiliations

PATHS BY HIERARCHY	PATHS BY AFFILIATIONS					
	realistic	subset	subseq.	= permut.		
elements ordered	● ● ○ ○ ● ● ○ ○	○ ○	● ●	● ●	○ ○	
elements complete	● ● ● ● ○ ○ ○ ○	● ●	● ●	○ ○	○ ○	
elements same	○ ○ ○ ○ ○ ○ ○ ○	● ●	● ●	● ●	● ●	
all path available	● ○ ● ○ ● ○ ● ○	● ○	● ○	● ○	● ○	
recovers hierarchy	○ ○ ○ ○ ○ ○ ○ ○	● ○	● ●	○ ○	○ ○	

The latter are either only subsets of, only subsequences of, only permutations of, or identical to the first. If all paths are available, permutations suffice to recover the affiliations' true hierarchy

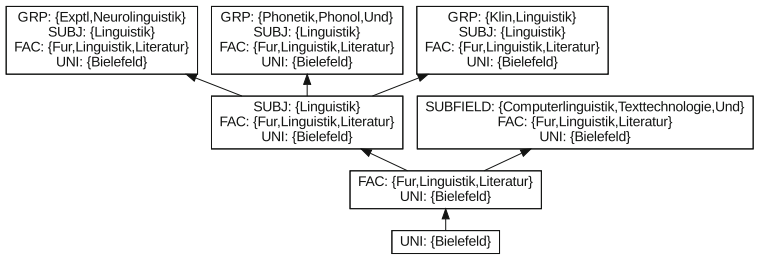
Realistic affiliations In reality, e.g. in WoS affiliation strings, it is likely that an affiliation only reports a subsequence or subset of its ancestors, e.g. *"Bielefeld University, Computational Linguistics Group"*. Further variations from the norm will happen within the node labels, as in *"Univ Bielefeld, Comp Ling"*. Finally, it is unlikely that there is an affiliation for each institution, in particular for intermediate ones like *"Faculty of Linguistics and Literature, Bielefeld University"*. In summary, given perfect affiliations (paths complete and homogenized), we can recover their hierarchy, if the paths are sorted by hierarchy level or if we have an affiliation for each node in the hierarchy (see Table 1). In this paper, we do not assume to know the path as sequence (affiliation strings are rather arbitrary in this regard) and instead represent affiliations as *sets*, which is a major simplification and allows use of the subset partial order. Therefore, efforts dealing with realistic affiliations should try and achieve the following:

1. Homogenize equivalent names/terms across paths
2. Interpolate to get affiliations for most nodes
3. Find ancestors to complete an affiliation's path

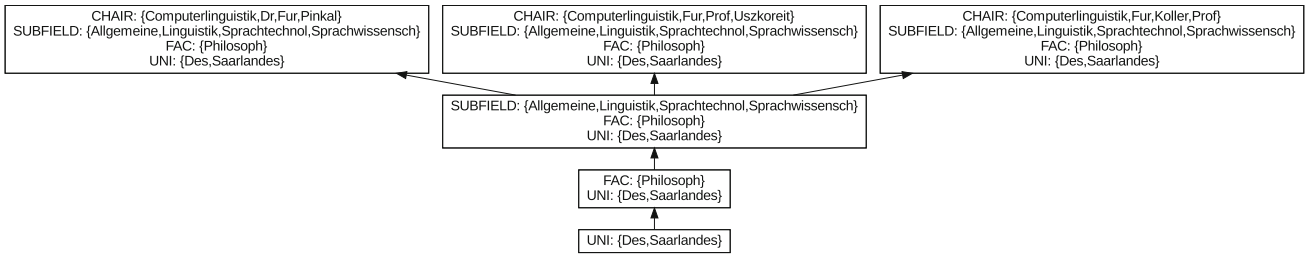
In the following, we briefly describe a framework that provides opportunities to achieve the first two means and recreates the hierarchy where possible (the first goal remains partially out of scope in this work). In Sect. 3, this framework is applied to real affiliations.

Integrated Framework Figure 3 visualizes a framework that produces institutional hierarchies from affiliation strings. With perfect affiliations, the steps in Fig. 3a, d suffice. First, we represent each affiliation as the set of nodes in the path that it describes. We call this step **Representation**. For example, *"Computational Linguistics Group, Degree Course in Linguistics, Faculty of Linguistics and Literature, Bielefeld University"* consists of *"Degree Course in Linguistics"*, *"Bielefeld University"*, *"Faculty of Linguistics and Literature"* and *"Computational Linguistics Group"*. The hierarchy is reconstructed by the *subset partial order* and visualized as the *Hasse diagram* of the partially ordered set of affiliation representations. We refer to this step as **Collocation**. If all hierarchy nodes are given as affiliations and all affiliations are perfect, then the hierarchy should be recreated perfectly in almost all cases. With real affiliations, a number of restrictions apply:

1. Not all path nodes are referenced in the same way, e.g. "Comp Ling Grp" vs. "Computerlinguistik."



(a) This version not yet reflects that computational linguistics has been moved below linguistics as of 2021.



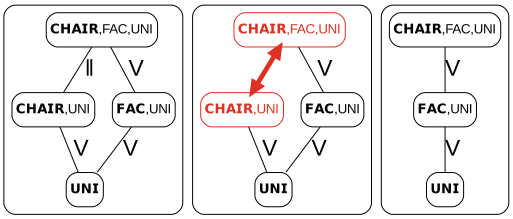
(b) The subfield name is still different from Figure 2 due to a recent merger not being fully covered in this GERIT version.

Fig. 4 Two examples from Fig. 2, here obtained through the subset partial order over the representations obtained from complete affiliation strings similar to the ones given in Fig. 2. The GERIT dataset used here is from 2019, which explains why some institutional merges visible in Fig. 2 have not been realized yet

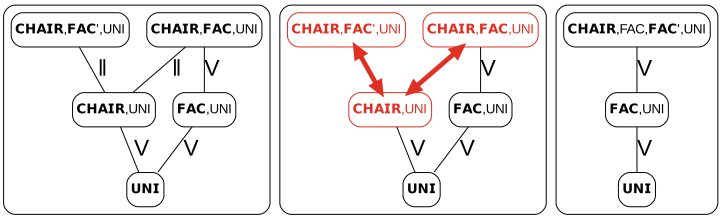
2. Not all hierarchy nodes are referenced by affiliations, e.g. faculties are rarely referenced as such.
3. Not all paths specified by affiliations are complete, e.g. intermediate levels like faculty are often ignored.

The first point can be addressed during the representation step. As many similarities among equivalent but different affiliation strings are reflected in a high term overlap, in this work, we experiment with representing each affiliation not simply by the set of (normalized) path nodes it references, but also split each string expressing one such node into multiple terms. At the same time, we try to label these terms by the institutional function that the respective node fulfills—e.g. *("Comp,"group)* and *("Ling," group)*—meaning all lower levels in the hierarchy need to share all of these terms (see Fig. 4). This mixes descriptive overlap with hierarchical overlap and also requires conflation. Any normalization is appreciated when turning affiliations into representations. The second point can be addressed by **Interpolation**. Here, we guess intermediate affiliations from observed affiliations, e.g. *"Faculty of Linguistics and Literature, Bielefeld University"* from *"Computational Linguistics Group, Faculty of Linguistics and Literature, Bielefeld University"*. The labels for institutional functions are essential at this point because they tell us which elements of an affiliation representation can be dropped to obtain higher-level institutional nodes. The third point can to some extend be addressed by **Conflation**: given two adjacent representations in the induced hierarchy, decide if they are equivalent or not. As shown in Fig. 5, functional labels like *FAC* or *CHAIR* can be used to identify equivalent affiliations. For example, in Fig. 5a, *CHAIR,FAC,UNI* specifies *CHAIR,UNI*, but we know that faculty is higher-level than chair, so they must be equivalent. However, if there is simply no affiliation that associates a certain *CHAIR* with a certain *FAC*, then conflation cannot possibly solve our problem, and external resources are required. This we do not investigate in our work. Over perfect affiliations, the subset partial order will produce individual hierarchies for each top-level institution, thus solving top-level resolution on the fly. However, for efficiency reasons it is desirable to first separate all affiliations by their top-level institution after the representation step and then apply the remaining steps for each of these separately. This we refer to as **Separation**. In summary, the following steps are applied:

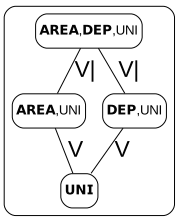
Representation	Discover and tokenize all parts of the affiliation string that fulfill an institutional function and label them correspondingly.
Interpolation	By iteratively removing the elements of a representation with the lowest-level label, interpolate unobserved intermediate institutional levels.
Separation	Like in ER *blocking*, reduce complexity by partitioning the set of affiliations as separately processed top-level institutions.
Collocation	Order the representations of different affiliations in the subset partial order to reveal their original hierarchical relationships.
Conflation	Merge representations that are in a hierarchical relationship after collocation, but



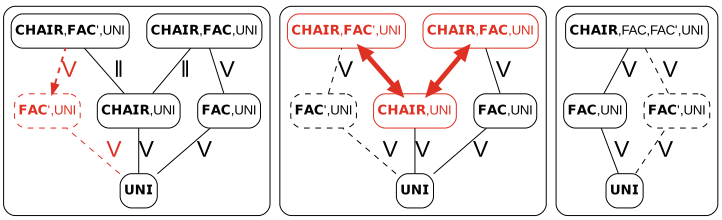
(a) Under *UNI*, one affiliation does not report *FAC* for *CHAIR*, while another does. As *CHAIR,FAC,UNI* cannot refer to a lower level than *CHAIR,UNI*, one can assume the two affiliations to be co-referring and thus merge their nodes.



