

Vergabe von eindeutigen Autoren-IDs in WoS

Endbericht

Nadine Bethke, Rainer Frietsch, Sonia Gruber, Carolin Mund

Fraunhofer Institut für System- und Innovationsforschung

Breslauer Straße 48

76139 Karlsruhe

Januar 2016

1. Motivation

Wenn, wie im Falle des KB, Zugriff auf die beiden Datenbanken Web of Science und Scopus besteht, stellt sich bei jeder Projektbearbeitung die Frage, welche Datenbasis am besten für die anstehenden Aufgaben geeignet ist. Ein besonderer Vorteil von Scopus¹ ist die sogenannte author-id, die es ermöglicht, Personen in den Publikationsdaten zu verfolgen. Allerdings sind auch diese IDs nicht immer eindeutig und merge- und split-errors bekannt (Moed et al. 2013). Dennoch ist das Vorhandensein dieser ID häufig ausschlaggebend bei der Entscheidung für Scopus bei der Auswahl der Datenbank für einzelne Projekte. Es wird deshalb eine Erweiterung der von Thomson Reuters gelieferten Autoren-Daten um eine eindeutige ID vorgeschlagen, die die Auswahl der passenden Datenbank flexibler gestaltet.

Das Problem kann nicht durch eine einfache Integration der ORCID² gelöst werden. Die ORCID ist stark von Angaben und (aktuellen) Änderungen der Nutzer abhängig. Sie ist für Analysen erfahrener Wissenschaftler sicherlich ausreichend, aber viele bibliometrische Analysen zielen auch auf Nachwuchs- oder Kurzzeit-Wissenschaftler ab, bei denen eine Nutzung und Pflege der ORCID nicht unbedingt gegeben ist.

Ziel des hier dokumentierten Vorgehens ist es daher, mithilfe von in der Datenbank verfügbaren Informationen eine Zusammenführung von Publikationen (EIDs) eines Autors / einer Autorin unter einer eignen Autoren-ID. Hierzu werden anhand eines Testdatensatzes einzelne Indikatoren (Informationen innerhalb der Datenbank) bewertet. Zunächst wird hierzu ein so genannter Gold-Standard definiert, der als Vergleichsinformation mit "richtigen" Zusammenführungen von Autorennamen dient. Die einzelnen Indikatoren werden anschließend darauf getestet, wie gut sie diesen Gold-Standard replizieren können. Dazu werden die beiden Maße Recall (Anteil richtiger Treffer unter allen möglichen Treffern) und Precision (Anteil von richtigen Treffern unter allen Treffern) berechnet und dargestellt. Auf dieser Basis wird eine Reihenfolge von zu verwendenden Informationen generiert, die am Ende dann in ihrem Zusammenspiel ebenfalls am Gold-Standard gemessen werden. Hierdurch erhält man eine Bewertung der zu erwartenden Abdeckung und der Güter der Vorgehensweise. Diese kann dann unter der Annahme der Repräsentanz auf die gesamte Datenbank angewendet werden.

2. Durchführung

In der KB-Datenbank sind bereits vereinheitlichte Schreibweisen der Autorennamen vorhanden. Diese werden genutzt, um Autoren mit ähnlichen Namen, d.h. Namen, die sich gleichen (in einem weiteren Schritt kann auch zugelassen werden, dass sie sich bspw. bis auf ein Zeichen ähneln) zu identifizieren. Solche Autorenpaare, im Folgenden „Zwillinge“ genannt, können genutzt werden, um verschiedene Einträge zu der gleichen Person zu sammeln.

Das Verfahren geht zu Beginn davon aus, dass es sich bei allen Autoren um unterschiedliche Instanzen handelt. Bei einem solchen bottom-up Approach werden Instanzen so lange als unterschiedlich behandelt, bis ein Merge-Kriterium greift. Ein solches Merge-Kriterium hängt von verschiedenen bibliometrischen (Meta-)Daten ab, die in diesem Projekt evaluiert werden sollen.

