



Teil I: Kurzbericht

Komparative Analyse und Kuratierung Deutscher Metadaten in Offenen Bibliometriedaten (KB); Teilprojekt: Datenbereitstellung und Autor:innendisambiguierung

Förderkennzeichen: 16WIK2301B

1 Ursprüngliche Aufgabenstellung und Stand der Forschung

Ziel des Vorhabens war der Aufbau einer offenen Bibliometriedatenbank innerhalb des Kompetenzzentrums Bibliometrie (KB), wobei dieses Teilprojekt einen gesonderten Fokus auf die Analyse und Bereitstellung von Dokumententypen, Open Science Evidenzen und Zitationsdaten hatte. Die Grundlage der Kuratation bildete die Datenbank OpenAlex, die einen umfangreichen Datenbestand und einen kostenfreien Zugang bietet. Die mit diesem Projekt verbundene Öffnung des KBs hin zu offenen Bibliometriedaten sollte zudem eine Alternative zu proprietären Datenbanken wie Web of Science und Scopus unterstützen, deren Nutzung mit erheblichen Lizenzkosten einhergeht. Zunächst wurden im Projekt Abdeckungsanalysen durchgeführt, die den gegenwärtigen Stand in Open-Alex mit den aktuell genutzten Datenquellen im KB Scopus und Web of Science vergleichen. In einem nächsten Schritt sollte OpenAlex im KB vollumfänglich zur Verfügung gestellt und mit kuratierten Daten angereichert werden. Gleichzeitig sollten die kuratierten Daten einem breiteren Publikum im Sinne des Open Data Prinzips zugänglich gemacht werden. Das KB konnte dabei auf jahrelange Erfahrung mit bibliometrischen Datenbanken zurückgreifen.

Das Teilprojekt bei GESIS konzentriert sich auf Autor:innendisambiguierung. Die präzise Unterscheidung von Autorennamen (Author Name Disambiguation, AND) stellt für bibliometrische Datenbanken weiterhin eine Herausforderung dar. Unterschiedliche Gepflogenheiten hinsichtlich der Namensreihenfolge, Synonymie (bedingt durch Publikationspraktiken und Namensvariationen) sowie Polysemie und Homonymie können dazu führen, dass verschiedene Schreibweisen desselben Autors beim Abgleich von Datensätzen fälschlicherweise nicht als identisch erkannt werden. Gegenwärtig stärken manuell kuratierte Systeme – wie sie etwa von OpenAlex, Scopus, DBLP und Web of Science betrieben werden – sowie unabhängige Initiativen wie ORCID das Bewusstsein innerhalb der wissenschaftlichen Gemeinschaft für die Bedeutung einer präzisen Namensunterscheidung. Ziel dieses Projekts war es daher, die Unterscheidung von Autorennamen innerhalb des Kompetenznetzwerks zu verstehen und zu optimieren, indem relevante und effektive Methoden auf Daten aus großen bibliografischen Datenbanken wie Web of Science, Scopus und OpenAlex angewandt wurden.

2 Ablauf des Vorhabens

Den Auftakt des Projekts bildete eine gemeinsame Publikation der KB-Partner unter der Leitung von Jack Culbert (GESIS); daraus entstand der in der Fachzeitschrift *Scientometrics* veröffentlichte Artikel „Reference Coverage“. Dieser Beitrag hat sich zu einem zentralen Beitrag für den Vergleich von Umfang und bibliometrischen Eigenschaften der drei wichtigsten bibliometrischen Datenbanken entwickelt – sowohl innerhalb als auch außerhalb der bibliometrischen Fachgemeinschaft. Die Studie vergleicht verschiedene Attribute hinsichtlich der Datenqualität und Vollständigkeit von Open-Access-Publikationen und unterstützt damit das Projektziel, die Qualitätssicherung zu verbessern und das Verständnis für die Grenzen von OpenAlex als Alternative zu proprietären bibliometrischen Datenbanken zu schärfen.

