# Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems

**Paul Donner**[1] iD **, Christine Rimmert**[2]**, and Nees Jan van Eck**[3] iD

[1]German Centre for Higher Education and Research Studies, DZHW, Berlin, Germany
[2]Bielefeld University, Germany
[3]Centre for Science and Technology Studies, Leiden University, Netherlands

## ABSTRACT

The present study is an evaluation of three frequently used institution name disambiguation systems. The Web of Science normalized institution names and Organization Enhanced system and the Scopus Affiliation ID system are tested against a complete, independent institution disambiguation system for a sample of German public sector research organizations. The independent system is used as the gold standard in the evaluations that we perform. We study the coverage of the disambiguation systems and, in particular, the differences in a number of commonly used bibliometric indicators. The key finding is that for the sample institutions, the studied systems provide bibliometric indicator values that have only a limited accuracy. Our conclusion is that for any use with policy implications, additional data cleaning for disambiguating affiliation data is recommended.

## 1. INTRODUCTION

Scientometric studies at the level of research institutions face the challenge of the correct attribution of publications to institutions. This task, here referred to as *institution name disambiguation*, comprises systematically standardizing the heterogeneous address data of the author-provided affiliation information present in publications and recorded in bibliographic databases. At present, institutional affiliation information in academic publications is not standardized and unique identifiers for research institutions have not yet been adopted. Therefore, in order to generate valid primary data on publications for studies at the meso level, the assignment of address strings to known real institutional entities is crucial. Institution name disambiguation belongs to a class of problems known as *named entity normalization*, in which variant forms have to be matched to the correct preferred form. Another prominent member of this class is author name disambiguation. Disambiguated affiliation information can contribute to the performance of author name disambiguation systems that employ affiliations as background information.[1]

In the recent past, a nearly complete institutional disambiguation for German research institutions was developed and implemented at the Institute for Interdisciplinary Studies of

---

[1] Likewise, disambiguated author information could potentially be used as additional input information for institutional disambiguation. However, we are not aware of any literature on this approach.

Science at Bielefeld University, as a major component of a national bibliometric data infra-structure for research and monitoring (Rimmert, Schwechheimer, & Winterhager, 2017). The system has been tested and improved over a number of years and is now in production use. We are therefore in a position to study the degree to which the use of a sophisticated disambiguation system with near-complete national-scale coverage leads to different biblio-metric indicator values compared to a situation in which no such system is available and sim-pler alternatives to the attribution problem have to be used. We consider here (a) the case where a simple unification strategy using ad hoc lexical searches in the address data fields of a bibliographic database is conducted in order to collect publications of the target institu-tions (based on vendor preprocessed affiliation data in Web of Science [WoS]); and (b) the use of bibliographic database vendors' own institution disambiguation systems (in both WoS and Scopus). We believe that these two situations are common in practice outside of specialized research or evaluation units with access to the raw data of bibliographic databases. The per-formance and implications of these approaches are therefore relevant and of wide interest to the bibliometrics and research evaluation communities. Prominent examples with profound science-political consequences of the use of bibliometric data of institutions derived from WoS or Scopus are the international university rankings, which generate much public atten-tion and elicit considerable debate.

The remainder of the article is structured as follows. We begin by providing an overview of the prior work on institutional disambiguation. Next, we briefly outline the institution name disambiguation systems that we study and describe the publication data and institution samples that we use, as well as the comparison scenarios, the bibliometric indicators that are calculated for the institutions, and our approach of assessing the differences in indicator values. In the next section we present the results of our comparisons. In particular, we assess the distributions of errors in indicator values over institutions arising when applying alternative disambiguation systems in contrast to the reference values obtained from the presented dis-ambiguation system for German research institutions, which can be assumed to be complete and nearly error free for the data. The results and their implications are summarized in the Discussion section.

## 2. RELATED WORK

Unification of author affiliation information and the allocation to clearly identified research institutions has been recognized as a challenging task in the bibliometric research community and beyond. Accurate disambiguation of heterogeneous affiliation data is crucial for institution-level scientometric research and bibliometric evaluation. Disambiguation systems connecting heterogeneous author affiliations to known research institutions have been constructed in sev-eral projects—usually for project-specific purposes and not to be made available publicly. They may be roughly divided into rule-based and machine learning approaches. However, this division is not a strict one, as approaches often use combinations of methods (e.g., rules and some manual work are used in addition to a machine learning approach to improve pre-cision, especially for problematic cases).

### 2.1. Rule-based Approaches

A substantial amount of work on this topic has been done at the Centre for Science and Technology Studies (CWTS) at Leiden University. For the case of universities, this began with De Bruin and Moed (1990). They performed a unification of about 85,000 affiliations from 75 journals (data from SCISEARCH) using the first part of the addresses. Using structural

information and reference books, they assigned units on lower hierarchical levels (e.g., departments), appearing in the first part of addresses, to the corresponding main organization. They found that many problems remained, and to solve these they used external information from encyclopedias, university handbooks, specialists, and staff lists of universities. This is a time-consuming method, and they only did this for selected countries (in particular the Netherlands). In a follow-up study, Moed, De Bruin, and Van Leeuwen (1995) reported on a bibliometric database constructed from all articles published by authors from the Netherlands using data from the Science Citation Index. To store unified affiliations, they improved their earlier procedure for Dutch addresses by, among other things, adding a classification of institutions to research sectors: that is, types of organizations such as universities, hospitals, and firms. They noted problematic affiliations that could not be handled correctly by their procedure. CWTS continued to maintain and improve its disambiguation system, in particular for its university ranking, for which all name variants that occur at least five times in the WoS are cleaned (Waltman et al., 2012). This system pays special attention to the way publications by academic hospitals are assigned to universities (Reyes-Elizondo, Calero-Medina, Visser, & Waltman, 2016).

The Swedish Research Council performed affiliation disambiguation for its bibliometric database, which was also constructed based on WoS data (Kronman, Gunnarsson, & Karlsson, 2010; Swedish Research Council, 2017). They used a deterministic approach based on a catalog of string rules, mapping address strings to 600 known Swedish research organizations. Organizations were also classified by the research sector. Their procedure was able to assign over 99% of Swedish address strings. A single address may be matched to more than one organization in the case of affiliations containing information on more than one organization, usually indicating collaborations.

### 2.2. Machine Learning Approaches

French, Powell, and Schulman (2000) described a number of institutional disambiguation experiments with different address string distance metrics and a one-pass heuristic clustering procedure. The clearly stated goal was not a complete automatic disambiguation, but rather the reduction of manual reviewing of the most difficult cases. Among other things, they introduced a new, domain-specific affiliation comparison function, based on normalized and sorted words, minimizing edit distances between aligned words across possible permutations. Jonnalagadda and Topham (2010) reported on their disambiguation of institution names extracted from PubMed data. The presented approach utilized agglomerative clustering, for which the entity similarity is computed with an edit distance, building on the work of French et al. (2000). In particular, their approach was a hybrid of a sequence alignment measure over word sequences (Smith–Waterman algorithm) and the Levenshtein distance between single words. Furthermore, similar clusters were merged. They reported sample precision values of 99.5% (4,135 affiliation strings related to "Antiangiogenesis," only US organizations) and 97.9% (1,000 affiliation strings related to "Diabetes," organizations from any country) for organization normalization. Although these values are high, it is not possible to extrapolate them to less restricted data sets. Galvez and Moya-Anegón (2006, 2007) reported on a new approach using finite-state graphs, developed with WoS data and also tested on data from Inspec, Medline, and CAB Abstracts. Although this is a promising approach, the authors outlined the limits of automatic classifications for problematic affiliations, which requires expert knowledge to classify. Jiang, Zheng, Wang, Lu, and Wu (2011) discussed an experimental approach of agglomerative clustering of affiliations using string compression distance. Their evaluation of the method is questionable, as they use the publication affiliations of mostly students

and staff from a single university. Their affiliation string pool is, therefore, dominated by name variants of that university, although the remainder are affiliations of coauthors. They extracted a reference corpus of 217 "affiliations" (variants) of 105 "categories" (true organizations). In any case, their clustering quality results are not encouraging. This also holds true for the application of supervised and semisupervised machine learning methods, tested by Cuxac, Lamirel, and Bonvallot (2013) on French CNRS addresses. Huang, Yang, Yan, and Rousseau (2014) proposed an algorithm using author information to classify affiliations that received high precision values but low recall.

