Nick Haupka*

Analyse der Abdeckung wissenschaftlicher Publikationen auf Semantic Scholar im Kontext von Open Access

https://doi.org/10.1515/bfp-2023-0057

Zusammenfassung: In diesem Beitrag soll der Service Semantic Scholar untersucht und der Frage nachgegangen werden, inwiefern sich Informationen des Services für das OA-Publikationsmonitoring an Bibliotheken und Hochschulen nachnutzen lassen. Ausgehend von einer Vollerhebung von Zeitschriftenartikeln der Jahre 2000 bis 2020 wird der OA-Anteil innerhalb verschiedener wissenschaftlicher Disziplinen bestimmt. Ferner soll eine Aussage bezüglich der Qualität der OA-Nachweise in Semantic Scholar getroffen werden.

Schlüsselwörter: Open Access; Semantic Scholar; Publikationsmonitoring; Datenanalyse; Big Scholarly Data

Analysis of the Coverage of Scientific Publications on Semantic Scholar in the Context of Open Access

Abstract: In this article, the Semantic Scholar service will be examined, and the question will be clarified to what extent information from the service can be reused for OA publication monitoring at libraries and universities. Based on a complete survey of journal articles from 2000 to 2020, the OA share within different scientific disciplines will be determined. Furthermore, the quality of OA references in Semantic Scholar will be evaluated in this work.

Keywords: Open access; Semantic Scholar; open access monitoring; data analysis; big scholarly data

1 Einleitung

In den letzten Jahren wurde das Angebot nicht proprietärer Datenbanken im Bereich der Bibliometrie erheblich ausgebaut. Neben den etablierten kommerziellen und zugangsbeschränkten bibliometrischen Datenbanken Scopus und Web of Science (WoS) haben sich diesbezüglich neue Akteure und Dienste auf dem Markt verbreitet, die im bibliothekarischen Diskurs zunehmend an Akzeptanz und Popularität gewinnen. Hierzu zählen OpenAlex (vormals Microsoft Academic Graph (MAG)), Crossref, Unpaywall und Dimensions. Viele Bibliotheken, Hochschulen und Forschungseinrichtungen nutzen diese Services, um Aussagen über das eigene Publikationsverhalten zu treffen. Dieser Umstand ist oftmals mit institutionellen und nationalen Richtlinien verbunden, die eine systematische Erfassung und Auswertung des Publikationsaufkommens fordern. In diesem Zusammenhang stehen die Institutionen vor verschiedenen Herausforderungen. Ein Schwerpunkt hierbei liegt im verantwortungsvollen Umgang mit bibliometrischen Daten und Verfahren.[1]

Parallel zu dieser Entwicklung hat die Transformation des wissenschaftlichen Publikationswesens in den Open Access (OA) zahlreiche Veränderungen in der akademischen Landschaft angestoßen. Neben einem anhaltenden Anstieg von OA-Veröffentlichungen in der wissenschaftlichen Literatur[2] sorgt die zunehmende Komplexität der strukturellen Förderung und Auswertung von OA-Publikationen für Beratungsbedarf an wissenschaftlichen Einrichtungen.[3] Insbesondere das OA-Monitoring, welches im Zuge von Förderprogrammen wie „Open Access Publikationskosten“ der Deutschen Forschungsgemeinschaft (DFG) verpflichtend praktiziert wird, verlangt von teilnehmenden Einrichtungen eine kontinuierliche Beobachtung und Erhebung der eigenen Publikationsaktivität.[4] Damit einhergehend ist die Eingliederung und ständige Evaluierung von bibliografischen Datenquellen, um den Standards eines OA-Berichtswesens nachzukommen.

Bibliotheken und Hochschulen können bei der Einbindung von bibliografischen Datenquellen für das Publikationsmonitoring auf ein breites Portfolio schauen. Als Goldstandard werden immer noch die kommerziellen Anbieter Elsevier (Scopus) und Clarivate Analytics (WoS) angesehen.[5]

***Kontaktperson: Nick Haupka**, nick.haupka@sub.uni-goettingen.de.
https://orcid.org/0009-0002-6478-6789

1 Hicks et al. (2015).
2 Archambault et al. (2014), Piwowar et al. (2018), Robinson-Garcia et al. (2020).
3 Deutsche Forschungsgemeinschaft (2020) 9.
4 Mittermaier (2021).
5 Pranckutė (2021).