Darüber hinaus beteiligte sich GESIS an einer von Nick Haupka (Göttingen) geleiteten Studie zum Vergleich von Publikations- und Dokumenttypen, bei der die Klassifizierungen von Publikationen in fünf Open-Access- und proprietären Datenbanken gegenübergestellt wurden. Die Studie nutzte ein von Göttingen eingerichtetes und verwaltetes System – eine Google-BigQuery-Instanz, die eine Kopie der durch KB-Aktivitäten angereicherten OpenAlex-Daten beherbergt – in Kombination mit einer vom FIZ Karlsruhe gehosteten und verwalteten PostgreSQL-Infrastruktur. Parallel dazu zeigte sich im Austausch mit KB-Nutzern und der breiteren bibliometrischen Fachwelt, dass der Datenaustausch zu den zugrundeliegenden Studien durch Zugangsvereinbarungen eingeschränkt war. Daher wurde die Initiative ergriffen, ein Tool namens „Open Research Converter“ zu entwickeln und zu veröffentlichen. Dieses ermöglicht die massenhafte Umwandlung öffentlicher Identifikatoren in OpenAlex-IDs unter Einhaltung von API-Beschränkungen und Berücksichtigung der Datenherkunft (Provenance) für OpenAlex – Funktionen, die über die direkte Nutzung der API oder der Benutzeroberfläche nicht verfügbar waren. Ein Demonstrator hierfür wurde auf der Infrastruktur von GESIS bereitgestellt, und ein Systemartikel wurde beim *Journal of Open Source Software* eingereicht.

Zudem führte die Zusammenarbeit Yoed Kenett (Technion) zu zwei Publikationen, die ein neuartiges textbasiertes Maß für Kreativität untersuchen – insbesondere dessen Vorhersagekraft für die Zitationshäufigkeit unter Kontrolle anderer bibliometrischer Variablen. Ein Beitrag zu diesem laufenden Forschungsvorhaben wurde für die „20th International Conference on Scientometrics and Informetrics“ angenommen, und ein vollständiger Forschungsartikel wurde bei *Scientometrics* eingereicht (derzeit im Begutachtungsverfahren). GESIS trug zudem zur Open-Data-Veröffentlichung bei, indem das Institut eine spezifische Lösung zur Disambiguierung von Autorennamen für Autoren mit deutscher Affiliation entwickelte, implementierte und bereitstellte. Zu den entwickelten Modellen gehörten Varianten auf Basis von Graph-Neuronalen Netzen sowie klassische Verfahren des maschinellen Lernens; die Ergebnisse des schließlich ausgewählten Modells wurden im Rahmen der beiden Tabellen auf Autoren- und Werkebene in der Open-Data-Veröffentlichung bereitgestellt. Die Abfragen für das Modellschema wurden im GitHub-Repository der Wissensdatenbank (KB) veröffentlicht, ebenso wie eine dauerhafte Version der Daten im Zenodo-Open-Data-Datensatz, der über die drei veröffentlichten Versionen hinweg 9.701-mal heruntergeladen wurde.

3 Wesentliche Ergebnisse (und ggf. die Zusammenarbeit mit anderen Forschungseinrichtungen)

Im Rahmen des Hauptziels dieses Projekts haben wir zwei Ergebnisse erarbeitet: den OpenKB-Datensatz sowie das zugehörige Modell zur Disambiguierung von Autorennamen und die OpenKB-Graphdatenbank (gehostet vom FIZ Karlsruhe). Letztere unterstützt künftige Vorhaben im Bereich groß angelegter Netzwerkanalysen – etwa zu Zitations- oder Koautorenschaften – sowie die Entwicklung von Modellen für maschinelles Lernen und statistische Analysen. Das OpenKB-Modell zur Disambiguierung von Autorennamen konzentriert sich auf Autorenschaften mit deutscher Affiliation und basiert auf einem Clustering-Heuristik-Verfahren.