(b) The situation as described under 5a, but with another faculty *FAC'* associated with the same *CHAIR*. Thus the result of conflating all affiliations of *CHAIR* references both *FAC* and *FAC'*, although only subordinate to *FAC,UNI*. This is unfortunate as we might interpret the case as one where a chair has two faculties above it.

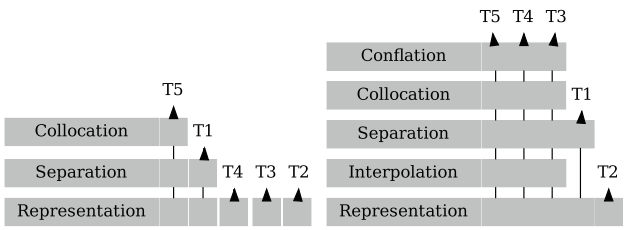


(c) Here it remains unclear if *AREA* should be above *DEP* or the other way round.



(d) If we interpolate *FAC',UNI* from *CHAIR,FAC',UNI*, the procedure from 5b returns a hierarchy with *FAC* and *FAC'* above *CHAIR*.

Fig. 5 Four examples for simple logic-based operations to improve an incomplete institutional hierarchy



(a) For perfect affiliations. (b) For realistic affiliations.

Fig. 6 Framework components and tasks: a task's arrow crosses the components required for its fulfillment

are actually refer to the same institutional branch.

Figure 6 shows how they fulfill tasks T1–T5.

Some mathematical terms During representation, we turn a set of affiliations into a set of sets of attribute-value pairs. The *subset/superset partial order* over this set can be drawn as a *directed acyclic graph* that has an edge for each subset/superset relationship. Its *transitive reduction* corresponds to a *Hasse diagram* for the set of affiliation representations. During conflation, we *contract edges* between nodes deemed equivalent, which returns a *graph minor*. During separation, we compute *connected components* of the graph, creating a *partitioning* of the affiliation representations, where each component has one or more *minimal elements* that correspond to the respective top-level institution(s).

4 Method

In the above preliminaries, we have briefly sketched the fundamental concepts of our approach. There are five main subtasks T1–T5 and five modular components in our framework. Their relationship is depicted in Fig. 6. Assuming perfect affiliations, assigning affiliations to a *given* top-level institution (**T2**) is the easiest task, because we simply need **representation** by feature sets to select all supersets of the top-level institution's representation as belonging to it (cf. Fig. 4). This also discovers all lower-level institutions (**T3**) because each unique representation corresponds to one. Not all affiliation representations, however, describe a top-level institution. Identifying those (**T1**) is achieved through **separation** in terms of minimal element search, which also speeds up T2. The superset-approach also works to assign affiliations to lower-level institutions (**T4**) if their true representation is known. **Collocation** essentially precomputes the superset relations and thereby speeds up T4 in addition to creating the hierarchical ordering of affiliations (**T5**). With realistic affiliations, we add **interpolation** and **conflation**. Interpolation adds additional lower-level institutions, so it changes the result of T3. Conflation merges equivalent adjacent nodes, thereby extending the representation of some nodes (T3) and their assignment in T4 (e.g. in Fig. 5a, *CHAIR,UNI* becomes *CHAIR,FAC,UNI*, so now it becomes a superset of *FAC,UNI*). This can also change the ordering of corresponding affiliations (T5).

In the following, we attempt to bridge the gap between the theoretical assumptions and the reality as presented

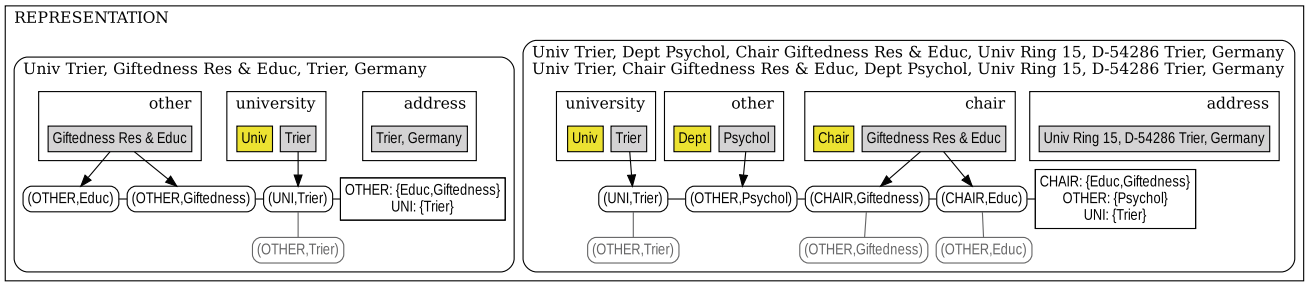


Fig. 7 Representing affiliations: grey boxes are labelled substrings, yellow boxes standardized labels. In the next row, individual attribute-value pairs are derived based on the labels and extracted terms. Bottom row features are added for generalization but not shown in the displayed node labels shown in the bottom-right corners

by the 'dirty' affiliation strings from the WoS. As stated above, these efforts focus on homogenizing affiliation strings, interpolating intermediate hierarchies and finding equivalent representations adjacent in the postulated hierarchy. For each of our framework's modular components, we discuss its goal, general functioning, practical challenges and a baseline implementation.

4.1 Representation

The goal of the representation step is to create a set of features for each affiliation string. The representation step satisfies the requirements for task **T2** *Assign affiliations to a given top-level institution*, as one can define a linking candidate representation—i.e. $UNI : \{Bielefeld\}$—and return all its supersets.

General Functioning The general representation parsing approach is depicted in Fig. 7. Each part (e.g. *"Chair Giftedness Res & Educ"*) of the institutional path (e.g. *"Univ Trier, Dept Psychol, Chair Giftedness Res & Educ"*) described by the affiliation should be separated from the other parts (e.g. from *"Dept Psychol"*). With perfect affiliations, this would be sufficient. To account for realistic affiliations, we make two additional assumptions: First, we also split each part into its terms, e.g. *"Giftedness"* and *"Educ"* (*"Res"* is dropped like few other terms and stopwords with little to no meaning). Second, we label each part by the institutional function it seems to fulfill (e.g. *(chair, "Giftedness")* and *(chair, "Educ")*. The parsing process includes determining boundaries between segments of the string (e.g. comma) and between words (e.g. whitespace) and finding function-indicating terms to label each part (e.g. "Group" triggers a labelling by *community*). The parser's ability to generate expressive, clean and normalized representations from affiliation strings has great impact on downstream performance. As a further step of addressing realistic affiliations, terms can be normalized to resolve equivalent variations like typos.

Table 2 Component labels with hierarchy levels and example triggering keywords

label	lvl	key	label	lvl	key
unknown	0	–	university	0	Univ
association	1	Assoc	academy	1	Acad
college	1	College	clinic	1	Klin
faculty	1	Fak	centre	1	Ctr
agency	1	Amt	factory	1	Werk
company	1	Ltd	site	1	Campus
field	1	Sect	institute	2	Seminar
lab	2	Anstalt	collection	2	Archive
subfield	3	Area	other	4	Dept
community	4	Group	chair	4	Ls
subject	4	Fach			

Practical challenges Affiliation strings in the WoS are heterogeneous where distinct surface forms often represent the same concept. Prominent variations are:

a Different languages like *gesellschaft/society*
b Abbreviations like *society/soc*
c Acronyms like *rwth/rhein-westfal-hs*
d OCR mistakes like *Diisseldorf/Düsseldorf*
e Typos like *Wesfalen/Westfalen*
f Umlaut handling like *Tuebingen/Tubingen*
g Flectional forms like *Energien/Energie*

In addition, there is great ambiguity around the description of institutional functions, for example, *chair*, *department*, *institute* and *group* might in some cases all refer to the same hierarchy node, while in others describe essential differences.

Baseline implementation As shown in Fig. 8, our pipeline obtains representations from affiliations by

(i) Splitting a string on comma
(ii) Finding address components by regular expressions
(iii) Classifying the remaining components

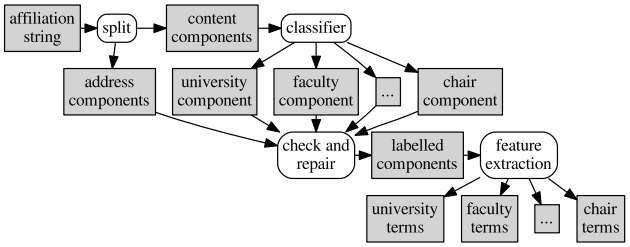


Fig. 8 Parsing affiliations into representations. Starting in the top-left corner, grey boxes represent in- and output, rounded boxes represent processing steps. The result is a labelled set of terms from the affiliation string

(iv) Extracting terms for each component
(v) Normalizing terms

Most component boundaries can be detected by comma. Address components can be detected by regular expressions. Almost all components contain some keyword that indicates their label (see Table 2). By browsing the WoS affiliation strings and repeatedly viewing still unlabelled components, we have manually created a list with over a hundred keywords of the kind $Fachhsch \Rightarrow university[FH]$ meaning "upon *Fachhsch*, replace by *FH* and use as *UNI*." This means that the rules trigger upon exact string match of the left-hand side. These rules cover a large part of the potential hierarchical substructures, but can be easily extended or learnt automatically in the future. Over the labels in this list, a total order is defined that indicates which labels cannot be above others in a hierarchy. We allow suffix-matching for compounds like Krebs*zentrum*. This hierarchy is also used to chose among multiple keywords in the same component, e.g. if "Univ" and "Hosp" co-occur, classify as *hospital* because it is more specific. We have noticed that departments (e.g. *Dept*) can refer to subdivisions on practically all levels of the hierarchy and are often synonymous to other, more precise descriptors like *faculty*. Therefore, we label these components by *OTHER*, like those without identified keywords. In fact, all terms are labelled *OTHER* in addition to their detected label. Thereby, label information complements term information. In visualizations, we only display *OTHER* terms without any label. For each label, we sort all terms alphabetically and compare each term w_i with all other terms w_j within window-size 20, computing (a) *Damerau distance* $d(w_i, w_j)$ and (b) length of *longest common prefix* $l(w_i, w_j)$. If

$$\frac{d(w_i, w_j)}{\max(|w_i|, |w_j|)} < \frac{1}{8} \text{ or } \frac{l(w_i, w_j)}{\max(|w_i|, |w_j|)} > \frac{4}{5}$$

we consider terms equivalent, get the transitive closure and replace all terms in each equivalence class by the most frequent term. Then we repeat until convergence.