¹ Neben anderen, s. auch Kulkarni 2009

² Vgl. <http://orcid.org/>

Das Verfahren durchläuft die folgenden Schritte, um Zwillinge zu identifizieren:

1. Es werden Homonyme identifiziert. Hierzu werden die standardisierten Schreibweisen in der KB-Tabelle genutzt. Ein Trainingsdatensatz hilft zu entscheiden, welche Schreibweisen zusammengefasst werden sollen und welche nicht. Email-Adressen geben zusätzliche Hinweise zu Zwillingen mit unterschiedlicher Namensschreibweise. In einem weiteren Schritt, den wie an dieser Stelle nicht explizit dokumentieren, werden auch unterschiedliche Namensschreibweisen zugelassen, indem sie über die Levenshtein Distance hinsichtlich leichter Schreibvarianten zusammengeführt werden.
2. Als nächstes erfolgt eine Ähnlichkeitsberechnung der Zwillinge auf Basis der folgenden Kontextinformationen, also alle, die das Kriterium erfüllen, werden unter einer ID zusammengeführt. Dies wird jeweils nur für diejenigen durchgeführt, die nicht bereits zugeordnet wurden. Die einzelnen Kriterien sind:

1. Namensmatching:

Wir matchen Vor- und Nachname, sowie Nachname mit abgekürztem Vornamen. In einem weiteren, in der folgenden Analyse nicht explizit dokumentierten Schritt, lassen wir eine Abweichung im Namen zu, bsp.: Muller, Müller.

Die so erhaltenen Zwillinge dienen als Datenbasis für alle weiteren Schritte. Bezogen auf die Namensvariationen sind noch viele Fehler enthalten (Müller, Meller und Miller gehören jetzt zusammen, sind aber mit hoher Wahrscheinlichkeit nicht dieselbe Person). Diese Fehler werden mit Hilfe der folgenden Vergleichsmerkmale so weit wie möglich reduziert.

2. Emailmatching:

Exakt gleiche E-Mailadressen gehören zu ein und derselben Person (ohne weitere Merkmale zu betrachten).

3. Affiliation³:

Haben Zwillinge die gleiche Affiliation, dann sind sie mit hoher Wahrscheinlichkeit die gleiche Person. Umgekehrt würde dieses Kriterium allerdings nicht funktionieren, denn Personen können schließlich im Lauf ihrer wissenschaftlichen Karriere die Affiliation wechseln. Da wir aber zunächst von grundsätzlich distinkten Instanzen starten und nur jene Instanzen zusammen führen, die auch tatsächlich ein Kriterium erfüllen, können wir die Affiliationen hier verwenden. Mit anderen Worten bedeutet das, wenn zu einem Namenszwilling die gleiche Affiliation vorliegt, dann gehen wir davon aus, dass es sich um die gleiche Person handelt. Umgekehrt schließen wir aber nicht aus, dass ein und die selbe Person unterschiedliche Affiliationen hat.

4. Referenzen (Bibliographic coupling):

Es wird ausgezählt, wie viele gleiche Publikationen die Zwillinge zitieren. Da die Anzahl von Zitaten sehr unterschiedlich ist und 2 von insgesamt 3 gleichen Zitierungen „mehr wert“ sind als

³ Hierzu müssen die Adressangaben der Autoren in Web of Science bereinigt werden. Derzeit ist in unserer Implementierung der WoS-Datenbank die Information der Affiliation und des Autors / der Autorin nicht eindeutig verknüpft, weshalb dieses Kriterium keine sonderlich hohe Präzision erreicht (s.u.). In zukünftigen Umsetzungen werden wir dies jedoch berücksichtigen.

4 von 10, wird der Anteil von Zitaten gleicher Publikationen an Zitate aller Publikationen berechnet und nicht etwas die absolute Zahl verwendet. Es wird dann ein Schwellwert bestimmt, ab dem das Kriterium erfüllt ist. Wir verwenden hier drei Schwellwerte in absteigender Form, nämlich 90%, 60% und 30%.

5. Selbstzitate:

Zitieren sich die Zwillinge gegenseitig, so ist dies ein weiterer Hinweis auf die gleiche Person.

6. Co-Autorenetzwerke:

Wie bei den Referenzen wird auch hier der Anteil von gemeinsamen Co-Autoren an allen Co-Autoren betrachtet. Auch hier kann auf Grund von hohen Co-Autoren-Zahlen auf einzelnen Papieren bzw. in einzelnen Disziplinen eine Verzerrung auftreten. Wir bewerten daher die Kooperation anhand der Gesamtzahl der Ko-Autoren. Auch hier verwenden wir die Schwellwerte 90%, 60% und 30% absteigend.

3. Trainingsdatensatz (Gold-Standard)

Zur Bewertung der gesamten Autorendisambiguierung, aber auch zur Beurteilung der Passfähigkeit und Sinnhaftigkeit der einzelnen Schritte der Disambiguierung ist ein Trainingsdatensatz notwendig – ein so genannter Gold-Standard.