- Ausgewählte Publikationen:
 - Jack H. Culbert, Anne Hobert, Najko Jahn, *et al.* Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics* **130**, 2475–2492 (2025).
<https://doi.org/10.1007/s11192-025-05293-3>
 - Nick Haupka, Jack H. Culbert, Alexander Schniedermann, Najko Jahn, Philipp Mayr; Analysis of the publication and document types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. *Quantitative Science Studies* **7**, 179–194 (2026). doi:
<https://doi.org/10.1162/QSS.a.406>
 - Jack H. Culbert, Yoed N. Kenett, & Philipp Mayr. Investigating the originality of scientific papers across time and domain: A quantitative analysis. Arxiv, (2025)
<https://doi.org/10.48550/arXiv.2512.07892>
 - Submitted to *Scientometrics* December 2025, under review



Teil II: Eingehende Darstellung

Komparative Analyse und Kuratierung Deutscher Metadaten in Offenen Bibliometriedaten (KB); Teilprojekt: Datenbereitstellung und Autor:innendisambiguierung

Förderkennzeichen: 16WIK2301B

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter den Förderkennzeichen 16WIK2301B gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei dem/der Autor/in bzw. den Autor/inn/en.

Inhaltsverzeichnis

1	Aufgabenstellung	2
2	Wissenschaftlicher und technischer Stand, an den angeknüpft wurde	2
3	Planung und Ablauf des Vorhabens	4
4	Inhaltliche Ergebnisse	4
5	Wichtigste Positionen des zahlenmäßigen Nachweises	5
6	Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten	5
7	Voraussichtlicher Nutzen, Verwertbarkeit der Ergebnisse und zukünftige Planungen im Sinne des Verwertungsplans	5
8	Fortschritt auf dem Gebiet des Vorhabens während der Durchführung des Vorhabens bei anderen Stellen	6
9	Erfolgte oder geplante Veröffentlichungen der Projektergebnisse	7
10	Literaturverzeichnis	8

Gliederung

1 Aufgabenstellung

Ziel des Vorhabens war der Aufbau einer offenen Bibliometriedatenbank innerhalb des Kompetenz-zentrums Bibliometrie (KB), wobei dieses Teilprojekt einen gesonderten Fokus auf die Analyse und Bereitstellung von Dokumententypen, Open Science Evidenzen und Zitationsdaten hatte. Die Grundlage der Kuration bildete die Datenbank OpenAlex, die einen umfangreichen Datenbestand und einen kostenfreien Zugang bietet. Die mit diesem Projekt verbundene Öffnung des KBs hin zu offenen Bibliometriedaten sollte zudem eine Alternative zu proprietären Datenbanken wie Web of Science und Scopus unterstützen, deren Nutzung mit erheblichen Lizenzkosten einhergeht.

Zunächst wurden im Projekt Abdeckungsanalysen durchgeführt, die den gegenwärtigen Stand in OpenAlex mit den aktuell genutzten Datenquellen im KB Scopus und Web of Science vergleichen. In einem nächsten Schritt sollte OpenAlex im KB vollumfänglich zur Verfügung gestellt und mit kuratierten Daten angereichert werden. Gleichzeitig sollten die kuratierten Daten einem breiteren Publikum im Sinne des Open Data Prinzips zugänglich gemacht werden. Das KB konnte dabei auf jahrelange Erfahrung mit bibliometrischen Datenbanken zurückgreifen.

Das Teilprojekt bei GESIS konzentriert sich auf Autor:innendisambiguierung. Die präzise Unterscheidung von Autorennamen (Author Name Disambiguation, AND) stellt für bibliometrische Datenbanken weiterhin eine Herausforderung dar. Unterschiedliche Gepflogenheiten hinsichtlich der Namensreihenfolge, Synonymie (bedingt durch Publikationspraktiken und Namensvariationen) sowie Polysemie und Homonymie können dazu führen, dass verschiedene Schreibweisen desselben Autors beim Abgleich von Datensätzen fälschlicherweise nicht als identisch erkannt werden. So könnte beispielsweise der Name „Max Hermann Schmidt“ als „Schmidt, M.H.“, „Schmidt M.“, „Schmidt, Max H.“, „Schmidt, Max.“ usw. abgekürzt werden. Diese Varianten müssen zusammengeführt werden; zudem erfordern Fälle, in denen eine Person ihren Namen ändert (etwa durch Heirat) oder in denen ein „Max Hienert Schmidt“ im selben Fachbereich tätig ist, eine präzise Unterscheidung und Zuordnung. Gegenwärtig stärken manuell kuratierte Systeme – wie sie etwa von OpenAlex, Scopus, DBLP und Web of Science betrieben werden – sowie unabhängige Initiativen wie ORCID das Bewusstsein innerhalb der wissenschaftlichen Gemeinschaft für die Bedeutung einer präzisen Namensunterscheidung. Ziel dieses Projekts war es daher, die Unterscheidung von Autorennamen innerhalb des Kompetenznetzwerks zu verstehen und zu optimieren, indem relevante und effektive Methoden auf Daten aus großen bibliografischen Datenbanken wie Web of Science, Scopus und OpenAlex angewandt wurden.