We can conclude that the problem of institution name disambiguation is far from being solved. For the objective of achieving highly accurate disambiguation, it seems that simple methods have not yet been replaced by fully automatic methods, despite the experimental application of several sophisticated approaches with partly promising results on small scales. However, significant progress has been made on affiliation string similarity calculation methods. Both rule-based and machine learning methods can be used to minimize the necessary amount of manual human decisions. Nevertheless, the necessarily higher amount of labor required by rule-based methods means that they have only been applied to parts of all author affiliations, typically to those from one country or discipline. No standard evaluation data set is available for this task. Furthermore, none of the studies have investigated the effects of institutional disambiguation on the quality of bibliometric indicator scores.

## 3. SYSTEMS, DATA, AND METHODS

### 3.1. Institution Disambiguation Systems

In this section we summarize the disambiguation system that was developed for German institutions. For a full description of the system we refer readers to Rimmert et al. (2017). The system, which we call the *KB system*,[2] is comprised of (a) a set of known and uniquely identified German research institutions, (b) a mapping of institutions to affiliation records identified as belonging to each institution from the two data sources WoS and Scopus, (c) a hierarchical classification of the institutions into sectors, and (d) a change history of the institutions which record the splitting and merging and incorporation of institutions and sector changes. The KB system is thus built on the affiliation data provided in WoS and Scopus, respectively, and belongs to the category of rule-based systems. The tracking of structural changes affords the necessary flexibility in handling such changes required by different project contexts. In the KB system, two different analytical views are implemented (item 4 above). With Mode S (for synchronic allocation), we can perform analyses that take into account the institutional structures as they were at the time of publication for each paper. Institutions that have later come to be related to another institution through structural changes, such as through mergers or splits, are treated as different entities. On the other hand, with Mode A (asynchronic, current perspective), we can analyze the publication records of institutions as they are constituted at present; that is, including publications of predecessor units. The mapping of institutions to affiliation records (item 2 above) is a deterministic, rule-based classification. The core of the institutional coding procedure is a mapping of author addresses to the corresponding uniquely identified research institutions and their subdivisions, using a large library of regular expressions. This library currently contains some 45,000 expressions and is continuously being expanded and improved.

---

[2] The system was created in a project called *Kompetenzzentrum Bibliometrie* (Competence Centre for Bibliometrics).

The sector classification (item c above) contains the classes of higher education sectors (universities and universities of applied sciences), four major nonuniversity research organizations (Fraunhofer-Gesellschaft [FHG], Helmholtz Association [HGF], Leibniz Association [WGL], and the Max Planck Society [MPG]), private companies, registered associations, government laboratories, and academies of science. For the sector information, structural changes over time and multiple assignments of research institutions to these sectors are also available.

The version of the KB system used for this study contained 2,097 institutions, which also included placeholder records for unidentified institutions for which only the sector could be determined. An evaluation of the KB disambiguation system was conducted prior to the main study. We provide a detailed overview of the system evaluation in Appendix A for German research institutions. We conclude that, based on the good results of this evaluation, the KB system is a valid, gold standard benchmark for German institutional affiliation disambiguation data. This is not to say, however, that the KB system or its rule-based approach are superior in general. In fact, its scope is limited to a single country and it would be difficult to extend the method to global scope because of the large effort and unreasonable expense required.

We deliberately do not attempt to describe the workings of the proprietary institution disambiguation systems of WoS and Scopus and regard them as black boxes, of which we only analyze the results. The reason for this is that both systems are not documented in any detail by the providers. What we can gather from the information of the platforms is that WoS Organizations Enhanced (OE) is based on lists of variant names mapped to preferred names.[3] WoS OE can therefore be seen as a rule-based system. Regarding the Scopus Affiliation Identifiers (AFIDs), the documentation merely informs us that "the Affiliation Identifier distinguishes between affiliations by assigning each affiliation in Scopus a unique number, and grouping together all of the documents affiliated with an organization."[4] No information is given about how the system works.

### 3.2. Data

The data used in the analyses are derived from the licensed source of WoS[5] and Scopus, obtained in spring 2017. The data were loaded into in-house relational databases, cleaned, and enhanced at the Competence Centre for Bibliometrics for Germany. The most important enhancement is the disambiguation of German author addresses to known German research institutions. This process is conducted separately for each data source using the KB disambiguation system described in the previous subsection. The units of the analysis for this study are German academic institutions, in particular universities, universities of applied sciences, and nonuniversity research institutes. Publications are restricted to articles and reviews published between 1995 and 2017. To be included, an institution needed to have at least 50 such publications associated with it according to the KB disambiguation of the WoS data. These restrictions resulted in a study sample of 445 institutions. The same institutions are used to investigate both WoS and Scopus.

---

[3] https://clarivate.libguides.com/woscc/institution; http://ips.clarivate.com/m/pdfs/UnifyingNameVariants.pdf, accessed April 13, 2018.

[4] https://service.elsevier.com/app/answers/detail/a_id/11215/supporthub/scopus/related/1, accessed April 13, 2018.

[5] The WoS data used in this study include data from the Science Citation Index Expanded, the Social Sciences Citation Index, the Arts & Humanities Citation Index, and the Conference Proceedings Citation Index.

### 3.2.1. Scopus AFID

For the Scopus data, we compare the KB system-derived reference data to sets of publications that have one or more common assigned AFIDs (affiliation identifiers), as provided by Elsevier. Some preprocessing steps to align the Scopus and KB disambiguation systems were performed in order to make them comparable, as they are conceptually and structurally somewhat different. To match AFIDs to the KB system IDs, the AFID for each institution in our sample was obtained by searching Scopus's online platform. It is not clear whether and how exactly the definition of an institution in Scopus differs from the one the KB disambiguation is based on. One difference that we have noticed is that the AFID system typically has separate IDs for university hospitals and the universities they belong to, which is not the case in the KB system. We have therefore merged those AFIDs to create more comparable and consistent publication record sets. Furthermore, in some cases more than one AFID for the same institution exists in Scopus, for instance, for multiple branch locations. If these are logically linked in the hierarchical relations in the Scopus system, we also merged these linked AFIDs. If not, we took only the most commonly used AFID per institution.

We found that in the AFID system, publications with affiliations referring to predecessor units are grouped with their current unit. Based on this observation, we compare the AFID results with those from the KB system's Mode A.

### 3.2.2. Web of Science (WoS) organization enhanced

The WoS OE system does not have unit identifiers but Preferred Names, which are additionally assigned as institution names to affiliations considered belonging to one real institution. In order to identify the WoS Preferred Name for the institutions in our set, we started by identifying all the Preferred Names of records with German addresses that occur more than 20 times. From this list, we chose the Preferred Name matching the target institution and otherwise excluded the institution from this part of the study. In fact, for our sample set, it was not possible to retrieve the corresponding publications on the main institutional level in a majority of cases. Although many universities are recorded in OE, the institutions of FHG, HGF, WGL, and MPG are almost all grouped such that only all publications of each of the respective head organizations can be found, but rarely those of their member institutes.[6]

Similar to AFID, also in the WoS OE system, predecessor units are grouped under the Preferred Name of the current institution. In consequence, we also compare the WoS OE system with Mode A of the KB system.