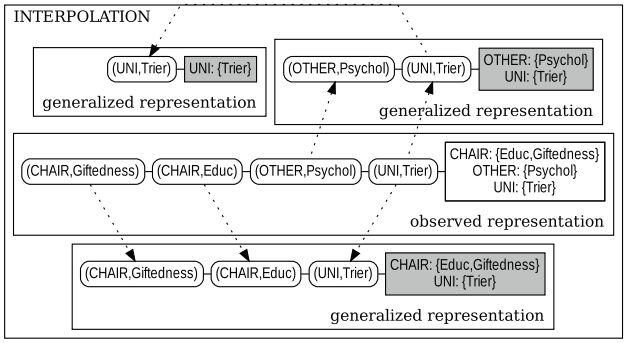


Fig. 9 Creating interpolations by removing the most specific elements. Grey box: unobserved interpolations. Arrows show which features are carried over. The middle row has two possible direct generalizations

4.2 Interpolation

The goal of interpolation is to infer intermediate levels in institutional hierarchies that are not observed in the sense that there is no representation that corresponds to them exactly, while there might be representations that refer to a specification. Interpolation is not really required for any of the tasks, but can help to improve the results for tasks **T3–5**.

General functioning An affiliation representation is generalized by removing its most specific elements. Given *Chair of Giftedness Research and Education, Psychological Department, University Trier* as in Fig. 9, we can infer that there also exists *University Trier, Psychological Department* as well as *Chair of Giftedness Research and Education, University Trier* and just *University Trier*—that is if we do not know whether *Psychological Department* is above or below *Chair of Giftedness Research and Education*. This is a main use of labelling terms with hierarchical functions because these can be used to determine the most specific part of the institutional path referenced by an affiliation.

Practical challenges As the challenge of interpolation is to decide which elements of an affiliation's representation to remove, problems arise when the information used for making this decision is unreliable. We have already stated above that the same institutional function is often described by different words. In addition, depending on the top-level institution, the official functions might not always be arranged in the same hierarchical way. For example, in Germany, sometimes a *Fachbereich* is equivalent to a *Fakultät*, sometimes it is one level above the latter. Therefore, the institutional function label, even if properly detected, is not always sufficient to know which is the most specific part of an affiliation representation.

Baseline implementation The question of finding generalizations of observed affiliation representations that actually correspond to higher level hierarchy nodes is about deciding which elements of the representation to remove. For

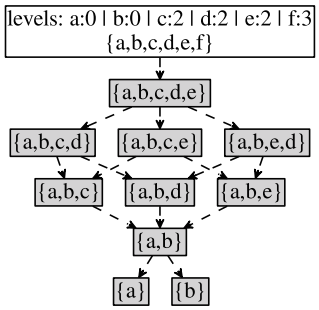


Fig. 10 Interpolation by rule-based recursive generalization through repeated removal of the most specific features. Top to bottom. Order of removal is based on the hardcoded level of labels (e.g. f is removed first)

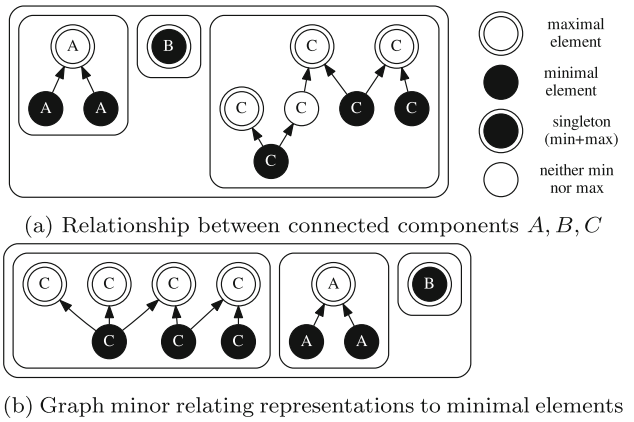


Fig. 11 Minimal and maximal elements in the subset relation. The crucial point is the connectivity, which is the same in the above subset partial order and the below graph minor based on minimal elements (minels). Separation by connected components reduces downstream computation complexity. The graph minor is built directly from the data in order to save memory by skipping the intermediate edges in the top before computing the connected components

example, we might have some affiliation UNI : $\{x\}$, FAC : $\{y\}$, CHAIR : $\{z\}$ from which we can infer there also is an institution UNI : $\{x\}$, FAC : $\{y\}$. We look at each observed representation separately and interpolate its superior levels as subsets. To find the right subsets, we use our total label order (Table 2) where we encode which labels cannot refer to a higher level than others. This list of hierarchical functions was manually determined while browsing international affiliation strings in the WoS. If a level number is lower than another, then its component is assumed impossible to refer to a more specific hierarchical level than the latter's. For example, we assume that *centre* cannot be above *university* ($1 > 0$), subfield cannot be more specific than group. Judging from that label order, for each representation, we remove all features with the most specific label and repeat recursively as in Fig. 10. If we do not know which is the most specific label, we stop. If an interpolated representation is also *observed*, then the interpolation vanishes. In practice the number of remaining interpolations is limited.

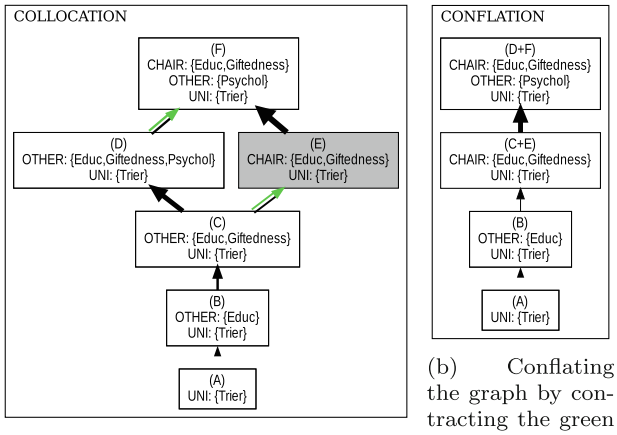
4.3 Separation

In Separation, we aim at segregating representations that have no relationships. On the one hand, this has computational benefits as it is more convenient to apply subsequent steps onto a number of smaller subsets rather than the entire data. On the other hand, it performs top-level resolution. The separation step satisfies the requirements for (**T1**) *Discover top-level institutions*, as optimally, each minimal element corresponds to one top-level institution.

General functioning The concept of *weakly connected components* in a directed acyclic graph (i.e. our hypothesized hierarchy) lends itself to partitioning, creating independent institutional subgraphs that are all somehow related internally, while unrelated to any other subgraph. With perfect affiliations, each such connected component corresponds to the institutional hierarchy of one (or multiple related) top-level institutions(s). The top-level institutions itself are identified as minimal elements in the subset partial order, that is affiliation representations of which no other representation is a subset. A connected component might have multiple minimal elements (top-level institutions) if they are somehow connected by a shared lower-level institution (for example, a joint research centre). Figure 11 illustrates the relationship between minimal elements (in black) and connected components.

Practical challenges With real affiliations, overly general representations (e.g. *{(university, "Tech")}*) become a problem. They result from parsing errors or insufficient affiliations and create incorrect minimal elements connecting separate institutional components. Another problem is the possibility of legitimate common lower-level connections between many institutions. It is not unlikely that any top-level institution shares at least one lower-level node with another top-level, so that a 'chain' of components evolves and ultimately almost all institutions are somehow connected.

Baseline implementation As shown in Fig. 11, mapping representations to their minimal elements does not change the connectivity of the corresponding graph. Its connected components remain the same. While a minimal element can only be in one connected component, each representation—and thereby each component—can have multiple minimal elements. Therefore, if we view each minimal element separately, some affiliations will show up in more than one view. This might be useful as we have discussed earlier that through examples like joint research centres, almost all top-level institutions might be somehow connected. Once the mapping between affiliation representations and their minimal elements has been computed, one can freely chose



(a) Collocation in the subset partial order. Arrows show the cover-relation. Green edges are determined contractable

(b) Conflating the graph by contracting the green edges. Deciding *which* edges are green is crucial

Fig. 12 Collocation and conflation in the same graph subset. On the left, collocation manifests in the partial-order arrangement of the representations displayed as nodes. On the right, the result of conflation is shown

between separating by connected components or by using minimal elements to obtain overlapping subsets corresponding to individual top-level institutions. We have developed an efficient algorithm that allows parallellized detection of connected components in the subset partial order. We abstain from describing it here in detail. In contrast with most other approaches, it does not require the built partial order, but the underlying sets as input—in this case our affiliation representations. It distributes the search space over multiple iterations and parallel processes to discover subset relations and ultimately all *minimal elements* of all representations.