2 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Der Stand der Technik bei der Autorennamendisambiguierung zum Zeitpunkt der Konzeption des Projekts befand sich an einem Punkt, an dem der Umfang der Disambiguierung zu einer großen Herausforderung für traditionelle AND-Modelle wurde, da offene Ressourcen wie der Microsoft Academic Graph und OpenAlex im akademischen Bereich zunehmend an Bedeutung gewannen. Die erste angegangene Herausforderung war die Skalierbarkeit: Explorative Arbeiten zur Auswahl minimaler Variablen für Disambiguierungsmodelle führten zum Aufbau und Hosting der Graphrepräsentationen von OpenAlex, Web of Science und Scopus durch GESIS und FIZ Karlsruhe.

Dies geschah teilweise im Rahmen eines Skalierbarkeitstests für auf wissenschaftlichen Wissensgraphen basierende Modelle zur Autorennamensdisambiguierung, wie RF-LRC (Rehs 2021), LAND (Santini et al. 2022) oder DND (Chen et al. 2023). Die meisten dieser Modelle wurden auf kleineren Datensätzen entwickelt, etwa der AMiner-Familie oder benutzerdefinierten Datensätzen aus DBLP, CiteSeerX und Web of Science. Die Entwicklung einer universell einsetzbaren Autorennamensdisambiguierung stellte sowohl aufgrund der Größe als auch wegen des Mangels an verlässlichen Ground-Truth-Daten in OpenAlex, Scopus und Web of Science eine Herausforderung dar – rechtliche Datenbeschränkungen verhinderten eine Kreuzvalidierung zwischen den Datensätzen während des Modelltrainings und zur Laufzeit sowie die Erstellung eines gemeinsamen Datensatzes, der die Daten zusammenführt.

Daher scheiterten erste Versuche des Modelltrainings aufgrund unzureichender Rechenressourcen. Bemühungen, Zugang zu großskaligen Rechenressourcen im wissenschaftlichen Rechencluster zu erhalten, sowie weitere Versuche blieben erfolglos. Daraufhin verlagerte sich der Fokus auf die Konzeption und Erstellung eines eingeschränkten, aber relevanten Datensatzes, auf dem die lokale Entwicklung eines Modells innerhalb der zeitlichen Projektbeschränkungen realistisch abgeschlossen werden konnte.

Dies führte zu einer Neugestaltung des Modells nach dem Vorbild heuristikbasierter Clustering-Modelle wie SAND (Ferreira et al. 2014). Dabei wurde ein vereinfachtes Modell mit Blocking nach Nachname und erstem Initial entwickelt, kombiniert mit einem ablativen heuristischen Auswahlprozess auf einem Datensatz, der auf Grundlage einer Übereinkunft der KB-Technik-Gruppe ausgewählt wurde, wonach ein auf in Deutschland affilierte Autorinnen und Autoren fokussierter Datensatz den größten Nutzen bieten würde. Dies führte zur Veröffentlichung der Autorentabellen im Open Data Release im Dezember 2025.

GESIS hat zu den Disseminationsbemühungen des KB beigetragen, indem Blogbeiträge über die Erfolge des Reference Coverage Paper verfasst wurden und Ergebnisse, Daten, Demonstrator-Veröffentlichungen und Projekterfolge über akademische Kanäle auf BlueSky, LinkedIn und X verbreitet wurden.

Als Ergebnis dieser Disseminationsaktivitäten lud Frédérique Sachwald im April 2024 Jack Culbert ein, die Ergebnisse des Reference-Coverage-Papers in einem Seminar vorzustellen, das sich mit der Evaluierung von OpenAlex als potenzieller Ersatzdatenbank für bibliometrische Studien befasste.