### 3.2.3. WoS institution name search

As well as the comparison of WoS OE data with the KB disambiguation, we also investigated the performance of a lexical search using the institution name in the WoS affiliation data. As pointed out above, the coverage of institutions in the WoS OE system is far from complete (since only head organizations are covered, not their member institutes), which supports the notion that such an alternative approach might often be required in practice. The institution name search method makes use of WoS disambiguation efforts, because institution names extracted from affiliation information in papers are not indexed identically to how they are given in the original publication but are normalized. Because the affiliations in Scopus are not

---

[6] After reaching out to Clarivate Analytics for comments, we were informed that the decision to include the MPG as a whole but not its member institutes was based on a request by MPG. The same approach was then also taken for the other nonuniversity research organizations.

transformed or normalized, we do not apply a similar search strategy to Scopus data. In fact, it is not possible to conduct comparable searches across these two databases because WoS only contains normalized address strings, while Scopus only contains the original address strings.

In this scenario, we model a hypothetical user who has a list of the names of the German research institutions available, which is used as a basis for generating search terms. We also assume that the user is familiar with searching in WoS data to a sufficient degree. This scenario further requires a definition of the name list, the search terms, and the search parameters.

In order to generate a plausible name list, we begin by using the KB institutional disambiguation results to find the most common normalized name in the WoS data for each real institution in our initial set, because in principle there should be only one normalized name for

**Table 1.** Overview of selected bibliometric indicators

| Domain | Indicator | Remarks |
|---|---|---|
| Publication output | P: Number of publications (full count) | The number of publications of an institution |
| Collaboration | P(collab): Number of collaborative publications | The number of an institution's publications that have been coauthored with one or more other institutions |
| | PP(collab): Proportion of collaborative publications | The proportion of an institution's publications that have been coauthored with one or more other institutions |
| | P(int collab): Number of international collaborative publications | The number of an institution's publications that have been coauthored by two or more countries |
| | PP(int collab): Proportion of international collaborative publications | The proportion of an institution's publications that have been coauthored by two or more countries |
| Citation impact | TCS: Total citation score | The total number of citations of the publications of an institution |
| | MCS: Mean citation score | The average number of citations of the publications of an institution |
| | TNCS: Total normalized citation score | The total number of citations of the publications of an institution, normalized for field (WoS: Subject Category; Scopus: ASJC) and publication year |
| | MNCS: Mean normalized citation score | The average number of citations of the publications of an institution, normalized for field (WoS: Subject Category; Scopus: ASJC) and publication year |
| | P(top 10%): Number of highly cited publications | The number of an institution's publications that, compared with other publications in the same field and the same year, belong to the top 10% most frequently cited |
| | PP(top 10%): Share of highly cited publications | The proportion of an institution's publications that, compared with other publications in the same field and the same year, belong to the top 10% most frequently cited |

**Table 2.** Coverage of sample institutions by the studied disambiguation systems

| Sector | Number of institutions | Covered in WoS OE | Covered in WoS search | Covered in Scopus AFID |
|---|---|---|---|---|
| Fraunhofer-Gesellschaft | 62 | 1 | 62 | 57 |
| Helmholtz Association | 23 | 9 | 23 | 16 |
| Max Planck Society | 86 | 0 | 86 | 78 |
| Universities | 107 | 66 | 107 | 96 |
| Leibniz Association | 83 | 8 | 83 | 56 |
| Universities of applied sciences | 90 | 9 | 90 | 75 |
| Total | 445 | 91 | 445 | 376 |

*Note.* The figure in the row "Total" may differ from the sum of the above cells because some institutions are assigned to more than one sector.

each institution. We manually assess the lists side by side with the real names and discard any WoS name that cannot be deduced from the name list, using instead the next most common name variant iteratively until all WoS normalized names are mapped to KB system IDs based on the names in the two systems. This relates to our decision of going beyond a completely naïve and automatic procedure and including a realistic degree of user common sense and domain familiarity. We use the search term list thus obtained as retrieval inputs, while also ignoring capitalization and allowing truncation at the end of the term, and searched the full address information field. This came reasonably close to an informed, but nonspecialized, search for an institution on the online platform of WoS. It is general in the sense that all institutions are treated in the same way and no special knowledge of affiliation idiosyncrasies is included. It is limited in the sense that we only consider one name variant per institution.

Because we directly use the normalized affiliation data as it is indexed in WoS, it is clear that we use the normalized versions of the institution names at the time of publication. Thus, we use Mode S of the KB system for comparison.

### 3.3.  Methods

To assess the performance of the studied systems in terms of being able to identify the correct publications of the research institutions we use the information retrieval measures of precision and recall. For this task, precision is calculated as the share of correctly retrieved publications among the total number of retrieved publications. Recall is the share of correctly retrieved publications among all relevant publications. The correct publications of an institution are those identified by the KB system.

In order to quantify the effect of the application of a specific institutional disambiguation on scores of bibliometric indicators, we calculated the indicator values based on the publications of each institution as retrieved by the KB system—considered a validated gold standard for the

**Table 3.** Summary statistics of the distributions of precision and recall of retrieved publications per institution for WoS OE ($n = 91$)

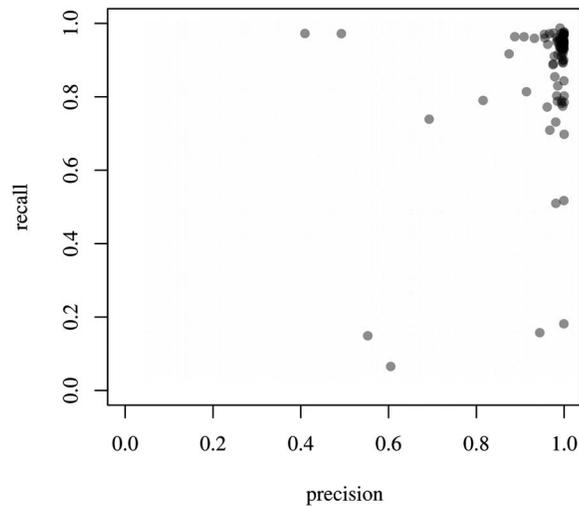| | Minimum | Median | Arithmetic mean | Weighted mean | Maximum | Standard deviation |
|---|---|---|---|---|---|---|
| **Precision** | 0.41 | 0.99 | 0.96 | 0.95 | 1.00 | 0.11 |
| **Recall** | 0.07 | 0.94 | 0.87 | 0.93 | 0.99 | 0.18 |

**Figure 1.**   Precision and recall for WoS OE (*n* = 91).

selected institutions in this study—and for each of the three alternative systems. The differences of indicator values are calculated, and the arising error distributions are displayed.

A number of commonly utilized bibliometric indicators are included in this study. We consider the three domains of publication output, collaboration, and citation impact. For the latter two domains we have selected indicators that are size dependent (absolute numbers) as well as size-independent indicators (ratios or averages). The citation indicators are all calculated for 5-year citation windows which include the year of publication. The indicators are summarized in Table 1. It is clear that the size-dependent indicator values are directly related to the number of correctly identified publications. However, it might be hypothesized that the values of size-independent indicators are less affected when only a part of the correct publication set is used as their input, because errors may cancel each other out.