Zur Entwicklung eines solchen Gold-Standards ist eine externe Quelle mit Informationen zu sicher zusammengeführten Autoren bzw. deren jeweiligen Publikationen notwendig. Eine Möglichkeit zur Erarbeitung eines Gold-Standards wäre es, die Publikationslisten von Universitätsmitarbeitenden oder Mitarbeitenden von Forschungseinrichtungen herunterzuladen und die jeweils genannten Publikationen in der Datenbank WoS zu identifizieren. Dieses Verfahren wäre allerdings einerseits aufwändig und andererseits könnte so nur eine überschaubare Zahl an Publikationen erfasst werden. Hinzu kommt, dass keine Möglichkeit der Überprüfung auf Vollständigkeit der Listen besteht. Fehlende Publikationen würden einem anderen Autor zugeordnet. Daneben besteht das Sprachproblem bspw. bei asiatischen Autoren, die ggf. die Publikationslisten (teilweise) in ihren Landessprachen auf den Websites haben. Weiterhin könnten schließlich nur jene Autoren erfasst werden, die auch tatsächlich eine Publikationsliste veröffentlicht haben. In einigen Forschungseinrichtungen ist es aber nicht Usus, die Mitarbeitenden nach außen auszuflaggen, wodurch auch keine Möglichkeit der Veröffentlichung von Publikationslisten besteht.

Eine zweite Möglichkeit besteht in der Nutzung bestehender Zusammenführungen. Hier bietet sich zunächst Google Scholar an. Allerdings ist die Qualität der Zusammenführung nicht überprüfbar sowie ggf. die Unterscheidung zwischen einem konsolidierten und einem nicht-konsolidierten Autorenprofil in Google Scholar nicht möglich. Über die Vollständigkeit der in Google Scholar erfassten Publikationen eines Autors bestehen daneben auch keine Informationen. Hinzu kommt, dass eine Zusammenführung der WoS und der Google Scholar Daten aufwändig wäre.

Als weitere Datenquelle mit zusammengeführten Autoreninformationen böte sich auch Scopus an. Dort sind die Publikationen eines Autors unter einer Autoren-ID realisiert. Allerdings basiert die Zusammenführung in Scopus auf einem Algorithmus bzw. verschiedenen automatischen Prozessschritten, die Elsevier durchführt. Eine explizite Qualitätskontrolle der Ergebnisse ist dabei

nicht möglich. Würden wir nun also die Autoren-IDs aus Scopus verwenden, so würden wir ggf. bestehende Fehler bzw. Ungenauigkeiten übernehmen, diese zum Gold-Standard erheben, und schließlich unsere eigene Vorgehensweise möglichst bzgl. diesen ggf. fehlerhaften Benchmarks zu optimieren versuchen. Dies wäre nicht wünschenswert. Daneben verbieten die Lizenzverträge mit Elsevier bzw. Thomson Reuters eine Zusammenführung der beiden Datenbanken. Eine solche Datenbank wäre jedoch auf Basis der document ID (DOI) leicht umsetzbar.

Es könnte eine Vielzahl weiterer Datenbanken wie bspw. die von der GESIS für ihre Analysen verwendete Dagstuhl-Datenbank für den IT-Bereich verwendet werden. Da dies aber nur einen kleinen Ausschnitt der WoS darstellt und zudem auch hier ein großer Aufwand bei der Zusammenführung mit WoS anfallen würde, haben wir uns für eine Datenbank-interne Lösung entschieden.

In der Datenbank sind sowohl die ORCID als auch die R_ID enthalten. Beides sind Systeme, die die Autorinnen und Autoren selbst "füttern und pflegen", d.h. die jeweiligen Autoren führen ihre Publikationen unter einer ID zusammen. Die Daten haben damit eine hohe Sicherheit bei der Zuordnung. Allerdings zeigt sich in der Nutzung der Informationen, dass sie nicht immer aktuell und vollständig sind. Es finden sich einige Veröffentlichungen ohne R_ID, die mit hoher Wahrscheinlichkeit einem Autor / einer Autorin mit einer R_ID zugeordnet werden können.

In Tabelle 1 sind die Anteile der Publikationen, die einer Person mit ORCID bzw. R_ID zugeordnet werden können, für die Publikationsjahre ab 2000 abgebildet. Von den jährlichen Publikationen haben ca. 7% eine ORCID, während ca. 17-20% eine R_ID aufweisen. Die Anteile steigen in beiden Fällen über die Zeit an, was in der erst kurzen Implementierungszeit der beiden Systeme begründet ist. Ältere, nicht mehr aktive Wissenschaftlerinnen und Wissenschaftler sind somit weniger stark abgedeckt.