Ein zentrales Vorhaben innerhalb des Projekts war die Entwicklung eines Modells zur Autorennamensdisambiguierung. Die daraus hervorgegangene Forschung führte zu einer Graphdatenbankrepräsentation der Verknüpfungen des Modells, die nun den KB-Partnern zugänglich ist, sowie zu den disambiguierten Autorentabellen im Open Data Release.

GESIS hat zudem eine Arbeitsbeziehung zu leitenden Mitgliedern des OpenAlex-Teams aufgebaut und gepflegt, darunter Kyle Demes, und an technischen Diskussionen mit ehemaligen Teammitgliedern wie Jason Portenoy teilgenommen.

3 Planung und Ablauf des Vorhabens

Der Projektvorschlag umfasste zwei Arbeitspakete: F.5.1 – die Anreicherung der bibliometrischen Rohdaten der KB mit persistenten Identifikatoren – und F.5.2 – die Ergänzung um Open-Access-Referenzen sowie Studien zum Datenvergleich.

F.5.1.

Im Rahmen dieses Arbeitspakets sollten die im Projekt zu berücksichtigenden Verfahren zur Disambiguierung von Autorennamen anhand von vier Kriterien hinsichtlich ihrer Praxistauglichkeit bewertet werden: Verständlichkeit, Allgemeingültigkeit, Effizienz und Anpassbarkeit. Die aus dem Modell resultierenden Daten wurden in den regulären Prozess zur Datenübernahme in die Wissensbasis (KB) integriert; zudem wurde ein Datenstand im Rahmen der Open-Data-Veröffentlichung festgehalten, während ergänzende Informationen in Tabellen der PostgreSQL-Datenbank aufgenommen wurden. Diese Arbeiten erfolgten in Zusammenarbeit mit dem FIZ Karlsruhe, das beim Aufbau und Hosting der auf Graphdatenbanken basierenden Repräsentation der Wissensbasis (unter Einbeziehung der Datenquellen OpenAlex, Web of Science und Scopus) unterstützte.

F.5.2

Diese gemeinsam mit dem Göttinger Team durchgeführte Arbeit zur Anreicherung der KB mit Open-Access-Referenzen mündete in mehrere Publikationen, darunter die beiden Vergleichsstudien von Culbert et al. (2025) und Haupka et al. (2026). In diesen Arbeiten wurden Vergleiche der Granularitätsebene von Publikationen und Zeitschriften zwischen den Datenbanken OpenAlex, Web of Science und Scopus (sowie Semantic Scholar und PubMed in der Studie von Haupka et al. 2026) angestellt. Die Ergebnisse dieser Studien wurden zudem im Rahmen von Verbreitungsmaßnahmen, etwa in Blogbeiträgen zur KB, näher erläutert.

4 Inhaltliche Ergebnisse

Unsere Forschung umfasste die Untersuchung verschiedener Architekturen zur Disambiguierung von Autorennamen, darunter neuartige, auf Graphen und heuristischem Clustering basierende Verfahren; letztere erwiesen sich hinsichtlich vier praktischer Bewertungskriterien als am effektivsten. Daraus resultierten zwei primäre Ergebnistabellen sowie fünf sekundäre Tabellen mit Zwischenergebnissen und Metriken. Das Blocking-Verfahren, das auf dem Nachnamen und dem Anfangsbuchstaben des Vornamens basiert, erzeugte auf dem kleineren, deutschsprachig geprägten Datensatz über 869 Millionen Autorenpaare; das daraus resultierende Modell erreichte auf einem zurückgehaltenen Testdatensatz eine Genauigkeit von 96,94%. Die beiden Haupttabellen bieten unterschiedliche Perspektiven auf die Ergebnisse: Eine Tabelle aggregiert die Daten nach Autoren und enthält den vom Modell ausgegebenen Sicherheitsscore, während die andere nach Werken aggregiert und so den Datenzugriff für Studien erleichtert, die mit Werk-IDs arbeiten.