4.4 Collocation

The goal of collocation is to hierarchically order all lower-level institutions under the same top-level based on their representations. Despite the common assumption of a hierarchy being a tree, both the true and the hypothesized hierarchy might in fact be more general, i.e. directed acyclic graphs, because some lower-level institutions could have multiple 'parents.' Collocation (partially) solves (**T5**) *Order affiliations hierarchically*. Further, it enables merging equivalent representations through conflation, which affects T3 and T4.

General functioning Given perfect affiliations in sufficient numbers, the institutional hierarchy can be reconstructed by applying the subset/superset partial order over their representations. The same principle is applied to imperfect affiliation representations, so that shortcomings will be fully observable in the result. We do not attempt to change anything about the partial ordering since it is the most central principle of our approach. A first example is given in Fig. 12a. Here, any two representations are related that are in the subset/superset relation. The partial order can be visualized as a Hasse diagram or directed acyclic graph. This does not display the entire partial order but its *transitive reduction*, so that given $a \rightarrow b \rightarrow c$, we omit $a \rightarrow c$. If tree-shaped hierarchies are strictly required, this can be enforced by computing a *spanning tree*.

4.5 Conflation

Despite all efforts to achieve clean feature-sets during the representation step, differences in affiliation representations can be the result of redundancy, which is caused by variation and incompleteness. Often, incomplete representations appear as generalizations in the subset partial order, or variations correspond to different specifications of the same parent. The goal of conflation is to classify affiliation representations that are adjacent in the hypothesized hierarchy as to whether they are equivalent or not. This means deciding whether the additional features of the superset describe a hierarchical specification or only some additional information regarding the same lower-level institution. The iterative application of this step can potentially integrate missing information into incomplete representations, thus not only reveals incorrect dominance that is actually equivalence (regarding **T5**), but also improves the set of assumed lower-level institutions (**T3**) and their assignment to affiliations (**T4**).

General functioning In general, any binary classifier can be applied to the task. Since it is only applied on representations that are adjacent in the hypothesized hierarchy, the computational complexity of this pairwise comparison is limited. We assume that the institutional function labels like *chair* or *department* are crucial because any real dominance relation between two affiliation representations requires for the superset to contain information about an additional lower-level function label. For example, *uni, faculty* might be specified by *uni, faculty, chair*. However, we know that not all additional labels are more specific. For example, *uni, faculty, chair* only completes *uni, chair* because *faculty* must be above chair.

Practical challenges Unfortunately, there is great variation in the use of hierarchical function words. The same lower-level institution might be referred to as *department*, *chair*, *group*, etc. Therefore, it is difficult to decide the hierarchical order of those types. In addition, even the official labels might be used differently across top-level institutions. As a result, it can be difficult to distinguish 'legitimate' supersets that describe actual hierarchy and those that only add some details to the same level. For example, in Fig. 12a we have (C) $\{Educ, Giftedness\}$ once as some underspecified level *OTHER* and (E) once as *CHAIR*. As we label all terms not only by their detected label but also with *OTHER*, (E) is a superset of (C). Still, both actually refer to the same level in the institutional hierarchy. This is not as clear for (D) and

in fact their common specification (F) suggests that *Psychol* does not refer to the chair, but to a different level. The conflation result for this example is shown in Fig. 12b.

Baseline implementation With perfect labelling, it is possible to simply merge all representations that are in the subset partial order and have the same set of attributes. In practice, this is not as easy and more conservative rules as applied in the above example are recommended. Conflation creates a graph minor by edge contraction. We want to contract edges that do not correspond to actual hierarchical relationships. These 'redundant' edges are the result of a representation being a superset of another, but none of the features in the set difference describing a new hierarchy level, e.g.

$$\text{UNI} : \{\text{Heidelberg}\} \rightarrow \text{UNI} : \{\text{Karl, Ruprecht, Heidelberg}\}$$

If all representations have unambiguously labelled terms, an aggressive, but sound method is to contract all edges between nodes with the same set of *attributes/labels*. This does *not* merge all representations with the same set of labels since most of them do not have any edge between them. It means that a true hierarchical specification must introduce a new label, for example, some

$$\text{UNI} : \{x\}, \text{FAC} : \{y\}$$

cannot have another $\text{UNI} : \{x\}, \text{FAC} : \{j\}$ under it, but

$$\text{UNI} : \{x\}, \text{FAC} : \{y\}, \text{CHAIR}, \{z\}$$

would be possible for a number of distinct z. This fails occasionally due to overly general representations or ambiguous labels and merges unrelated representations like UNI : {Heidelberg}, OTHER : {Comp} vs. UNI : {Heidelberg}, OTHER : {Comp, Ling}. We use a conservative method that contracts all edges between representations with the same set of *values/ terms*, regardless of their labels. Both conflation methods have the convenient property that they are order-invariant, in that the order of contractions does not make a difference to the final result.

5 Experimental evaluation

For two reasons, evaluating (hierarchical) affiliation resolution is not trivial. First, there is no large-scale hierarchical gold dataset where affiliation strings with realistic variations and mistakes are assigned to individual nodes. Second, there are multiple possible solutions for comparing the hypothesized hierarchy to a gold hierarchy. Figure 13 shows our evaluation framework. Top-level resolution as a byproduct of separation and top-level linking as realized through representation can be evaluated on a gold labelling of separate top-level institutions, which is easier to obtain.

5.1 Data

We deploy the Web of Science corpus with $58M$ publication metadata records from 1980–2019 and $230M$ author mentions with $95M$ affiliations. We have identified two major benchmarks that can act as gold standards:

a The top-level institution assignment to all German WoS affiliation strings provided by [19,43], which resolves $6.5M$ affiliations to $2K$ top-level institutions
b The institutional hierarchies (*GERiT*) of all institutions that have ever applied for funding with the DFG (German Research Foundation), which is practically the entire German research landscape

The latter contains 1978 top-level German research organizations with 29,196 sub-organizations over seven hierarchy levels. The GERiT gold hierarchies are not directly applicable, as they order not (synonymous) affiliations but only map a single name to each node (e.g. *Subject Hydrology*). Therefore, we have to map 'dirty' WoS affiliation strings to the nodes in the gold hierarchy. For this annotation task, we apply an automatic retrieval-based linking method that suggests the most likely matches, which are then verified or rejected manually. We have thus processed all of the suggested matches for University Trier as well as 1000 for University Heidelberg and University Bonn. An example subset of the annotation result is depicted in Fig. 14, which shows for each node the GERiT label on top and below the manually assigned WoS affiliation strings.

5.2 Evaluation objectives

The contribution of our work is not an overall solution for the problem, but rather a first analysis of the interactions between a large collection of realistic affiliation strings, a curated hierarchy and first baseline methods. Hence, a number of different experiments provide different pieces to the puzzle introduced by the new task. These aspects contribute to a basic impression:

1. **General functioning of proposed framework components** (proof of concept)
2. **Adequacy of baseline implementations** and **selective assessment** of the returned hierarchies
3. **Scalability** of the framework components
4. Insights about the **realistic affiliations** and their **relationship to true lower-level institutions**

These aspects are assessed by objective evaluation:

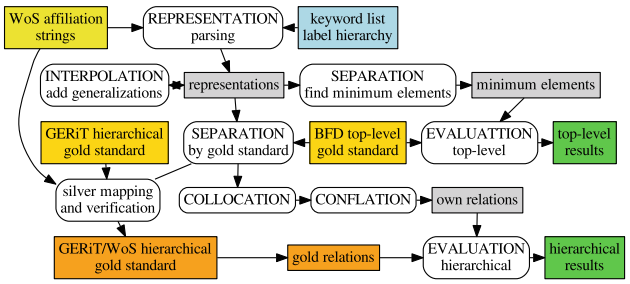


Fig. 13 Evaluation framework using hierarchical- (GERiT [17]) and top-level- (BFD/"Bielefeld" [43]) gold standards. Starting in top-left corner. Yellow: input; blue: parameters; golden: gold standard; orange: combined gold standard; green: results (color figure online)

5. **Top-level resolution** by separation component
   Grouping affiliations by minimal elements
6. **Top-level linking** by superset query
   Grouping affiliations by known minimal elements
7. **Lower-level resolution**
   Discovering equivalence between affiliations
8. **Hierarchical resolution**
   Discovering hierarchical relations between affiliations

Objectives 1–4 are addressed informally by inspecting the processes and their outputs. Questions 5–8 are measured by the experiments described next.

5.3 Experiments

We run three major experiments:

1. **Top-level resolution**: compare a top-level gold standard against grouping of affiliations under the same minimum element
2. **Top-level linking**: compare the set of all affiliations with representations that are supersets of a top-level institution's canonical representation against the set of affiliations assigned by the gold standard
3. **Lower-level resolution** and **Hierarchical resolution**: evaluate the result of the entire framework against a hierarchical gold standard

In the first experiment, we create a mapping between the German WoS affiliations and their minimal elements using the superset partial order of their representations. For any pair of affiliations, we check if it shares one or more minimal elements (counting towards P, the positive pairs) and if it has the same Bielefeld gold label (counting towards T, the true pairs). If both apply, then it also counts towards TP, the true-positive pairs.

In the second experiment, we use UNI : {Trier}, UNI : {Bonn} and UNI : {Heidelberg} as representations for three top-level institutions. For each, we take the number of affiliations that are represented by supersets as P, thereof the number of affiliations with the respective gold identifier as TP and the overall number of affiliations with the gold identifier as T.