Tabelle 1: Anteile der Publikationen mit ORCID und R_ID nach Publikationsjahren

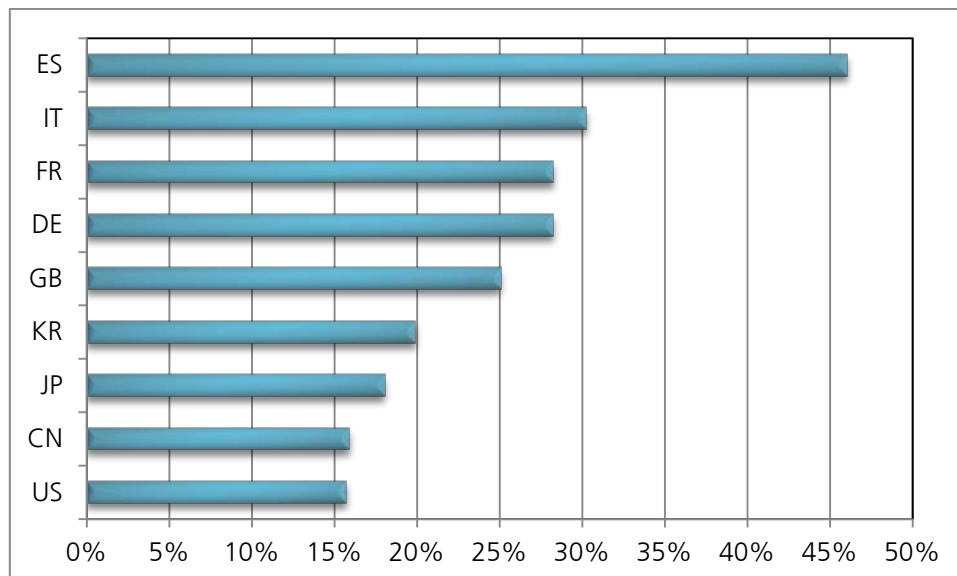
Jahr	# Publ.	ORCID	R_ID	Anteil ORCID	Anteil R_ID
2000	1324824	58293	168744	4.4%	12.7%
2001	1307705	63276	181902	4.8%	13.9%
2002	1356228	61154	177701	4.5%	13.1%
2003	1412127	76221	220897	5.4%	15.6%
2004	1504920	85821	249037	5.7%	16.5%
2005	1606270	95012	275945	5.9%	17.2%
2006	1681989	104979	303312	6.2%	18.0%
2007	1821413	120653	347926	6.6%	19.1%
2008	1921694	130883	375014	6.8%	19.5%
2009	2010166	142886	397675	7.1%	19.8%
2010	1978632	147041	402654	7.4%	20.4%
2011	2045593	156160	405912	7.6%	19.8%
2012	2172518	168573	385922	7.8%	17.8%
2013	2239152	158636	317949	7.1%	14.2%
2014	2108968	95373	174807	4.5%	8.3%

Quelle: Thomson-Reuters WoS; Berechnungen des Fraunhofer ISI

Es bietet sich somit das Thomson-Reuters eigene System der ResearcherID als Trainingsdatensatz bzw. als Gold-Standard an. Eine Zusammenführung von ORCID und R_ID macht keinen Sinn, da das gleiche Problem wie bei der generellen Autorendisambiguierung besteht. Welche tatsächlich zusammen gehören ist nicht zu identifizieren. Man könnte zwar über den Überlapp von Publikationen jeweils die ORCID und die R_ID einer Person ermitteln und dann die nicht durch R_ID erfassten Publikationen mit ORCID zuspähen. Da aber einerseits ein hohes Maß an Überlappung erwartet werden kann – beide Systeme beruhen auf aktivem Handeln der Autoren – und andererseits 300-400 Tausend Veröffentlichungen pro Jahr eine mehr als ausreichende Zahl an Daten sind, aus denen wir unseren Gold-Standard extrahieren können, haben wir uns dazu entschlossen, ausschließlich die R_ID zu verwenden.

Wie Abbildung 1 zeigt sind die Anteile nach unterschiedlichen Sprachen ähnlich und jedenfalls hinreichend abgedeckt, mit hohen Anteilen in Spanien – was zum Teil auch den relativ niedrigen Gesamtzahlen geschuldet sein kann. Großbritannien deckt 25% ab, während in den USA, das sicher eine größere sprachliche Heterogenität aufweist als Großbritannien, immerhin noch 15% erfasst sind.