Arbeitsbereich 2: Datenbankvergleich

In diesem Arbeitspaket wurde die Datenbank OpenAlex mit den proprietären Datenbanken Scopus und Web of Science hinsichtlich des Publikationsvolumen, Anzahl der Referenzen sowie weiterer Kriterien verglichen. Hierbei zeigt sich, dass OpenAlex eine vergleichbare Anzahl an *source-Referenzen* wie die etablierten Datenbanken Scopus und Web of Science hat.

source-Referenzen sind Referenzen, bei denen die Referenz selbst auch in der jeweiligen Datenbank als Publikation enthalten ist. Der Vergleich zeigte auch, dass die Abdeckung von Förderinformationen in den klassischen Datenbanken ausgeprägter ist als bei OpenAlex. Die Abdeckung von ORCID-Identifiern ist wiederum ausgeprägter in OpenAlex.

- Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., Mayr, P. (2025). Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*.

In einer weiteren Abdeckungsanalyse zu den Publikations- und Dokumenttypen in OpenAlex wurden große Diskrepanzen bei der Klassifikation jener Entitäten innerhalb verschiedener Datenbanken festgestellt. Konkret wurde hier OpenAlex mit den Datenbanken Web of Science, Scopus, PubMed und Semantic Scholar verglichen. Dabei zeigte sich, dass zwar bei den Publikationstypen in OpenAlex, Web of Science und Scopus grobe Überschneidungen existieren, jedoch deutliche Unterschiede bei der Dokumenttypenklassifikation zwischen den Datenbanken existierte, da OpenAlex fast durchgängig Dokumente als Artikel klassifizierte, während andere Datenbanken wie Web of Science oder PubMed differenzierter vorgehen und auch Reviews und klinische Studien zuordneten.

- Haupka, N., Culbert, J., Schniedermann, A., Jahn, N., Mayr, P. (2026). Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. *Quantitative Science Studies*.

5 Wichtigste Positionen des zahlenmäßigen Nachweises

Die bewilligten Personalmittel sind wie geplant verausgabt worden.

6 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Das Vorhaben ist von hoher Relevanz für die bibliometrische Landschaft in Deutschland, aber auch international. Die Bereitstellung von offenen Daten im KB erleichtert den Zugang und die Arbeit mit Bibliometriedaten, wodurch Analysen und Entscheidungsprozesse gestärkt werden können. Gleichzeitig bieten die erstellten Abdeckungsanalysen Transparenz beim Umgang und Umfang der neuen Datenquelle OpenAlex. Die im Projekt entwickelten Verfahren zur Datenkuratierung sind öffentlich verfügbar und ermöglichen weiterführende Studien und können zudem für spezifische Anwendungen nachgenutzt werden.

Alle Arbeiten im Projekt sind im ursprünglichen Projektantrag konzeptionell beschrieben worden. Zum Zeitpunkt des Antrags existierten andere Versionen der Datenbasis OpenAlex; aus diesem Grund wurden gesonderte Arbeiten notwendig, diese konnten aber in das Projektvorhaben eingegliedert werden.

7 Voraussichtlicher Nutzen, Verwertbarkeit der Ergebnisse und zukünftige Planungen im Sinne des Verwertungsplans

Die im Projekt entstandenen Ergebnisse, insbesondere die Publikationen, Datensätze, und Tools stellen für die deutsche, aber auch internationale Community wichtige Bausteine zum besseren Verständnis von offener Bibliometrie bereit. Die Nutzung der im Projekt erzielten

Ergebnisse, gemessen in Zitationen und Downloadzahlen der entstandenen publizierten Studien, zeigen ihre Relevanz und Anschlussfähigkeit. Bei der Erstellung der Publikationen wurde berücksichtigt, die zugrundeliegenden Daten zu dokumentieren und frei zur Verfügung zu stellen.

Im Projekt konnte OpenAlex als zusätzliche Datenquelle im KB etabliert werden, womit eine Alternative zu kostenpflichtigen Angeboten wie Scopus und Web of Science angeboten werden kann. Die im Projekt geleisteten Abdeckungsanalysen legen dabei Stärken und Schwächen des neuen Service offen und können Nutzenden helfen können zwischen OpenAlex und proprietären Datenquellen abzuwägen. Gleichzeitig helfen die im Projekt erstellten Datenkurationen dabei, das Angebot von OpenAlex gezielt zu verbessern.