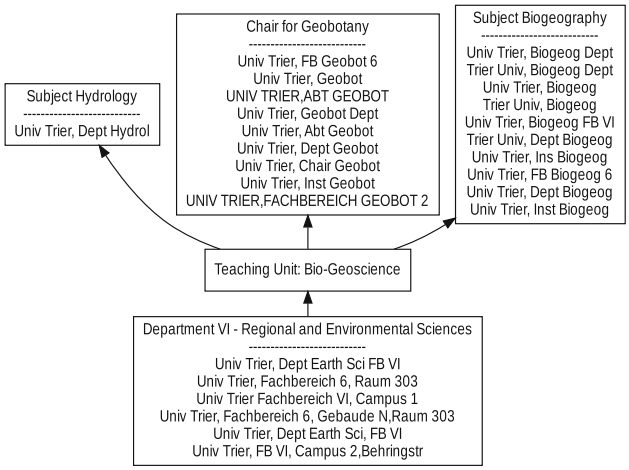


Fig. 14 A small subset of the hierarchical gold standard. In each node the GERiT label is given above the line. Below it, the WoS affiliation strings are listed that have been manually attached to it. This combination of the GERiT hierarchical gold standard and the WoS affiliation strings achieves a hierarchical gold standard over realistic "dirty" affiliations

For the third experiment, we first separate the data by the known top-level institutions. For each of the three annotated top-level institutions, we apply our framework to order the annotated affiliations, producing results like that in Fig. 15. We then compare the hierarchical relation between two affiliations ($=$ or $<$) computed by our method (Fig. 15) against that annotated in the hierarchical gold standard (Fig. 14). Each of the two relations has their own P, T and TP. Looking only at equivalence ($=$) evaluates lower-level resolution as it assigns affiliations to lower-level institutions without sorting the latter hierarchically.

5.4 Evaluation measures

As stated above, we implement the comparison of two hierarchies by converting them into mathematical relations (i.e. sets of pairs). The result can be compared in terms of set overlap, which translates well to common measures like precision and recall. For evaluation, we use Precision and Recall computed as $\frac{TP}{P}$ and $\frac{TP}{T}$, respectively, where TP is the number of true-positives, P that of positives and T that of true pairs. TP is the number of pairs that are both true and positive. In top-level evaluation, a true pair is a pair of affiliations that refers to the same top-level institution according to the gold standard. A positive pair is one that our method *claims* to be coreferring, i.e. that shares at least one minimal element. In hierarchical evaluation, we compare two hierarchies, each of which is encoded by the subset relation $<$ over the men-

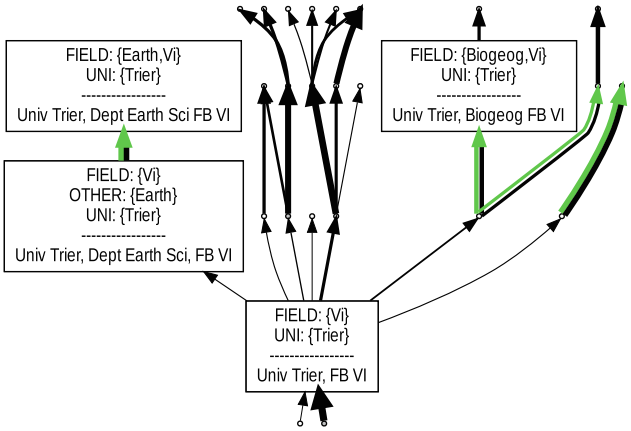


Fig. 15 Some annotated affiliation strings from University Trier in the hierarchy built over their representations. Each node label features the representation (above line) and corresponding affiliation (below). Small circles depict nodes without linked annotated affiliations. Green edges are determined contractable (color figure online)

tions' representations in the hierarchy and the equivalence relation $=$. Then we compare $<$ and $=$ of our hierarchy as positive pairs against $<$ and $=$ of the gold hierarchy as true pairs. Perfect results can only be achieved with the annotated representations ordered exactly as in the gold hierarchy.

6 Results

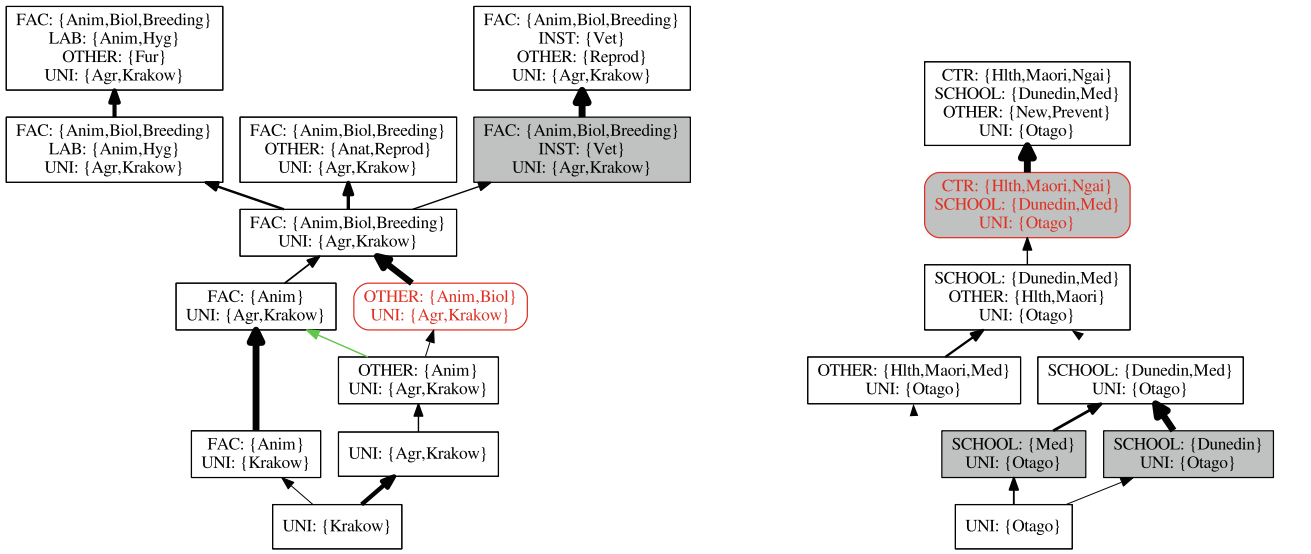
In this section, we observe the results of applying our framework instantiated with the described baseline methods to the WoS affiliations. We do so by answering the eight research questions listed in the previous section, which are separated into four questions of basic observations and general impressions on the one hand as well as four objective evaluation scenarios on the other.

6.1 Basic observations and general impressions

General functioning of proposed framework components All the framework components have been successfully implemented with relatively simple baseline methods. Each component is realized as a separate program. The representation step takes affiliation strings as input and outputs a set of attribute-value pairs for each string. The interpolation step takes representations as input and tries to output other representations that correspond to ancestor institutions. The separation step takes representations as input and outputs a mapping between representations and their minimal elements. This can also produce the subset/superset partial order on the fly. The collocation step can be run for any subset of the representations, for example, for all supersets of a minimal element (top-level institution). The results are displayed as graphs. Figure 16 shows two subgraphs of the hierarchies under the (Agricultural) University of Krakow and University of Otago, respectively. The conflation step is applied directly on the graphs and is therefore more complicated to implement: when two adjacent nodes are determined equivalent, the (green) edge between them is removed and they are merged. This means that the respective nodes are replaced by a new node that has as incoming edges all edges going into any of the two nodes and likewise for outgoing edges. The corresponding representation is the union of the merged nodes' representations.

Adequacy of baseline implementations Using the baseline affiliation parser, most of the produced representations look correct. It is only through the creation of hierarchies that occasional oddities become apparent. The advantage of the interpolation step is that problematic generalizations can be avoided by only interpolating obvious cases. In many cases, there is no obvious solution (cf. Fig. 5c). The consequence is that the number of interpolated nodes is rather limited. As will be discussed under top-level resolution, the result of separation suffers from a small number of overly general minimal elements like *UNI:Tech*, which points to a problem in the representation step—or in the affiliations themselves. Detecting these cases is not trivial. These graphs are built over all affiliations from the 1980s to 2014, during which time many hierarchies or subdivision names have of course changed. Fortunately, the subset of affiliations to be ordered hierarchically can also be further broken down by custom ranges over the WoS publication years. We have experimented with a number of different conflation schemes and unfortunately all effective ones at some point overshoot, leading to a self-enforcing process of oversize nodes absorbing more and more adjacent nodes. This does not happen with our conservative conflation baseline, but then, many edges that should be contracted remain. For example: the entire graph in Fig. 17 refers to the same branch, underscoring the amount of synonymy. Hence, collocation is not at all a trivial task.