We compare two vendor-provided disambiguation system results and one search-based result with the KB system's results, which we take as the correct result providing reference values. We divide the system evaluation into two parts. First, for each institution in the evaluation set, we would like to find all its publications, without retrieving any publications it was not involved in. This is a basic information retrieval task, which can be measured with precision and recall. We also use retrieval performance, including the absolute number of retrieved institutions in the evaluation set, to analyze the coverage of the systems with respect to our sample of 445 institutions. The second component of the evaluation concerns the bibliometric

**Table 4.** Institutions with low recall for WoS OE

| Institution Name (KB system) | WoS OE Preferred Name | Recall |
|---|---|---|
| Helmholtz-Zentrum für Infektionsforschung | Helmholtz Center Infectious Research | 0.07 |
| Hochschule Fresenius | Fresenius University of Applied Sciences | 0.15 |
| Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB | Fraunhofer Optronics, System Technologies & Image Exploitation Ettlingen | 0.16 |
| Leibniz-Institut für Festkörper- und Werkstoffforschung Dresden | Leibniz Institute Solid State & Materials Research Dresden | 0.18 |

**Table 5.** Deviation of indicator scores of WoS OE from KB system (*n* = 91)

| Indicator | Percent deviation within ±5% | Median absolute deviation (%) |
|---|---|---|
| P | 40.7 | 4.5 |
| P(collab) | 38.5 | 4.3 |
| PP(collab) | 93.4 | 0.6 |
| P(int collab) | 37.4 | 3.4 |
| PP(int collab) | 86.8 | 1.3 |
| TCS | 52.6 | 4.0 |
| MCS | 79.1 | 1.3 |
| TNCS | 40.7 | 3.7 |
| MNCS | 85.7 | 1.3 |
| P(top 10%) | 48.4 | 3.7 |
| PP(top 10%) | 100.0 | 0.04 |

indicator scores calculated from the retrieved institution publication sets. In general, the numerical discrepancy between the indicator values, using the KB disambiguation (reference values) and the other methods, will be expressed as relative deviation in percent, calculated as

$$deviation = (observed\ system\ score - KB\ system\ reference\ score)/KB\ system\ reference\ score \times 100$$

The deviation has a lower bound at −100% and is unbounded in the positive direction. For example, let the reference MCS of a unit be 5.5 (calculated based on the KB disambiguated data), and the focal value obtained from a simple institution search in WoS be 4.2. Then the deviation as defined above is (4.2 − 5.5)/5.5 × 100 = −23.6%. In this case, the correct result would be underestimated by 23.6%.

For each indicator, the computed deviations for each institution are collected. Our main measure of accuracy is the percentage of values within a range of ±5% of the reference score.

## 4. RESULTS

An overview of the coverage of German institutions in the WoS and Scopus institution disambiguation systems and the lexical search method in WoS is provided in Table 2. We are able to find only 91 of our 445 (20%) evaluation sample institutions in the OE system. The coverage of OE is the lowest among the systems considered. To a significant extent, this is a consequence of the choice not to include the member institutes of nonuniversity research organizations in WoS OE. The set of covered institutions in WoS OE is comprised mostly of universities. However, also for the universities, in particular for the universities of applied sciences, a significant number of institutions are not covered in WoS OE. Using the search strategy, we can find one normalized form for each institution, achieving complete coverage of the institutions. The Scopus AFID system covers 376 (85%) of the institutions with no conspicuous differences between sectors.
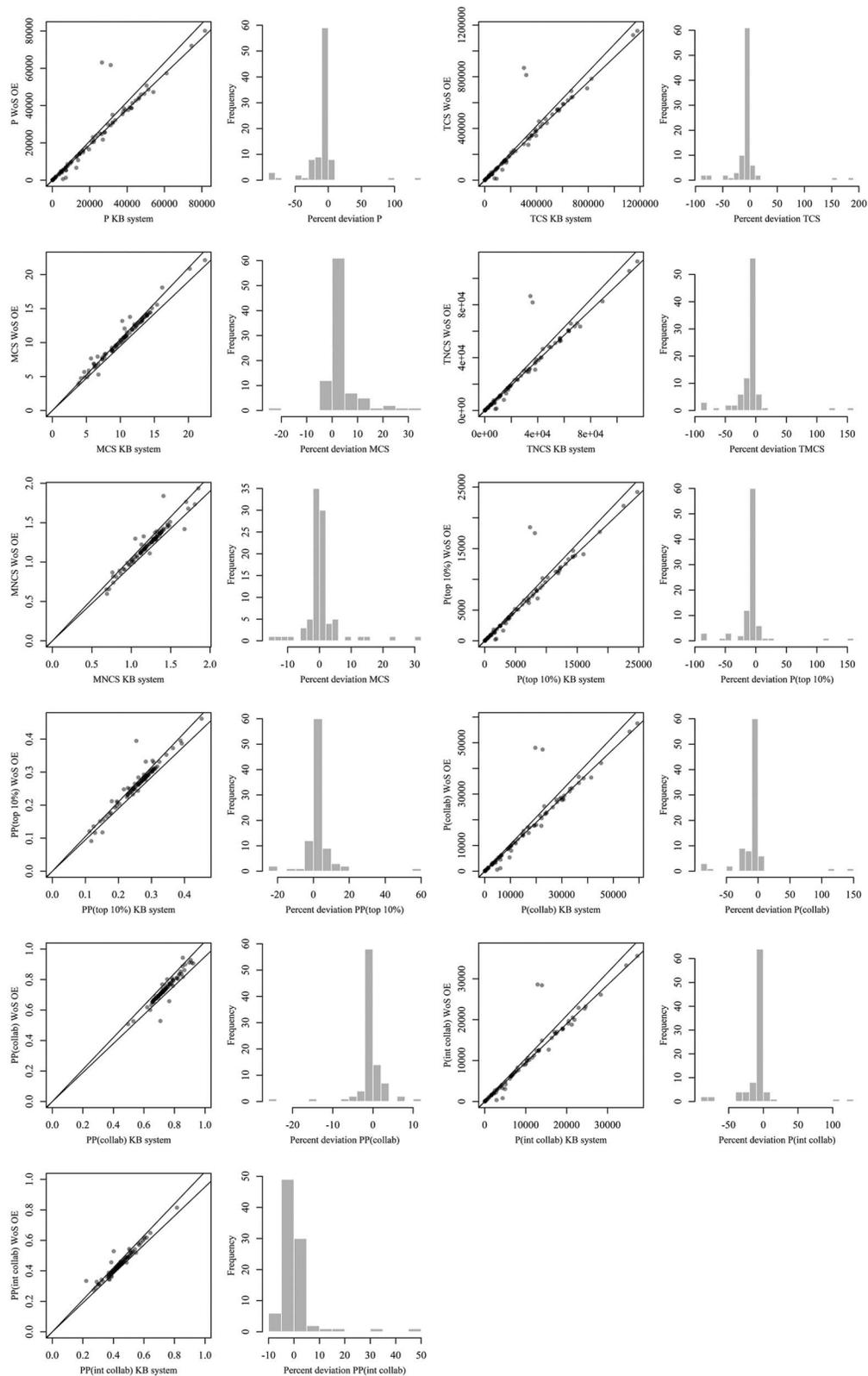
**Figure 2.** Distributions of indicator score deviations of WoS OE from KB system in Mode A. Diagonal lines indicate the ±5% error margin for indicator values.