8 Fortschritt auf dem Gebiet des Vorhabens während der Durchführung des Vorhabens bei anderen Stellen

Während der Projektlaufzeit gab es umfangreiche Änderungen am Datenbestand seitens OpenAlex. Insbesondere die Klassifikation von Dokumententypen und das AND-System wurde durch OpenAlex erheblich optimiert. Unter anderem wurden Dokumenttypen aus PubMed in OpenAlex integriert, was zu einer Verbesserung bei der Identifizierung von Editorials und Letters in OpenAlex geführt hat. Des Weiteren hat OpenAlex weitere Dokumenttypen implementiert, darunter Reviews. Während der Projektlaufzeit wurden zudem neue Systeme zur Disambiguierung von Autorennamen implementiert; dazu gehörten ein Modell des maschinellen Lernens, eine verbesserte Einbindung von ORCID und anderen Quellen für Autoren Daten sowie nachfolgende Maßnahmen zur besseren Kuratierung der Autoren Daten durch manuelle Optimierung. Diese Entwicklungen wurden im Projekt durch Blogposts aufgegriffen und dokumentiert.

Auch von externer Seite wurden Analyseergebnisse bezüglich der Dokumenttypen in OpenAlex veröffentlicht. Mongeon et al. (2025) hat Diskrepanzen bei der Klassifikation von Dokumenttypen zwischen OpenAlex und Web of Science untersucht und dabei festgestellt, dass OpenAlex bei der Klassifikation von Forschungsartikeln und Reviews zu Missklassifikationen neigt. Ein Vergleich der Dokumenttypen in OpenAlex und Scopus wurde von Alperin, Portenoy, Demes, Larivière und Haustein (2024) durchgeführt, wobei hier auf die unterschiedlichen Klassifikationssysteme in beiden Datenbanken hingewiesen wird. So benutzt OpenAlex den Dokumenttyp Artikel als Auffangbegriff für alle Dokumente, die sich keinem speziellen Dokumenttyp im Klassifikationssystem von OpenAlex zuordnen lassen. Auch Maisano, Mastrogiacomo, Ferrara und Franceschini (2025) weisen auf Genauigkeitsprobleme bei der Klassifikation von Dokumenttypen in bibliometrischen Datenbanken hin. Gleichzeitig stellen sie ein Verfahren vor, mit dessen fehlerhafte Dokumenttypenklassifikationen identifiziert werden können.

In Zusammenarbeit mit der Technion University hat GESIS zudem zur Entwicklung neuartiger Indikatoren für die bibliometrische Analyse beigetragen – im Einklang mit den übergeordneten Zielen des KB-Netzwerks. Dies betrifft insbesondere die beiden Arbeiten von Jack H. Culbert, Yoed N. Kenett und Philipp Mayr (2025a, 2025b), in denen die Verwendung von DSI als Metrik zur Bestimmung des Neuheitsgrads wissenschaftlicher Publikationen sowie dessen Korrelation mit der Zitationshäufigkeit und anderen bibliometrischen Variablen untersucht wird.

Im Laufe des Projekts hat GESIS zudem mehrere Datensätze erstellt.

Das Thema „offene Bibliometriedaten“ hat im Projektzeitraum deutlich an Bedeutung und Aufmerksamkeit gewonnen. Offene Daten fördern Transparenz, Nachvollziehbarkeit und Reproduzierbarkeit wissenschaftlicher Forschung, indem sie freien Zugang zu Forschungsdaten ermöglichen. Sie stärken die kollaborative Wissenschaft, beschleunigen Innovationen und ermöglichen es, Erkenntnisse effizienter und breiter innerhalb der wissenschaftlichen Gemeinschaft und darüber hinaus zu teilen. Insbesondere GESIS hat durch die Entwicklung und Bereitstellung des „Open Research Converter“ zu diesem Vorhaben beigetragen. Dieses Tool ermöglicht den offenen Austausch bibliometrischer Daten innerhalb der Wissenschaftsgemeinschaft, indem es Digital Object Identifiers (DOIs) mit OpenAlex-Identifikatoren verknüpft; dies wiederum erlaubt die reproduzierbare Berechnung bibliometrischer Kennzahlen sowie die gemeinsame Nutzung von Datensätzen im Zusammenhang mit wissenschaftlichen Publikationen. Die Ergebnisse dieser Arbeit wurden auf einschlägigen bibliometrischen Konferenzen vorgestellt und haben so eine breite Akzeptanz in der Fachwelt gefördert.