Scalability The representation step has linear complexity as it processes each affiliation independent of the others. It can be easily parallellized by splitting up the set of input affiliations. The same holds for the generalization step. Separation amounts to minimal element search, which has been subject to a number of works [9,20,33,35,39–42,45,48,49]. In practice, we have managed to deploy a parallellizable implementation that can solve the task in a matter of minutes, even with the arguably slow Python programming language. Collocation can be applied to any subsets. The graph visualizations are only practical if applied to a smaller subset. The hierarchical relations as such can already be produced by the separation step without additional costs (except storing them). Currently, conflation is only applied in the context of



(a) Selected affiliation from the Agricultural University Krakow (b) Selected affiliation from the University Otago

Fig. 16 Graphs depict the cover relation of the subset partial order over affiliation representations around a selected affiliation (in red). On the bottom is the most general representation (institutional top level) with supposed hierarchical specification towards the top. Line width corresponds to strength of association, which is currently not exploited. The green edge is determined contractable by a rule-based classifier (color figure online)

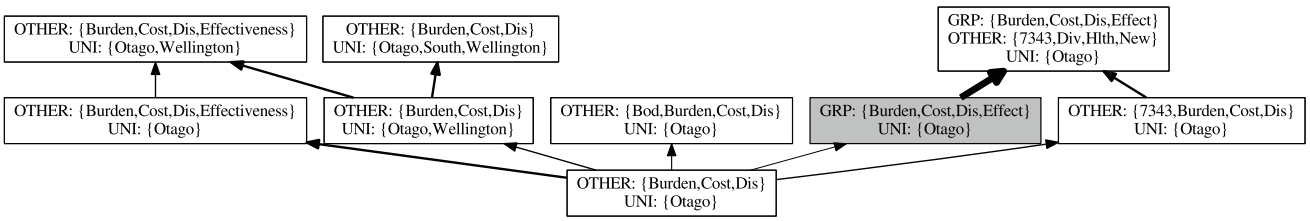


Fig. 17 An extreme case of synonymy: all edges should be contracted as all nodes refer to the same institution

the graph visualization program. However, it can be applied directly on the node pairs that make up the hierarchical relations by comparing the respective representations. Due to the generality of this task, it can be assumed that the consequences of interative edge contraction can be computed efficiently. Overall, with little optimization efforts, using around 16 cores, the entire Web of Science is processed within 24 hours, but we expect this can be reduced substantially by cutting corners.

Realistic affiliations and true lower-level institutions Our approach reveals interesting relationships between affiliation strings referring to international institutions. It can partially order and visualize any subset of the WoS affiliations. Figure 16a, b displays cases where this process uncovered hierarchical relationships, e.g. the *Faculty of Animal Breeding and Biology* at the *University of Agriculture in Krakow* has an animal hygiene lab (with some fur-related subdivision), some department of anatomy/reproduction and a veterinary institute (with a specialized reproduction subdivision). In many cases where hierarchical relations are not properly resolved, the result allows to inspect systematically the variations among equivalent lower-level institutions. For example, Fig. 17 displays 9 representations that all refer to the same lower-level institution, the ***Burden*** *of* ***Dis****ease Epidemiology, Equity and* ***Cost****-Effectiveness Programme* at the *Department of Public Health* of the *University of* ***Otago*** in *Wellington, NZ*. The following variations can be observed: The *effectiveness* part is only mentioned in some affiliations. University of Otago is sometimes specified as being the Wellington site. In one case, *'South'* is included, possibly referring to South Otago. In one case *'Bod'* refers to the acronym of the Program ($BODE^3$). In some cases, the *effectiveness* part is mentioned. In two cases the postcode was accidentally included in the representation. Finally, one affiliation even references the *Health Science Division*, which is actually above the Department of Public Health and even above the Wellington Site. We only know this from the Web-

site[1] as it is not apparent from the affiliations themselves. In Fig. 15, we can observe the effect of a missing comma between *Dept Earth Sci* and *FB VI* (left top node vs. the one below it). Fortunately, here even the conservative conflation method recognizes the representations to be equivalent. In summary, a subjective inspection of the output hierarchies suggests that they can be used well to get an overview of some of the hierarchical relationships and many of the systematic variations among equivalent affiliations. It is, however, obvious that the complete true underlying hierarchy is not clear until a reference like the official website is compared to the extracted relation. In addition, without a specific date range applied, the hierarchies also include outdated structures.

6.2 Objective evaluation

Top-level resolution Separation achieves 63% recall and 23% precision as top-level disambiguation. The main problem is overly general representations such as CLI : {Univ}. Although it is possible that some affiliation strings simply do not contain all relevant information, usually the problem can be traced back to missed or miss-classified components from the representation step. Often this happens when some address information has to be used to complete information in the rest of the string, as in *Univ Hosp, INF 672, 69120 Heidelberg*.

In Fig. 18a, we see a number of overly general minimum elements in orange that surely all create incorrect top-level relations, e.g. UNI : {Tech} or INST : {Geschichte}. None of these should have been created, whether through representation or interpolation. On the other hand, UNI : {Dresden} actually is correct but links two affiliations that are considered separate in the gold standard. The reason is that the gold standard cannot assign an affiliation to multiple top-level institutions. Other examples include a joint affiliation by University Göttingen and the Charite or by University Würzburg and a clinic in Bad Mergentheim. In Fig. 18b, we show cases where our method misses true connections between affiliations as no minimal elements are shared. In the first, one would have to know that the paleontological collection is part of the University of Munich. In the next example, there is a typo, et cetera. We conclude that the primary source of error for top-level disambiguation lies in the representation step, which can be improved in a standalone project.

Top-level linking In our framework, a simple linking baseline for individual top-level institutions may use a manually defined linking candidate like UNI : {Trier} as a top-level representation to select all its supersets by performing the corresponding database query. We test this for three universities (see Table 3). Recall of 81–99% and 95–99% precision are in the same range as results for the other fully automatic linking methods [24,44]. Imperfections in precision are usually due to absorption of similar institutions in the same city (e.g. *Bonn Rhein Sieg Univ Appl Sci, St Augustin, Germany* is not part of the *University of Bonn*), while recall mistakes are the result of difficulties during string parsing (e.g. *Univ Hals Nasen Ohrenklin, Heidelberg, Germany* is part of *Heidelberg University*).

Table 3 Results of the linking task

Univ Trier	TRUE id:86	
POSITIVE UNI:{Trier}	17571 (95%) 99%	18488
	17729	

Univ Bonn	TRUE id:168	
POSITIVE UNI:{Bonn}	318179 (99%) 89%	320873
	359330	

Univ Heidelberg	TRUE id:129	
POSITIVE UNI:{Heidelberg}	489437 (97%) 81%	503469
	603658	

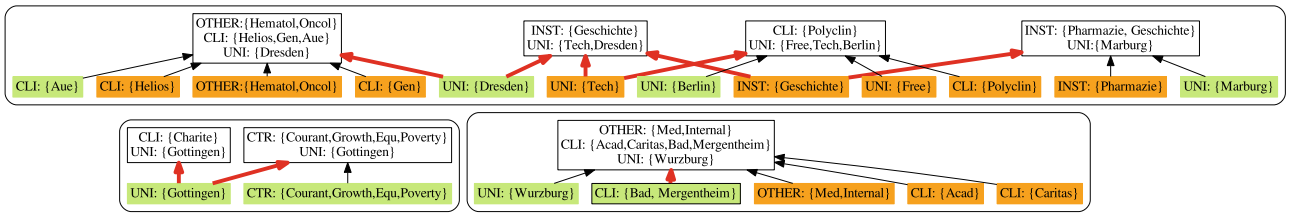
top-level	TRUE target	
POSITIVE predictor	TP Rec (Prec)	P
	T	

TP: true-positive pairs, P: positives, T: true, Prec: precision, Rec: recall. The bottom-right corner legend shows how the numbers are arranged. With positive pairs on the right and true ones on the bottom, precision and recall are displayed in between as proportion TP/P and TP/P, respectively

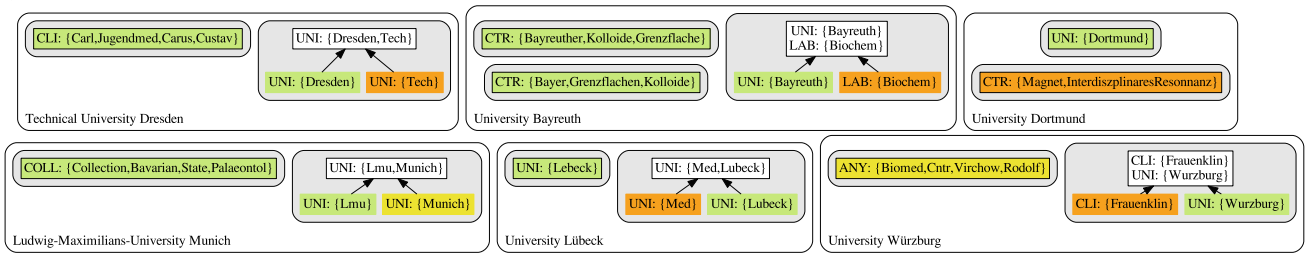
Lower-level resolution Results for hierarchical resolution in the sense of (**T5**) *Order institutions hierarchically* are displayed in Table 4. Percentages are rounded, so that 0% does not necessarily mean TP is zero. The legend in the bottom right corner of Table 3 explains that we use equivalence (=) and subset (<) relation as well as their combination ($\leq$) as predictors to target these same three relations in the hierarchical gold standard. True equivalence can be predicted well both by which affiliations are represented under the same node (100% precision and recall between 82 and 97% or a microaverage of 85%) and by which affiliations are supersets/subsets (100% precision and between 87 and 99% recall or a microaverage of 89%). The latter results are actually better since most predicted proper supersets are true equivalences (65–100% precision or a microaverage of 93%).

Hierarchical resolution Most correct hierarchical relations have been extracted for University Heidelberg with 30% precision (the rest are equivalences); however, still at only 4% recall. According to Table 4 only $3 + 278 + 2{,}415 = 2{,}696$ or 2.5% of all gold hierarchical relations are found. While, for example, the relation between *Univ Trier, Biogeog FB VI* and *Univ Trier, FB VI* is correctly identified in Fig. 15, many affiliations lack clues for such relations. For example,

[1] https://www.otago.ac.nz/wellington/departments/publichealth/research/bode3/index.html.