**Table 6.** Summary statistics of distributions of precision and recall of retrieved publications per institution for WoS institution name search (*n* = 445)

|  | Minimum | Median | Arithmetic mean | Weighted mean | Maximum | Standard deviation |
|---|---|---|---|---|---|---|
| **Precision** | 0.08 | 0.69 | 0.67 | 0.61 | 0.99 | 0.19 |
| **Recall** | 0.03 | 0.51 | 0.55 | 0.74 | 1.00 | 0.26 |

### 4.1. WoS Organizations Enhanced

We present the institution-level figures for precision and recall for WoS OE in Table 3 and Figure 1. All results should be interpreted with due caution because of the OE system's limited coverage of the selected institutions. The precision of WoS OE for these institutional publication sets is 0.95, on average, across institutions, weighted by publication numbers. Hence, typically about 5% of the returned publications in a result set of a specific preferred name will be false positives. The weighted mean of recall across institutions is 0.93, meaning that the result sets do not include about 7% of relevant publications, on average. The contrast between unweighted (0.87) and weighted mean for the recall shows that the results for larger institutions (in terms of number of publication) are better than for smaller institutions. We found poor results for recall for the four institutions presented in Table 4.

We now turn to the results of the comparison of the scores of bibliometric indicators between the WoS OE and the KB system. The results are presented in Table 5, in the form of summaries of the deviation score distributions, visualized in Figure 2. It can be seen that absolute indicator scores (number of publications, collaborative publications, and citations) are less often within the range of nearly correct values (±5%) than relative indicator scores.

### 4.2. WoS Institution Name Search

In this section, we compare the results of the WoS institution name search with those of the KB system. Note that the search makes use of the institution name normalization of WoS, and we
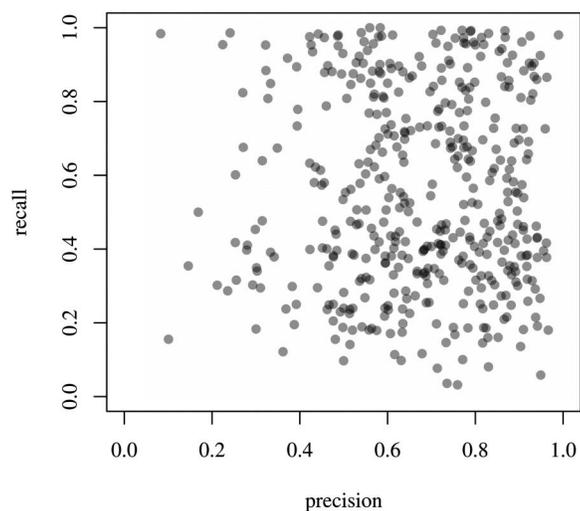


**Figure 3.** Precision and recall for WoS institution name search (*n* = 445).

**Table 7.** Institutions with low recall for WoS institution name search

| Institution name (KB system) | Most common WoS normalized name | Recall |
|---|---|---|
| Max-Planck-Forschungsgruppe Optik, Information und Photonik | INST OPT INFORMAT & PHOTON | 0.03 |
| Leibniz-Institut für Katalyse e. V. an der Universität Rostock (LIKAT) | LEIBNIZ INST KATALYSE EV | 0.04 |
| Friedrich-Miescher-Laboratorium für biologische Arbeitsgruppen in der Max-Planck-Gesellschaft | FRIEDRICH MIESCHER LAB | 0.06 |
| Berliner Elektronenspeicherring - Gesellschaft für Synchrotronstrahlung mbH | BERLINER ELEKTRONENSPEICHERRING GESELL SYNCHROTRO | 0.08 |
| Hochschule für angewandte Wissenschaft und Kunst, Fachhochschule Hildesheim/Holzminden/Göttingen | HAWK UNIV APPL SCI & ARTS | 0.08 |

have deliberately searched for the single most common WoS normalized name per institution, as mentioned above. Using this search method, we obtain vastly more institution publication sets than using WoS OE; in fact, full coverage of all sample institutions is achieved (see Table 2). The summary of the distributions of precision and recall is given in Table 6 and the values are displayed in Figure 3. We obtain rather poor results for the average precision of 0.61 when weighting institutions by the number of publications, and 0.67 as the arithmetic mean. Publication sets for this method will often contain many publications incorrectly assigned to the institutions in question. Recall is at 0.74 weighted mean and 0.55 arithmetic mean, which means that the publication lists returned by these queries will commonly be incomplete, but less so for the larger institutions. Tables 7 and 8 provide the five institutions with the lowest recall and precision scores.

These results for recall suggest that the normalization procedure of WoS is often unable to group most of the relevant institution name variants under one normalized form.

The results of the comparison of the bibliometric indicator scores between the WoS institution name search approach and the KB system for Mode S are provided in Table 9 and the deviation distributions are displayed in Figure 4. The shares of institutions for which the scores obtained with the institution name search approach are within ±5% of the reference score are low, especially for the absolute indicators. Dispersion of the deviations is high. Moreover, the ratio- and mean-based citation scores are comparatively less inaccurate. Evidently, the incomplete publication list result sets of this method lead to substantially inaccurate scores for all indicators.

**Table 8.** Institutions with low precision for WoS institution name search

| Institution name (KB system) | Most common WoS normalized name | Precision |
|---|---|---|
| Max-Planck-Institut für Biologie | MAX PLANCK INST BIOL | 0.08 |
| Max-Planck-Institut für Biophysik | MAX PLANCK INST BIOPHYS | 0.10 |
| Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB | FRAUNHOFER IOSB | 0.15 |
| ESCP Europe Wirtschaftshochschule Berlin e.V. | ESCP EUROPE | 0.17 |
| Forschungszentrum für Marine Geowissenschaften GEOMAR | GEOMAR | 0.21 |

**Table 9.** Deviation of indicator scores of WoS institution name search from KB system (*n* = 445)

| Indicator | Percent deviation within ±5% | Median absolute deviation [%] |
|---|---|---|
| P | 6.7 | 52.5 |
| P(collab) | 7.2 | 45.4 |
| PP(collab) | 29.0 | 12.5 |
| P(int collab) | 10.1 | 45.2 |
| PP(int collab) | 20.5 | 17.9 |
| TCS | 10.8 | 39.6 |
| MCS | 9.9 | 18.8 |
| TNCS | 5.7 | 55.5 |
| MNCS | 52.0 | 6.6 |
| P(top 10%) | 7.6 | 46.6 |
| PP(top 10%) | 13.7 | 14.7 |

### 4.3. Scopus AFID

The results for precision and recall of the Scopus AFID system, under the Mode A condition, are summarized in Table 10 and displayed in Figure 5. Precision is quite high, but, in contrast, recall is more moderate. Again, we find that the weighted mean precision and recall are slightly greater than the unweighted ones, suggesting that disambiguation quality is typically a little better for larger institutions. We also note that the coverage of our selected benchmarking institutions for the AFID system is 376 out of 445 (i.e., 85%) and therefore far from complete. Unlike the WoS OE system, the Scopus AFID system is not largely concentrated on universities (Table 2). Table 11 provides the five institutions with the lowest recall scores for the Scopus AFID system.

The direct comparison of the results for the indicator scores, calculated with the Scopus platform disambiguation system—AFID—on the one hand, and those calculated with the KB system on the other, in terms of distributions of percent deviation, are given in Table 12 and the deviation distributions are displayed in Figure 6. We find on average, for the absolute indicators, considerable shares of scores that are outside the range of accepted values. Relative indicators scores are less severely affected, but not within the accepted range often enough to be considered reliable. It is worth pointing out that in particular the total number of citations (TCS) is rarely within the allowed range, which, however, did not seem to overly affect the other citation indicators.