Abbildung 1: Anteile der Publikationen mit R_ID nach Ländern, Publikationsjahr 2012



Quelle: Thomson-Reuters WoS; Berechnungen des Fraunhofer ISI

Es gilt anzumerken, dass auch die hier verwendete R_ID in der Datenbank keine wirklich fehlerfreie Information liefert. Bei der Arbeit mit den Daten hat sich gezeigt, dass durchaus Publikationen enthalten sind, die – nach manueller Durchsicht einiger Beispieldaten – zu einer bestehenden R_ID zuzuordnen wären, in der Datenbank aber die entsprechende Zuordnung fehlt. Grund ist vermutlich, dass nicht alle Autoren ihr R_ID-Profil permanent pflegen, so dass Publikationen nicht oder noch nicht vom jeweiligen Autor identifiziert und zugeordnet wurden. Erahnen kann man dies durch Folgendes: In unserem Testdatensatz von den 3122 R_IDs des Gold-Standards tritt bei 243 der Fall auf, dass der gleiche Name mit der gleichen E-Mail-Adresse bei einer anderen Publikation genannt wird.

Dieses Problem kann für die Erarbeitung des Gold-Standards an dieser Stelle nicht gelöst werden und muss daher akzeptiert werden. Das bedeutet aber, dass der Recall über und Precision unterschätzt werden. Wie hoch diese Über- bzw. Unterschätzung ist, kann allerdings nicht berechnet werden.

4. Ergebnisse

4.1 Bewertung der Einzelindikatoren/-kriterien

Ein besonderes Augenmerk bei der Vergabe der IDs liegt auf der Präzision des Verfahrens, so dass merge-errors möglichst ausgeschlossen werden. Aus diesem Grund haben wir uns für eine Kaskade von Indikatoren/Kriterien entschieden, die wir entsprechend ihrer Präzision geordnet nacheinander für die Zuordnung der Autoren-IDs verwenden.

In Tabelle 2 sind zunächst Recall und Precision in Bezug auf den Gold-Standard-Datensatz dargestellt. In diesem Gold-Standard-Datensatz sind ausschließlich Publikationen von Personen mit einer R_ID enthalten. Die Werte für die einzelnen Indikatoren/Kriterien zeigen an, wie viele von den zu treffenden Veröffentlichungen getroffen wurden (Recall) und wie viele davon richtig sind (Precision). Die Indikatoren sind nach ihrer Precision absteigend geordnet.

Eine sehr hohe Präzision, bei niedriger Trefferquote, liefern die beiden Kriterien der Ko-Autorenschaft von jeweils mindestens 0,9 bzw. 0,6. Auch die Email-Adresse liefert eine hohe Präzision, bei allerdings nicht allzu großen Recall. Der Grund ist, dass nur wenige Veröffentlichungen überhaupt eine Emailadresse haben und dort häufig auch nur der korrespondierende Autor. Es folgen die Anteile gemeinsame Referenzen von 0,3 noch vor 0,6 und 0,9. Diese Reihenfolge lässt sich mit den niedrigen Fallzahlen von hohen Überschneidungen erklären. Hier sind die Nachkommastellen bei der Precision nicht hinreichend stabil. Die Reihenfolge zeigt aber auch, dass bereits eine niedrige Überschneidung der Referenzen zu hoher Qualität bei der Zuordnung führt. Allerdings ist auch hier der Recall bei diesem Indikator für sich alleine genommen sehr niedrig. Eine Überschneidung der Referenzen von 0,9 als Kriterium führt schließlich zu keiner sichtbaren Ausweitung der Treffermenge.

Tabelle 2: Bewertung der Indikatoren bezogen auf den von GESIS definierten Goldstandard von 1000 Namen mit R_ID

Kriterium	Recall	Precision
Anteil der Ko-Autoren $\geq 0,9$	0,05	1,00
Anteil der Ko-Autoren $\geq 0,6$	0,15	1,00
Email-Adresse	0,17	0,99
Anteil gemeinsamer Referenzen $\geq 0,3$	0,05	0,99
Eigenzitate	0,05	0,99
Anteil gemeinsamer Referenzen $\geq 0,6$	0,01	0,99
Anteil der Ko-Autoren $\geq 0,3$	0,29	0,98
Anteil gemeinsamer Referenzen $\geq 0,9$	0,00*	0,94
Affiliation	0,86	0,89
Subject Categories	0,49	0,56
Klasse (27 Felder)	0,74	0,44

* Der Grund, dass der Recall beim Anteil gemeinsamer Referenzen von mehr also 90% gleich Null ist zeigt, dass es so gut wie keine Publikationen gibt, bei denen eine solche hohe Übereinstimmung von Referenzlisten zu finden ist. Der sinkende und auch niedrigere Precision- Wert allerdings bezieht sich auf die wenigen Publikationen und hier ist dann die Fallzahl so gering, dass selbst eine falsche Zuordnung direkt einen deutlichen Ausschlag im Indikatorwert erzeugt.