(a) False-positives; bold red edges indicate representations of different institutions incorrectly linked by the respective minel



(b) False-negatives; the outer frame indicates what should be in the same component – the inner what is connected

Fig. 18 Top-level resolution errors; filled: minimum elements; green: correct minels; orange: overly general minels (color figure online)

Table 4 Numbers and rounded ratios of predicted (positive) and target (true) hierarchical relationships (Legend as in Table 3)

Univ Trier	TRUE =	TRUE ≤	TRUE <	
POSITIVE =	72768 97% 100%	72768 93% 100%	0 0% 0%	72768
POSITIVE ≤	73661 99% 100%	73664 94% 100%	3 0% 0%	73664
POSITIVE <	893 1% 100%	896 0% 100%	3 0% 0%	896
	75106	78643	3537	

Univ Bonn	TRUE =	TRUE ≤	TRUE <	
POSITIVE =	665422 82% 100%	665422 78% 100%	0 0% 0%	665422
POSITIVE ≤	709678 87% 100%	709956 83% 100%	278 1% 0%	710815
POSITIVE <	44256 5% 97%	44534 5% 98%	278 1% 1%	45393
	816372	854103	37731	

Univ Heid.	TRUE =	TRUE ≤	TRUE <	
POSITIVE =	641820 89% 100%	641820 81% 100%	0 0% 0%	641820
POSITIVE ≤	647101 89% 100%	649516 82% 100%	2415 4% 0%	649921
POSITIVE <	5281 0% 65%	7696 1% 95%	2415 4% 30%	8101
	723558	792005	68447	

In contrast with the latter table, here we use three relations as predictors and targets. Values in the diagonals are most informative: for example, in the $=, =$ cells, we predict true equality (same node in true hierarchy) by system-based equality (same node in inferred hierarchy), in $\leq, \leq$ true dominance or equality by system dominance or equality and in $<, <$ true strict dominance by system strict dominance. The other cells give results for using predictors for different targets and are mostly interesting for error analysis

in Fig. 14, *Univ Trier, Dept Hydro* contains no reference to *Department VI - Regional and Environmental Science*. The number of hierarchical relationships that *can* be found is also rather low (6% of all annotated pairs). Many correctly identified hierarchical relationships are not between annotated representations and thus not counted.

7 Discussion

In this section, we discuss our objective results and their implications for future work.

7.1 Top-level resolution/linking

The results presented in the previous section reveal that top-level resolution mainly suffers from overly general incorrect minimal elements such as *UNI:Tech*, which are not trivial to detect, because their attribute structure by itself does not indicate the problem. In the absence of a predefined hierarchical structure, the most difficult part of the resolution task is to detect top-level institutions. Assuming that the top-level institutions are known and that simple representations like *UNI:Trier* can be derived, top-level linking by superset queries works well, i.e. most of the lower-level affiliations' representations contain the respective feature(s). The semi-automatic linking method by Donner et al. [19] achieves close to 100% Precision and Recall. However, it includes plenty of manual interventions that do not scale to larger data and smaller effort. Considering the good linking results, one can see that the additional difficulty of the separation step lies in finding the exact linking candidates in an *unsupervised* fashion, a task which was not considered by previous linking methods that have instead extracted them from reference

datasets [24,44]. For the most part, the problem of overly general representations is rooted in the representation step. However, as one cannot preclude such mistakes and their effect can be devastating with only one overly general representation connecting numerous unrelated affiliations (e.g. all university hospitals), it seems reasonable to consider a filtering step that rejects under-specified representations. It remains to be tested if also the separation step could be adapted for this purpose by considering edge weights for a soft connectedness instead of a hard binary one.

7.2 Lower-level (hierarchical) resolution

The extracted hierarchies also contain information about affiliations that are assigned to the same node. This mapping corresponds to an equivalence relation and provides an effective lower-level resolution with a solid performance. This means that for any known lower-level institution, the set of all affiliations belonging to it (which usually includes also its specifications) can be approximated relatively well already. At the core, this work is an attempt to discover hierarchical relationships between affiliations. This task has been proved very difficult. On the one hand, almost all claimed hierarchical relationships actually describe true equivalences. On the other, practically none of the annotated hierarchical relationships have been discovered (0-4% recall). The reason why most claimed hierarchical relationships are actually equivalences lies in the conservative conflation method and could be addressed with a more aggressive conflation method on higher quality representations, which would also retrieve more true hierarchical relationships. The fact that almost no annotated hierarchical relationships have been found can be explained to some extend by the small size of the hierarchical annotation, but mainly suggests that the subset/superset partial order over the representations as generated by our baseline parser does no reveal these relationships. Figure 14 suggests that many of these are not found since the affiliation strings simply do not include this information (cf. *Univ Trier, Dept Hydrol* which should be under *Univ Trier, Dept Earth Sci FB VI*). Therefore, the relationship can only be discovered using additional information.

7.3 Linking versus resolution

On top of the more fine-grained subdivision of tasks T1–T5, the hierarchical affiliation resolution consists of two major challenges: (1) hierarchy extraction (i.e. the induction of a real world institutional hierarchy) and (2) hierarchical resolution (i.e. the assignment of affiliations to institutional nodes in an institutional hierarchy). Our focus was on inducing the hierarchy from the same affiliation data that were to be resolved hierarchically (thus *unsupervised* hierarchical resolution). Under the assumption of the preexistence of a solution for challenge (1), a linking setup may constitute an easier scenario where an equally basic methodology might produce a better outcome than that which was returned by our fully unsupervised method. In fact, our results support previous top-level resolution approaches in their preference of linking over induction. Obviously, this is only possible where such knowledge is available. All affiliations that cannot be mapped to preexisting hierarchies can still be ordered by our unsupervised method—whether to obtain a partial result or to help inspect the remaining unassigned data. To create our hierarchical gold standard, we have implemented a simple retrieval-based linking method. During verification, only about half of the links were determined correct, showing that linking is not trivial either. The problems can be explained by the same properties that our unsupervised approach exploits, i.e. the mentioning of higher-level institutions in affiliations referring to their descendants, which confuses the retrieval engine as to which aspect of the string is crucial. It suggests that hierarchical linking and resolution are in fact not contradictory, but complementary.

7.4 Limitations and future work

Our experiments have shown the difficulties in various smaller subtasks involved in unsupervised hierarchical affiliation resolution, in particular with respect to the WoS as a large source of heterogeneous affiliations. Based on these insights, future work can focus on individual aspects or framework components. Improved representation would greatly benefit the overall performance as perfect affiliations (or representations thereof) only require the trivial collocation step to reconstruct optimal hierarchies. One approach could be to minimize the need for conflation-based error-correction by directly learning representations such that they recreate a large hierarchical gold standard. While this would be very challenging both in terms of data requirements and in the design of the learning procedure, an equivalence classifier for conflation constitutes a more straightforward supervised learning problem. Finally, interpolation offers an interesting application for understanding and generalizing individual representations.

8 Conclusion

In this work, we have introduced and analysed the task of automatic hierarchical affiliation resolution, breaking it down into five major subtasks. We have underlined its importance in particular for the aggregation of academic performance measures by discussing in Sect. 2 how previous performance comparisons have relied either on top-level resolution or on manual efforts, where the latter limitation has—with the exception of few large-scale government-directed surveys—

led to a limited scale of such studies. We have identified an integrated framework in Sect. 3 that is guaranteed to solve the five subtasks for hypothetical perfect affiliations and, when applied on realistic affiliations, reveals their inherent peculiarities and difficulties through the observable discrepancies between the ideal and the real case. The framework consists of five pipeline components, whereof three are designed specifically to enable the focused handling of the uncovered irregularities. Many easier tasks like top-level resolution or hierarchical linking can be viewed as special cases that correspond to a subset of the tasks and components. We have described a specific ad-hoc baseline to instantiate each component as proof of concept in Sect. 4 and to allow for experimentation and evaluation. To ensure scalability, we have implemented all baselines towards application on the entire WoS with millions of affiliations. We have used a large scale gold standard for top-level resolution to design our experiments in Sect. 5 and created a first hierarchical one that can be extended easily in a semi-automatic fashion.

The findings from Sects. 6 and 7 regarding our first approach to tasks T1–T5 can be summarized as follows: (**T1**) Discovering top-level institutions based only on their representations is quite challenging, in particular due to overly general or underspecified representations. This explains why all previous research has used curated lists of top-level institutions. We only achieve 63% recall and 23% precision with our baseline implementation. These resolution errors could be addressed with better representations. (**T2**) Alternatively, affiliations can be assigned to a given top-level institution by defining a linking candidate representation and selecting all its supersets. Performance is good with precision from 95 to 99% and recall from 81 to 99% (see Table 3) and should improve with better representations. (**T3**) Lower-level institutions can be discovered on all levels by grouping them based on equal or equivalent representations. After conservative conflation, our baseline achieves 100% precision and 82–97% recall (see Table 4). (**T4**) As for T1, affiliations can be assigned to lower-level institutions by selecting all affiliations with representations more or equally specific than those corresponding to the result obtained in T3. We achieve 100% precision and 82–94% recall. This task satisfies the requirements for ordinary lower-level aggregation of performance measures. (**T5**) In addition, lower-level institutions (and their assigned affiliations) can be ordered hierarchically by Hasse diagrams. Although manual inspection confirms that a number of true hierarchical relations are evidently revealed by the current methods, practically none of the annotated hierarchical relationships were found. Error analysis suggests that this is due to the fact that the respective relationships are simply not encoded in the affiliations.

Our results underline the difficulty of hierarchical affiliation resolution and thereby support the implicit assumptions of previous work that has focused on linking to known institutions instead. However, hierarchical resolution is still important, as it offers an alternative when such knowledge is not available. As a result of our first approach, future work can focus on improving specific aspects or framework components, for example, by applying supervised and unsupervised learning in the representation, interpolation or conflation step.