### 5. DISCUSSION

We have investigated the accuracy of bibliometric indicator values for German publicly funded research organizations that can be obtained through a search strategy on vendor-normalized data (for WoS) and through the use of the database vendors' proprietary institution disambiguation systems (for both WoS and Scopus). These indicator values were compared with results from a nearly complete and independent institutional disambiguation for which detailed performance characteristics were provided.
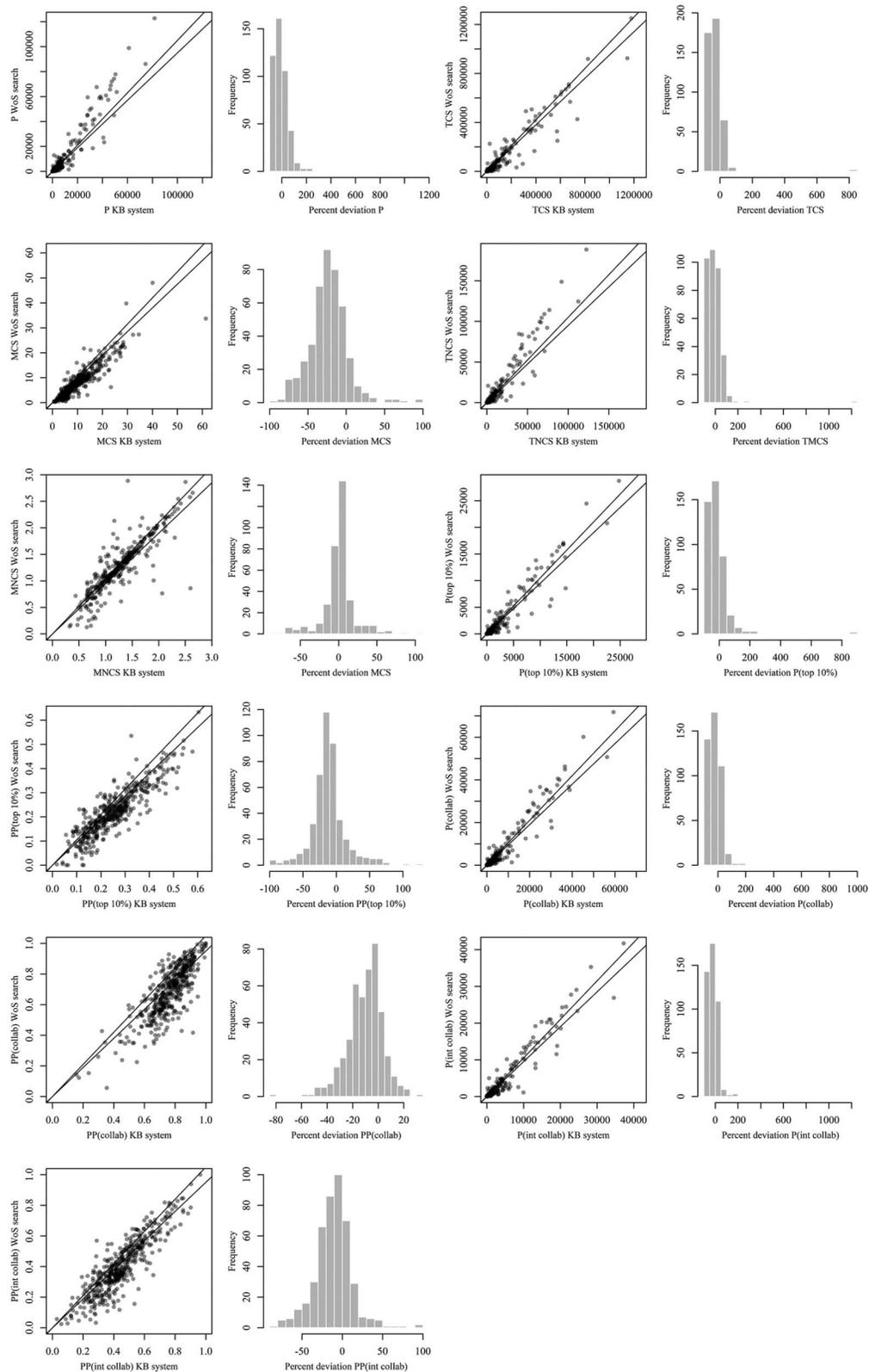
**Figure 4.** Distributions of indicator score deviations of WoS institution name search from KB system in Mode S. Diagonal lines indicate the ±5% error margin for indicator values.

**Table 10.** Summary statistics of the distributions of precision and recall of retrieved publications per institution for Scopus AFID ($n = 376$)

|  | Minimum | Median | Arithmetic mean | Weighted mean | Maximum | Standard deviation |
|---|---|---|---|---|---|---|
| **Precision** | 0.17 | 0.96 | 0.92 | 0.96 | 1.00 | 0.10 |
| **Recall** | 0.03 | 0.90 | 0.82 | 0.86 | 0.99 | 0.20 |

During our study, we found that conceptual differences between the three institution disambiguation systems and a lack of documentation of both the WoS OE system and the Scopus AFID system were obstacles to making straightforward comparisons. In particular, the definition of the basic institutional entity—which is a crucial point for comparing disambiguation systems—varied among the systems. For example, in Scopus, university hospitals were kept separate from university entities. They had different AFIDs, which were not connected in any way. This inhibits evaluations for universities including their academic hospitals or medical faculties. For a comparison with the KB system, these entities, academic hospitals and the universities to which they belong, had to be aggregated manually. A further issue was faced regarding the handling of predecessor institutions. In order to obtain valid results, we evaluated the systems on their own terms, adjusting the KB system as necessary, to include predecessor institutions. In WoS OE, the level at which institutional entities are defined (e.g., MPG as one single institutional entity), largely rules out a comparison on the institutional level, as defined in the KB system, for some KB sectors. Furthermore, there is no clear documentation on the handling of structural changes over time, such as splits or mergers. For analyses at the institutional level, this is a major limitation.

We find that WoS OE has the smallest coverage of our institution sample, at 20%, and is mainly restricted to universities. This reflects the choice made in WoS OE not to include the member institutes of nonuniversity research organizations. The coverage of Scopus AFID, on the other hand, is not largely limited to one institution type, but with 85%, it is also far from complete. These results show that the utility of the WoS and Scopus institution disambiguation systems for bibliometric analysis is limited, as they do not currently provide full coverage of disambiguated research organizations.
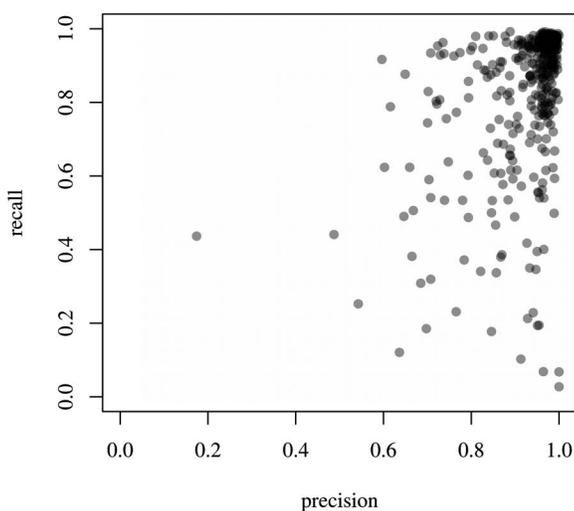


**Figure 5.** Precision and recall for Scopus AFID ($n = 376$).

**Table 11.** Institutions with low recall for Scopus AFID

| Institution name (KB system) | AFID and name | Recall |
|---|---|---|
| Rheinisch-Westfälisches Institut für Wirtschaftsforschung e.V. | 60029848: Rheinisch-Westfälisches Institut für Wirtschaftsforschung Essen | 0.03 |
| Technische Hochschule Mittelhessen—THM | 60000512: Technische Hochschule Mittelhessen | 0.07 |
| Fraunhofer MEVIS | 60107767: Fraunhofer-Institut für Bildgestutzte Medizin MEVIS | 0.07 |
| Hochschule Braunschweig/Wolfenbüttel, Ostfalia Hochschule für angewandte Wissenschaften | 60028838: Ostfalia Hochschule für angewandte Wissenschaften- Hochschule Braunschweig/Wolfenbüttel | 0.10 |
| Katholische Hochschule Nordrhein-Westfalen - Catholic University of Applied Sciences | 60006764: Katholische Fachhochschule Nordrhein-Westfalen | 0.12 |

In the WoS OE and Scopus AFID systems, precision of the obtainable publication sets was close to adequate levels at 0.95 and 0.96, respectively. However, neither system provided high recall rates (WoS: 0.93; Scopus: 0.86), which led to inaccurate indicator scores. Furthermore, we find substantial variation in precision and recall across institutions, indicating that within one system, these values are not systematically similar across the covered institutions but differ on a case-by-case basis. As for the tested name search method on normalized WoS data, precision and recall scores are poor, so this approach does not constitute a viable alternative.