Die Eigenzitate, d.h. in den Referenzlisten eines Autors genannte Veröffentlichungen mit dem gleichen Autorennamen (Homonym) tragen bei hoher Präzision nur wenig zur Treffermenge bei. Die Affiliation hingegen ist ein sowohl bei Recall als auch Precision hoher Wert, zumindest für sich alleine genommen. Die Information, dass die Publikation der selben Klasse (Disziplin) angehört, erzeugt zwar einen hohen Recall, allerdings bei sehr bescheidener Präzision. Dies bedeutet, dass zumindest innerhalb der 27 Klassen der hier verwendeten Zeitschriftenabgrenzung eine Vielzahl an gleichnamigen Autoren innerhalb einer Klasse vorkommt.

Nimmt man statt der 1000 Namen mit R_ID zusätzlich jeweils 20 Publikationen mit entsprechenden Homonymen hinzu, dann sinkt die Präzision deutlich. Der Recall ist in dieser Analyse nicht sinnvoll interpretierbar, da für die hinzugenommenen Veröffentlichungen keine Informationen zur „wahren“ Zuordnung vorliegen. Die Analyse soll aber zeigen, wie deutlich die Präzision sinkt, wenn – wie es in der gesamten WoS-Datenbank der Fall ist – eine größere Variation an Homonymen vorhanden ist. Bei allen Kriterien ist die Präzision deutlich niedriger.

Tabelle 3: Bewertung der Indikatoren bezogen auf den von GESIS definierten Goldstandard von 1000 Namen mit R_ID plus 20 Veröffentlichungen mit jeweiligen Homonymen (ohne R_ID)

Kriterium	Recall**	Precision
Anteil der Ko-Autoren $\geq 0,9$	0,17	0,77
Anteil der Ko-Autoren $\geq 0,6$	0,05	0,63
Email-Adresse	0,05	0,62
Anteil gemeinsamer Referenzen $\geq 0,3$	0,05	0,59
Eigenzitate	0,01	0,58
Anteil gemeinsamer Referenzen $\geq 0,6$	0,15	0,58
Anteil der Ko-Autoren $\geq 0,3$	0,29	0,56
Affiliation*	0,86	0,50
Anteil gemeinsamer Referenzen $\geq 0,9$	0,00	0,47
Subject Categories	0,49	0,28
Klasse (27 Felder)	0,74	0,19

* In der derzeitigen Version unserer WoS-Datenbank ist der eindeutige Link zwischen Autor und Affiliation, wie er in WoS seit dem Publikationsjahr 2008 möglich ist, bisher nicht implementiert. Daher sind die Affiliationen hier nicht eindeutig. Bei Herstellung des Links ist mit einer Verbesserung der Precision zu rechnen.

** Der Recall ist auf Grund der nicht im Gold-Standard erfassten zusätzlichen Veröffentlichungen nicht sinnvoll zu interpretieren.

4.2 Bewertung der Kombination der Indikatoren/Kriterien

Im Folgenden werden die einzelnen Indikatoren aufeinander aufbauend kombiniert und es werden jeweils die resultierenden Werte für Recall und Precision berechnet. Die verwendeten Kombinationen (Schritte) sind im Einzelnen:

- Schritt 1: Email-Adresse
- Schritt 2: Anteil der Ko-Autoren $\geq 0,9$
- Schritt 3: Anteil der Ko-Autoren $\geq 0,6$
- Schritt 4: Anteil gemeinsamer Referenzen $\geq 0,9$
- Schritt 5: Eigenzitate
- Schritt 6: Anteil gemeinsamer Referenzen $\geq 0,6$
- Schritt 7: Anteil der Ko-Autoren $\geq 0,3$
- Schritt 8: Anteil gemeinsamer Referenzen $\geq 0,3$
- Schritt 9: Affiliation
- Schritt 10: Subject Categories
- Schritt 11: Klasse (27 Felder)