Overall, we have made the following contributions: (**C1**) The five subtasks T1–T5 have been identified by us as the requirements for unsupervised hierarchical affiliation resolution. (**C2**) The five components of our proposed framework provide a systematic way to approach all of these subtasks. (**C3**) This framework has been instantiated with first baselines that have been evaluated through dedicated performance measures and error analysis. (**C4**) To do so, we have introduced a number of options for the evaluation of hierarchical affiliation resolution, including dedicated datasets. (**C5**) In the error analysis and the above concluding remarks, we have stressed which aspects are particularly challenging and pointed out clearly how future research may continue to improve the overall results. We have introduced and covered all aspects of the hierarchical resolution problem—from framework, common pitfalls, methods, implementation and scaling to evaluation. The description is accompanied by various examples and visualizations to underline our proposals and insights.

Acknowledgements We thank the German Research Foundation (DFG) for providing us with the GERiT dataset. We thank Dr. Philipp Mayr-Schlegel for initiating and defending the project proposal.

Funding Open Access funding enabled and organized by Project DEAL. This work was partially funded by the Ministry of Education & Research, Germany (BMBF) under grant number 01PQ17001 Kompetenzzentrum Bibliometrie. Open Access is funded by the DFG under project number 491156185.

Availability of data and material For researchers interested in using, reproducing or improving our results, we are happy to share as much of our output, code and datasets https://sites.google.com/view/hierarchical-affiliations as possible.

Code availability The website ibid. includes a link to a repository with the Python code of this project. Anyone may deploy or modify this code. The authors are available for assistance and discussion.

Declarations

Conflict of interest The authors have no conflict of interest.

References

1. Aksnes, D.W., Langfeldt, L., Wouters, P.: Citations, citation indicators, and research quality: an overview of basic concepts and theories. SAGE Open **9**(1), 2158244019829575 (2019)
2. Altanopoulou, P., Dontsidou, M., Tselios, N.: Evaluation of ninety-three major Greek university departments using Google Scholar. Qual. High. Educ. **18**(1), 111–137 (2012)
3. Clarivate Analytics: Institutional unification: getting the full picture
4. Clarivate Analytics: Organization name unification procedures
5. Aumueller, D., Rahm, E.: Web-based affiliation matching. In: ICIQ, pp. 246–256. Citeseer (2009)
6. Ball, R., Halwachi, J.: Performance indicators in higher education. High. Educ. **16**(4), 393–405 (1987)
7. Ball, R., Wilkinson, R.: The use and abuse of performance indicators in UK higher education. High. Educ. **27**(4), 417–427 (1994)
8. Baty, P.: The times higher education world university rankings, 2004–2012. Ethics Sci. Environ. Politics **13**(2), 125–130 (2014)
9. Bayardo, R.J., Panda, B.: Fast algorithms for finding extremal sets. In: Proceedings of the 2011 SIAM International Conference on Data Mining, pp. 25–34. SIAM (2011)
10. Birch, D.W., Calvert, J.R.: Performance indicators in higher education: a comparative study. Educ. Adm. **5**(2), 15–27 (1977)
11. Borgen, N.T., Mastekaasa, A.: Horizontal stratification of higher education: the relative importance of field of study, institution, and department for candidates' wages. Soc. Forces **97**(2), 531–558 (2018)
12. Chen, S.-P., Chang, C.-W.: Measuring the efficiency of university departments: an empirical study using data envelopment analysis and cluster analysis. Scientometrics **126**(6), 5263–5284 (2021)
13. Cuxac, P., Lamirel, J.-C., Bonvallot, V.: Efficient supervised and semi-supervised approaches for affiliations disambiguation. Scientometrics **97**(1), 47–58 (2013)
14. Davis, P., Papanek, G.F.: Faculty ratings of major economics departments by citations. Am. Econ. Rev. **74**(1), 225–230 (1984)
15. De Bruin, R.E., Moed, H.F.: Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications. Scientometrics **26**(1), 65–80 (1993)
16. De Bruin, R.E., Moed, H.F.: The unification of addresses in scientific publications. Informetrics **89**, 65–78 (1990)
17. DAAD DFG, HRK: GERiT: German research institutions. https://gerit.org (2019)
18. Dillon, E.W., Smith, J.A.: The consequences of academic match between students and colleges. J. Hum. Resour. **55**(3), 767–808 (2020)
19. Donner, P., Rimmert, C., van Eck, N.J.: Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. Quant. Sci. Stud. **1**(1), 150–170 (2020)
20. Fort, M., Antoni Sellarès, J., Valladares, N.: Finding extremal sets on the GPU. J. Parallel Distrib. Comput. **74**(1), 1891–1899 (2014)
21. Galvez, C., Moya-Anegon, F.: The unification of institutional addresses applying parametrized finite-state graphs. Scientometrics **69**(2), 323–345 (2006)
22. Huang, S., Yang, B., Yan, S., Rousseau, R.: Institution name disambiguation for research assessment. Scientometrics **99**(3), 823–838 (2014)
23. Huang, Y., Li, J., Sun, T., Xian, G.: Institution information specification and correlation based on institutional PIDs and IND tool. Scientometrics **122**(1), 381–396 (2020)
24. Jacob, F., Javed, F., Zhao, M., Mcnair, M.: sCool: a system for academic institution name normalization. In: 2014 International Conference on Collaboration Technologies and Systems (CTS), pp. 86–93. IEEE (2014)
25. Jiang, Y., Zheng, H.-T., Wang, X., Binggan, L., Kaihua, W.: Affiliation disambiguation for constructing semantic digital libraries. J. Am. Soc. Inform. Sci. Technol. **62**(6), 1029–1041 (2011)
26. Johnes, G.: Performance indicators in higher education: a survey of recent work. Oxf. Rev. Econ. Policy **8**(2), 19–34 (1992)
27. Johnes, G., Johnes, J.: Measuring the research performance of UK economics departments: an application of data envelopment analysis. Oxford Econ. Pap. **45**, 332–347 (1993)
28. Johnes, J.: Performance indicators and rankings in higher education. In: Valuing Higher Education: An Appreciation of the Work of Gareth Williams, pp. 77–105. UCL Institute of Education Press (2016)
29. Johnston, R.J., Jones, K., Gould, M.: Department size and research in English universities: inter-university variations. Qual. High. Educ. **1**(1), 41–47 (1995)
30. Jonnalagadda, S., Topham, P.: NEMO: extraction and normalization of organization names from PubMed affiliation strings. J. Biomed. Discov. Collab. **5**, 50 (2010)
31. Kells, H.R., Mundial, B.: Performance Indicators for Higher Education: A Critical Review with Policy Recommendations. Education and Employment Division, Population and Human Resources Department, World Bank (1992)
32. Kronman, U., Gunnarsson, M., Karlsson, S.: The bibliometric database at the swedish research council—contents, methods and indicators. Technical report, Stockholm: Swedish Research Council (2010)
33. Leiserson, C.E., Maza, M.M., Li, L., Xie, Y.: Parallel computation of the minimal elements of a poset. In: Proceedings of the 4th International Workshop on Parallel and Symbolic Computation, pp. 53–62 (2010)
34. Liebowitz, S.J., Palmer, J.P.: Assessing assessments of economics departments. Technical Report Working Paper 83-01 C.E.A.P.R. Department of Economics, University of Western Ontario (1986)
35. Marinov, M., Nash, N., Gregg, D.: Practical algorithms for finding extremal sets. J. Exp. Algorithmics **21**, 1–21 (2016)
36. Miroiu, A., Păunescu, M., Vîiu, G.-A.: Ranking Romanian academic departments in three fields of study using the g-index. Qual. High. Educ. **21**(2), 189–212 (2015)
37. Morillo, F., Aparicio, J., González-Albo, B., Moreno, L.: Towards the automation of address identification. Scientometrics **94**(1), 207–224 (2013)
38. Orduña-Malea, E., Ayllón, J.M., Martín-Martín, A., López-Cózar, E.D.: The lost academic home: institutional affiliation links in google scholar citations. Online Information Review (2017)
39. Pritchard, P.: Opportunistic algorithms for eliminating supersets. Acta Inform. **28**(8), 733–754 (1991)
40. Pritchard, P.: A simple sub-quadratic algorithm for computing the subset partial order. Inf. Process. Lett. **56**(6), 337–341 (1995)
41. Pritchard, P.: An old sub-quadratic algorithm for rinding extremal sets. Inf. Process. Lett. **62**(6), 329–334 (1997)
42. Pritchard, P.: On computing the subset graph of a collection of sets. J. Algorithms **33**(2), 187–203 (1999)
43. Rimmert, C., Schwechheimer, H., Winterhager, M.: Disambiguation of author addresses in bibliometric databases. Technical report, Bielefeld University (2017)
44. Shao, Z., Cao, X., Yuan, S., Wang, Y.: ELAD: an entity linking based affiliation disambiguation framework. IEEE Access **8**, 70519–70526 (2020)

45. Shen, H.: Fully dynamic algorithms for maintaining extremal sets in a family of sets. Int. J. Comput. Math. **69**(3–4), 203–215 (1998)
46. Sizer, J., Spee, A., Bormans, R.: The role of performance indicators in higher education. High. Educ. **24**(2), 133–155 (1992)
47. Su, J.-L.: The effects of the trial implementation of a departmental evaluation project in Taiwan. Qual. High. Educ. **1**(2), 159–172 (1995)
48. Yellin, D.M.: Algorithms for subset testing and finding maximal sets. In: Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 386–392 (1992)
49. Yellin, D.M., Jutla, C.S.: Finding extremal sets in less than quadratic time. Inf. Process. Lett. **48**(1), 29–34 (1993)