Our results show that indicator values will typically not be within tolerable error margins at the organizational level, which we have set at ±5% of the reference value. This holds both for size-dependent and size-independent indicators. Hence, bibliometric indicator values at the institutional level have only limited accuracy.

Relying on vendor disambiguation systems may incur serious inaccuracies in indicator values at the institutional level. Therefore we conclude that for any use with policy implications,

**Table 12.** Deviation of indicator scores of Scopus AFID from KB system in Mode A (*n* = 376)

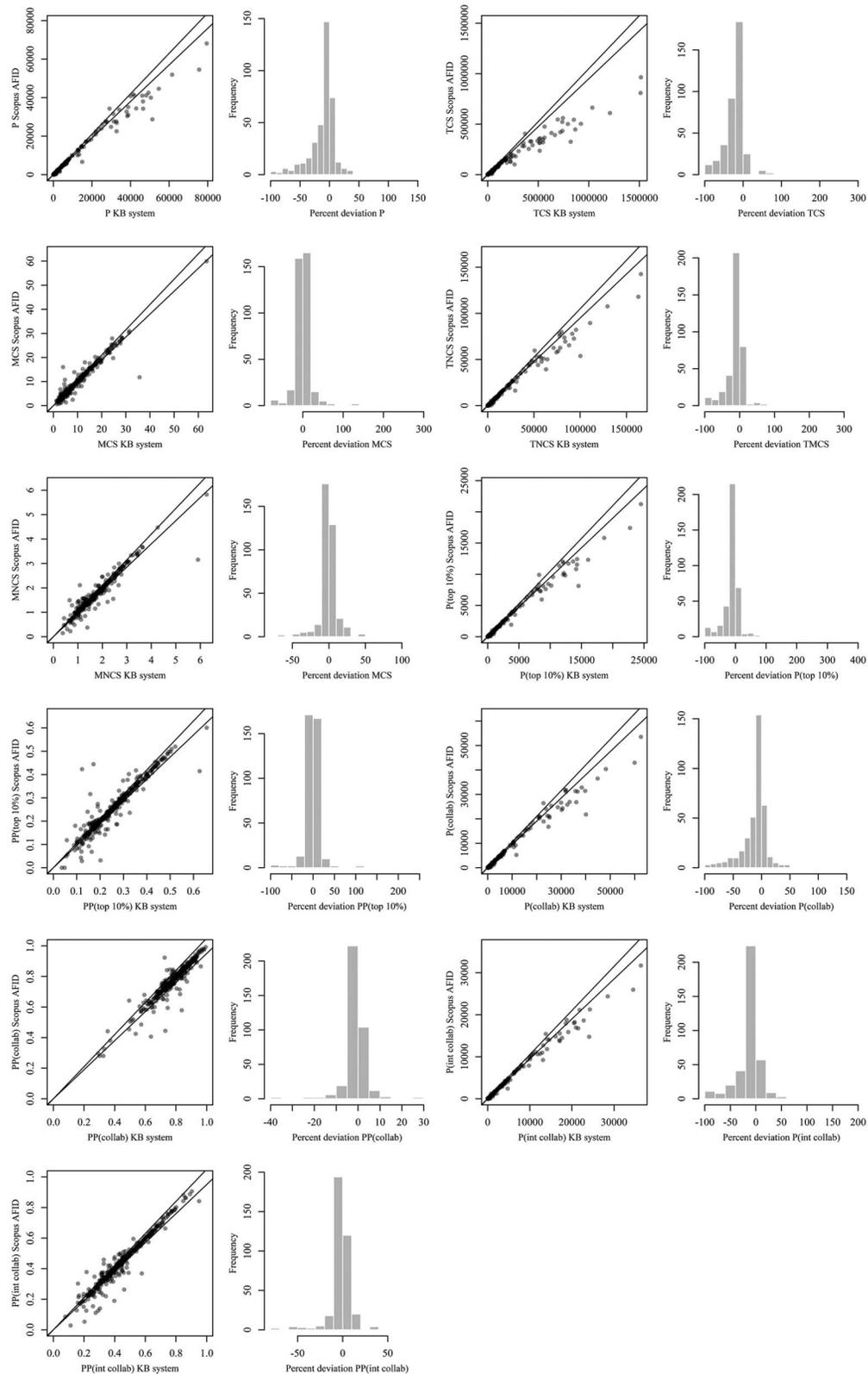| Indicator | Percent deviation within ±5% | Median absolute deviation (%) |
|---|---|---|
| P | 40.7 | 9.7 |
| P(collab) | 40.7 | 9.3 |
| PP(collab) | 86.7 | 1.4 |
| P(int collab) | 40.8 | 8.9 |
| PP(int collab) | 71.5 | 2.7 |
| TCS | 12.8 | 15.0 |
| MCS | 59.0 | 4.9 |
| TNCS | 41.0 | 9.7 |
| MNCS | 66.5 | 3.1 |
| P(top 10%) | 40.2 | 8.6 |
| PP(top 10%) | 68.1 | 3.1 |

**Figure 6.** Distributions of indicator score deviations of Scopus AFID from KB system in Mode A. Diagonal lines indicate the ±5% error margin for indicator values.

additional data cleaning for disambiguating affiliation data is recommended. We stress that any study such as the one presented in this paper shows only the current situation and that disambiguation systems may improve over time. The lack of adequate documentation of vendor institution disambiguation systems, including performance figures, is, however, another barrier impeding the adoption of these institution disambiguation systems in bibliometric studies.

## AUTHOR CONTRIBUTIONS

Paul Donner: Conceptualization, Methodology, Investigation, Formal Analysis, Visualization, Writing—original draft, Writing—review & editing, Funding Acquisition. Christine Rimmert: Conceptualization, Methodology, Investigation, Writing—original draft, Funding Acquisition. Nees Jan van Eck: Methodology, Investigation, Writing—review & editing, Supervision.

## COMPETING INTERESTS

Paul Donner applies the KB institution disambiguation system for WoS and Scopus data in research and commercial projects. Nees Jan van Eck is affiliated with the Centre for Science and Technology Studies (CWTS) at Leiden University, which has developed its own affiliation disambiguation system for WoS data and which uses this system in research and commercial projects.

## DATA AVAILABILITY

Data cannot be made available publicly due to the licensing contract terms of the original data.

## REFERENCES

Cuxac, P., Lamirel, J.-C., & Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, *97*(1), 47–58. https://doi.org/10.1007/s11192-013-1025-5

De Bruin, R. E., & Moed, H. F. (1990). The unification of addresses in specific publications. In L. Egghe & R. Rousseau (Eds.), *Informetrics 1989/90* (pp. 65–78). Amsterdam: Elsevier Science Publishers.