Tabelle 4: Recall und Precision bei kumulativer Nutzung der Indikatoren in Bezug auf den Gold-Standard

Name	Recall	Precision
Schritt 1 (Email)	0,17	0,99
Schritt 2 (Ko-Autoren 0,9)	0,25	0,99
Schritt 3 (Ko-Autoren 0,6)	0,36	0,99
Schritt 4 (Referenzen 0,9)	0,36	0,99
Schritt 5 (Eigenzitate)	0,45	0,99
Schritt 6 (Referenzen 0,6)	0,45	0,99
Schritt 7 (Ko-Autoren 0,3)	0,62	0,97
Schritt 8 (Referenzen 0,3)	0,64	0,97
Schritt 9 (Affiliation)*	0,95	0,69
Schritt 10 (Subject Categories)	0,99	0,29
Schritt 10 (Klasse)	0,99	0,23

* In der derzeitigen Version unserer WoS-Datenbank ist der eindeutige Link zwischen Autor und Affiliation, wie er in WoS seit dem Publikationsjahr 2008 möglich ist, bisher nicht implementiert. Daher sind die Affiliationen hier nicht eindeutig. Bei Herstellung des Links ist mit einer Verbesserung der Precision zu rechnen.

Die Ergebnisse der der kumulierten Nutzung der Indikatoren/Kriterien in Bezug auf den Gold-Standard sind in Tabelle 4 dargestellt. Es zeigt sich, dass bis einschließlich Schritt 8 eine hohe und zufriedenstellende Präzision erreicht werden kann, während mit Schritt 9 diese dann deutlich sinkt. Dies bedeutet, dass wir in unserem Ansatz nach Schritt 8 den Prozess der Vergabe von Autoren IDs beenden würden. Leider erreicht das Verfahren derzeit lediglich einen Recall von 0,64, d.h. lediglich ca. 64% der zuzuordnenden Autorennamen wurden auch tatsächlich zugeordnet. Dies ist sicherlich kein befriedigender Wert.

Tabelle 5: Recall und Precision bei kumulativer Nutzung der Indikatoren in Bezug auf den Gold-Standard plus 20 Veröffentlichungen mit jeweiligen Homonymen (ohne R_ID)

Name	Recall**	Precision
Schritt 1 (Email)	0,17	0,77
Schritt 2 (Ko-Autoren 0,9)	0,25	0,86
Schritt 3 (Ko-Autoren 0,6)	0,36	0,66
Schritt 4 (Referenzen 0,9)	0,36	0,65
Schritt 5 (Eigenzitate)	0,45	0,63
Schritt 6 (Referenzen 0,6)	0,45	0,63
Schritt 7 (Ko-Autoren 0,3)	0,62	0,57
Schritt 8 (Referenzen 0,3)	0,64	0,57
Schritt 9 (Affiliation)*	0,95	0,36
Schritt 10 (Subject Categories)	0,99	0,15
Schritt 10 (Klasse)	0,99	0,09

* In der derzeitigen Version unserer WoS-Datenbank ist der eindeutige Link zwischen Autor und Affiliation, wie er in WoS seit dem Publikationsjahr 2008 möglich ist, bisher nicht implementiert. Daher sind die Affiliationen hier nicht eindeutig. Bei Herstellung des Links ist mit einer Verbesserung der Precision zu rechnen.

** Der Recall ist auf Grund der nicht im Gold-Standard erfassten zusätzlichen Veröffentlichungen nicht sinnvoll zu interpretieren.

In Tabelle 4 sind die entsprechenden Werte für Recall und Precision angegeben, wenn nicht nur der Gold-Standard, sondern zusätzlich auch noch 20 Veröffentlichungen je Homonym hinzugenommen werden. Dies kommt dann zumindest bezüglich der Precision der tatsächlich zu erwartenden Precision in der Gesamtdatenbank näher. Der Recall ist hier nicht interpretierbar.

Guten Gewissens könnte man hier nur bis Schritt 2 gehen, also die Emailadressen und die Ko-Autorenschaften von 0,9 nutzen. Die niedrige Precision liegt vor allem daran, dass bei der Berechnung solche Matches als "falsch" eingeschätzt wurden, die mit sehr hoher Wahrscheinlichkeit richtig sind.

Ob hier unser Verfahren weniger geeignet ist und der Ansatz von GESIS eine höhere Präzision erreicht, kann auf Basis der uns vorliegenden Daten nicht beurteilt werden. Es wäre sicherlich wünschenswert, auch für diesen Ansatz eine entsprechende Einschätzung abgeben zu können.