9 Erfolgte oder geplante Veröffentlichung der Projektergebnisse

- Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., Mayr, P. (2025). Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*. <https://doi.org/10.1007/s11192-025-05293-3>
- Haupka, N., Culbert, J., Schniedermann, A., Jahn, N., Mayr, P. (2025). Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. *Quantitative Science Studies*.
- Jack H. Culbert, Yoed N. Kenett, & Philipp Mayr. Investigating the originality of scientific papers across time and domain: A quantitative analysis. Arxiv, (2025) <https://doi.org/10.48550/arXiv.2512.07892> Submitted to Scientometrics December 2025, under review
- Jack H. Culbert, Yoed N. Kenett & Philipp Mayr. "Originality in scientific titles and abstracts can predict citation count." (2025). ISSI 2025 Conference, Yerevan, Armenia. https://doi.org/10.51408/issi2025_106

Further materials:

- Open Research Converter (ORC) ([Paper](#), [demonstrator](#), [codebase](#))
- Indo-German Literature Dataset ([Zenodo](#))
- Open AI Literature 2010-2020 Dataset ([Zenodo](#))
- Hcéres invited talk at seminar for Observatoire des Sciences et Techniques, 28th February 2024, ([Zenodo](#))
- Culbert, J.H., Mayr, Philipp. (2025, June 23-27) Automating Reproducible Bibliometrics with the Open Research Converter. ISSI 2025 Conference, Yerevan, Armenia. ([Proceedings](#))
- Jack H. Culbert, Philipp Mayr. Open Research Converter. Global Tech Mining workshop at Science and Technology Indicators conference 2024, Berlin, Germany. ([Zenodo](#))
- Jack H. Culbert, Philipp Mayr. The Open Research Converter. Nordic Workshop on Bibliometrics and Public Policy 2024, Reykjavik, Iceland. ([Zenodo](#))

- Jack H. Culbert. Applied Graph Techniques for Bibliometrics. Kompetenznetzwerk Bibliometrie Network day 2024, Cologne, Germany. ([Zenodo](#))

10 Literaturverzeichnis

- Ferreira, A.A., Gonçalves, M.A., Laender, A.F. (2020). SAND: Self-Training Author Name Disambiguator. In: Automatic Disambiguation of Author Names in Bibliographic Repositories. Synthesis Lectures on Information Concepts, Retrieval, and Services. Springer, Cham. https://doi.org/10.1007/978-3-031-02322-4_5
- Santini, C., Gesese, G.A., Peroni, S. *et al.* A knowledge graph embeddings based approach for author name disambiguation using literals. *Scientometrics* **127**, 4887–4912 (2022). <https://doi.org/10.1007/s11192-022-04426-2>
- Chen, Y., Jiang, Z., Gao, J. *et al.* A supervised and distributed framework for cold-start author disambiguation in large-scale publications. *Neural Comput & Applic* **35**, 13093–13108 (2023). <https://doi.org/10.1007/s00521-020-05684-y>
- Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the Web of Science. *Journal of Informetrics*, **15**(3), 101166. <https://doi.org/10.1016/j.joi.2021.101166>
- Alperin, J. P., Portenoy, J., Demes, K., Larivière, V. & Haustein, S. (2024). An analysis of the suitability of OpenAlex for bibliometric analyses. arXiv. <https://10.48550/arXiv.2404.17663>
- Maisano, D. A., Mastrogiacomo, L., Ferrara, L. & Franceschini, F. (2025). A large-scale semi-automated approach for assessing document-type classification errors in bibliometric databases. *Scientometrics*. <https://doi.org/10.1007/s11192-025-05244-y>
- Mongeon, P., Hare, M., Krause, G., Marjoram, R., Riddle, P., Toupin, R. & Wilson, S. (2025). Investigating Document Type Discrepancies between OpenAlex and the Web of Science. Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI. <https://doi.org/10.29173/cais1943>