French, J. C., Powell, A. L., & Schulman, E. (2000). Using clustering strategies for creating authority files. *Journal of the American Society for Information Science*, *51*(8), 774–786.

Galvez, C., & Moya-Anegón, F. (2006). The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). *Scientometrics*, *69*(2), 323–345. https://doi.org/10.1007/s11192-006-0156-3

Galvez, C., & Moya-Anegón, F. (2007). Standardizing formats of corporate source data. *Scientometrics*, *70*(1), 3–26. https://doi.org/10.1007/s11192-007-0101-0

Huang, S., Yang, B., Yan, S., & Rousseau, R. (2014). Institution name disambiguation for research assessment. *Scientometrics*, *99*(3), 823–838. https://doi.org/10.1007/s11192-013-1214-2

Jiang, Y., Zheng, H. T., Wang, X., Lu, B., & Wu, K. (2011). Affiliation disambiguation for constructing semantic digital libraries. *Journal of the American Society for Information Science and Technology*, *62*(6), 1029–1041.

Jonnalagadda, S., & Topham, P. (2010). NEMO: Extraction and normalization of organization names from PubMed affiliation strings. *Journal of Biomedical Discovery and Collaboration*, *5*, 50–75.

Kronman, U., Gunnarsson, M., & Karlsson, S. (2010). *The bibliometric database at the Swedish Research Council—Contents, methods and indicators*. Stockholm: Swedish Research Council.

Moed, H., De Bruin, R., & Van Leeuwen, T. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, *22*(3), 381–422. https://doi.org/10.1007/BF02017338

Reyes-Elizondo, A., Calero-Medina, C., Visser, M., & Waltman, L. (2016). The challenge of identifying universities for a global university ranking [blog post]. Retrieved from https://www.cwts.nl/blog?article=n-q2w264

Rimmert, C., Schwechheimer, H., & Winterhager, M., (2017). Disambiguation of author addresses in bibliometric databases—technical report. Bielefeld: Universität Bielefeld, Institute for Interdisciplinary Studies of Science (I²SoS). Retrieved from https://pub.uni-bielefeld.de/publication/2914944

Swedish Research Council (2017). *The bibliometric database at the Swedish Research Council—Contents, methods and indicators. Version 2.1.* Retrieved from https://www.vr.se/download/18. 514d156f1639984ae0789d34/1529480565512/The+bibliometric+database+at+the+Swedish+Research+Council+-+contents,+methods+and+indicators.pdf

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C., Tijssen, R. J., Eck, N. J., …, Wouters, P. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419–2432. https://doi.org/10.1002/asi.22708

## APPENDIX A: EVALUATION OF THE KB DISAMBIGUATION SYSTEM FOR GERMAN RESEARCH INSTITUTIONS

A manual evaluation of the KB disambiguation system for German research institutions was conducted before the main study in order to assess its basic performance parameters and to inform any subsequent interpretations of results. The disambiguation quality is expressed in the information retrieval evaluation metrics of precision and recall. We begin by estimating the precision of the KB system for the data used in the study. In this task, precision is the proportion of correct document-address-combinations (assignments) to all assignments. From each result set of assignments produced by the KB disambiguation procedure for both WoS and Scopus data, a random sample of 1,000 assignments was checked manually for correctness of the mapped institution. The WoS sample contains 984 correct, 13 incorrect and three ambiguous assignments, leading to a precision of 0.9885 ± 0.0015. The Scopus sample contains 981 correct and 16 incorrect assignments and three ambiguous cases—leading to precision of 0.9825 ± 0.0015. We find precision to be similar for the two sets at approximately 98%.

We now consider the estimation of recall, which is more involved. Recall, in this task, is the proportion of correctly assigned addresses to all relevant addresses of an institution in the data. Because the real proportion of addresses belonging to the institutions considered here is unknown, it is not possible to directly calculate recall for the specific context of this analysis. Therefore, we first compute the recall for the total sets of German addresses in WoS and Scopus for publication years from 1996 until 2013. From this basis, we will extrapolate to the subset used in the study in a second step.

Overall recall is calculated as the number of assigned German document-address combinations divided by the number of all German document-address combinations. The exact figures are stated in Table A1.

The difference between WoS (0.91) and Scopus (0.82) in recall probably arises because the WoS addresses were processed with higher priority than the Scopus ones. Furthermore, there is an issue with incomplete address records. There are instances where only the town and country, but no institution name, are provided in an address. We found more of these cases in Scopus than WoS, but this preliminary observation requires further examination.

Because our main study only concerns a subset of all institutions, namely higher education institutions and nonuniversity research institutes, we are interested in an estimate of recall that is more specific to this subset. One way to obtain such an estimate is to look into the document-address combinations that were not successfully disambiguated and thus will not be retrieved. More specifically, we are interested in the share of records of the selected types of institutions among the nondisambiguated records. For this purpose, random samples of 100 addresses without assignments were checked manually, using both the WoS and Scopus data, to determine whether they belong to the research institutions of interest for this analysis. In the case of Scopus, 11 of 100 unassigned addresses belonged to institutions considered in this analysis; 86

**Table A1.** Assigned vs. all document-address combinations (recall), German addresses 1996–2013

| Document-address combinations of institutions (total set) | WoS | Scopus |
|---|---|---|
| All | 3,178,143 | 3,530,460 |
| Assigned to German research institutions in KB system | 2,891,537 (90.88%) | 2,908,500 (82.38%) |
| Overall recall | 0.91 | 0.82 |

addresses belonged to other institutions, with three ambiguous cases. The WoS sample contained 12 addresses of institutions of interest; 84 addresses belonged to other institutions and six cases were ambiguous. From these results, we obtain the following shares of the types of institutions of interest, in the set of not disambiguated records, including the ambiguous cases. WoS: 12 + 6 out of 100, a relative share of 0.18; Scopus: 11 + 3 out of 100, a relative share of 0.14. As the next step, we use these shares as factors for weighting the number of unassigned records to obtain estimates of the total numbers of records which cannot be recalled.

Recall for the selected set can then be estimated as the number of assigned document-address combinations in the selected set, divided by the number of all relevant document-address combinations in it. The unknown denominator—the total number of relevant records—is estimated on the basis of the sample described above: number of assigned addresses in the selected set + estimated number of addresses of interest in the set of unassigned addresses (derived from the sample).

We illustrate this calculation for the case of the WoS. The number of assignments for the types of institutions and publication years considered in the study is 2,211,795, which is a subset of the 2,891,537 total WoS assignments. This is the figure for the relevant recalled records—the numerator of recall. The number of relevant records—the denominator of recall—is not exactly known, but is estimated based on the checked sample. Of the 286,606 German address records not assigned to any institution, approximately 18% belong to the types of institutions considered in the study. Therefore, we arrive at an estimated number of relevant records of 2,263,384. The ratio of assigned relevant addresses to the estimated total relevant addresses is the estimated recall value specific to the publication set of the study; in this case, the value is 0.977.

The recall value for the Scopus data, following the same method, is 0.962, see Table A2.

**Table A2.** Assigned vs. all document-address-combinations (recall), selected set

| Document-address combinations of institutions in the selected set | WoS | Scopus |
|---|---|---|
| All (estimated) | 2,211,795 + (286,606 × 0.18) = **2,263,384** | 2,217,031 + (621,960 × 0.14) = **2,304,105** |
| Assigned | 2,211,795 | 2,217,031 |
| Estimated recall for the selected set | 0.977 | 0.962 |