5. Zusammenfassung und Ausblick

In diesem Projekt wurde versucht, mithilfe der in WoS vorhandenen Informationen eine Zusammenführung von Publikationen eines Autors / einer Autorin unter einer einheitlichen ID zu erreichen. Hierzu wurden 11 Indikatoren / Kriterien mit einem Gold-Standard-Datensatz verglichen, um deren Eignung anhand von Recall und Precision zu bewerten. Anschließend wurden die Indikatoren entsprechend ihrer Präzision "hintereinander geschaltet". Es hat sich dabei gezeigt, dass man bei hinreichend hoher Präzision mit diesem Verfahren derzeit lediglich einen Recall, also eine Abdeckung von etwa 64% erreicht. Hier schneidet der Ansatz der GESIS

mit einem gewichteten Verfahren, etwas besser ab. Allerdings bietet der Gold-Standard, den GESIS verwendet hat, nur eine eingeschränkte Aussage über die Präzision, wenn das Verfahren denn auf die gesamte Datenbank angewendet wird. Hier wäre sicherlich eine weitere Prüfung hilfreich. Wir konnten mit unserem erweiterten Gold-Standard-Datensatz jedenfalls belegen, dass die Präzision deutlich sinkt, wenn die Zahl der Variationsmöglichkeiten steigt.

Für die Zukunft könnte bei unserem Verfahren an zwei Stellen eine Verbesserung erreicht werden. Einerseits kann erwartet werden, dass durch die eindeutige Zuordnung der Affiliationen zu Autoren die Precision dieses Indikators erhöht wird, so dass auch der Gesamt-Recall bei zufriedenstellender Gesamt-Precision erhöht werden kann – zumindest für die Publikationsjahre ab 2008. Andererseits kann durch die Implementierung der Namensvariationen eine breitere Abdeckung erreicht werden. Wenn also bspw. Variationen auf Grund von Umlauten berücksichtigt werden wie z.B. Müller und Muller, dann könnte die Zuordnung von Publikationen zu Autoren-IDs ebenfalls erhöht werden – möglicherweise bei sinkender Precision. Dies wäre zu prüfen.

Eine erweiterte Bewertung und zusätzlich eine größere Variationsmöglichkeit könnte dadurch erreicht werden, dass nicht nur die Stichprobe mit 1000 Namen, sondern alle Veröffentlichungen mit einer R_ID verwendet werden. Auf Basis der in Tabelle 1 dargestellten Daten wären dies zwischen 13% und 20% aller Veröffentlichungen eines Publikationsjahres. Dies Schritt könnte man durchführen, ehe man die Umsetzung für alle Veröffentlichungen in der Datenbank angeht. Man könnte darauf aufbauend dann auch Analysen der Zuverlässigkeit des Ansatzes für einzelne Länder machen. Es ist davon auszugehen, dass unser Ansatz insgesamt eine höhere Präzision bei deutschen Autorinnen und Autoren erreicht als beispielsweise bei chinesischen oder koreanischen Autorinnen und Autoren, wo deutlich höhere absolute wie relative Anteile von Homonymen vorliegen.

Eine Umsetzung unseres Ansatzes für die gesamte Datenbank wäre technisch problemlos möglich. Hierzu würde man dann jedoch zunächst die R_ID als gesetzt annehmen und die übrigen Publikationen ohne Zuordnung einer R_ID entweder diesen R_IDs oder neuen IDs zuordnen. Inwiefern dies mit dem Ansatz von GESIS, der sicherlich mehr Rechner-Aufwand verlangt, möglich ist, können wir nicht beurteilen. Insgesamt zeigt sich aber, dass hinsichtlich der Abdeckung zum jetzigen Zeitpunkt der Ansatz von GESIS weitreichender und zufriedenstellender zu sein scheint.

6. Referenzen

- Gurney, Thomas; Horlings, Edwin; van den Besselaar, Peter (2012): Author disambiguation using multi-aspect similarity indicators. In *Scientometrics* 91 (2), pp. 435–449. DOI: 10.1007/s11192-011-0589-1.
- Kulkarni, Abhaya V. (2009): Comparisons of Citations in Web of Science, Scopus, and Google Scholar for Articles Published in General Medical Journals. In *JAMA* 302 (10), p. 1092.
- Moed, H.F./Aisati, M./Plume, A. (2013): Studying scientific migration in Scopus, *Scientometrics*, March 2013, 94, 929-942.