Mit der Einführung frei zugänglicher Datenquellen wie dem MAG, Crossref und Unpaywall schwindet allerdings der Einfluss kommerzieller Angebote. So werden dem MAG eine außerordentliche Abdeckung von Publikationen und Zitationen bescheinigt, die in einigen Bereichen Scopus und WoS übertreffen.[6]

Die seit 2015 existierende Suchmaschine Semantic Scholar bildet in diesem Kontext eine weitere kostenlos zugängliche Datenquelle für bibliometrische Analysen. Der Service wird von der Non-Profit-Organisation Allen Institute for Artificial Intelligence entwickelt und dient der KI-gestützten Recherche von wissenschaftlichen Veröffentlichungen. Die Herangehensweise von Semantic Scholar gliedert sich dabei in die aktuelle Entwicklung rund um graphengestützte Repräsentationen von wissenschaftlicher Literatur ein.[7] Der Schwerpunkt von Semantic Scholar liegt auf Publikationen im Bereich der Biologie, Medizin und Informatik. Während jedoch zu anderen bibliografischen Services bereits eine umfassende Literatur vorliegt, hat Semantic Scholar bislang kaum Beachtung in der bibliometrischen Forschung erfahren. Eine Analyse über die Abdeckung des Datendienstes könnte Bibliotheken und Hochschulen bei der Entscheidung unterstützen, Semantic Scholar als kostenfreie Datenquelle in ihr Publikationsmonitoring aufzunehmen.

In diesem Beitrag soll die wissenschaftliche Suchmaschine Semantic Scholar untersucht werden. Zu diesem Zweck wird ein Datenbanksnapshot des Dienstes analysiert, welcher den kompletten Umfang an wissenschaftlichen Publikationen in Semantic Scholar enthält. Dabei sollen Aussagen zum Datenbestand, zum OA-Stand sowie zur Nachnutzung von OA-Nachweisen in Semantic Scholar getroffen werden.

2 Hintergrund

2.1 Publikationsmonitoring an Bibliotheken und Hochschulen

Im Zuge der OA-Transformation sind immer mehr Bibliotheken am Veröffentlichungsprozess wissenschaftlicher Publikationen beteiligt. Diese Entwicklung ist mit öffentlich subventionierten Förderprogrammen verbunden, die das OA-Publizieren sowie dessen substanzielle Strukturen an beteiligten Einrichtungen fördern. Seit 2011 werden in Deutschland durch das Förderprogramm „Open Access Publizieren“ sowie durch dessen Nachfolgeprogramm „Open-Access-Publikationskosten“ hochschulweite Publikationsfonds aufgebaut. Vorausgegangen sind Entwicklungen in den Vereinigten Staaten, wo bereits 1998 erste Maßnahmen zur Einrichtung von Publikationsfonds getroffen wurden. Mit der „Compact for Open-Access Publishing Equity“ sollte ein neues Finanzierungsmodell etabliert werden, welches auf Publikationsgebühren beruht und damit hohe Subskriptionsgebühren ausgleichen und ersetzen sollte.[8] Ein (OA-) Publikationsfond wird nach Pampel und Tullney als ein Finanzierungs- und Steuerungsinstrument wissenschaftlicher Einrichtungen zur Übernahme von (OA-)Publikationsgebühren definiert.[9] Ziel dieser Publikationsfonds ist es, öffentlich zugängliche wissenschaftliche Publikationen zu fördern.[10] Mit der Implementierung von Publikationsfonds an wissenschaftlichen Institutionen ist auch eine umfassende Beschäftigung mit der internen Publikationsaktivität verbunden.[11] Unter anderem soll durch ein sogenanntes Publikationsmonitoring, definiert als die systematische Erfassung und Dokumentation des Publikationsoutputs einer Einrichtung,[12] der künftige Bedarf an OA-Publikationsgebühren ermittelt werden.[13] Das Publikationsmonitoring zielt zudem auch auf eine strategische Komponente ab, nämlich die Rechenschaftspflicht gegenüber investierten Mitteln, anhand derer der Anteil von OA-Publikationen gesteigert werden soll.

Eine Evaluierung des Förderprogramms „Open Access Publizieren“ hat gezeigt, dass viele Institutionen Schwierigkeiten damit haben, fundierte Zahlen zur eigenen Publikationsaktivität zu erheben.[14] So konstatiert Pampel, dass die Strategien zur Finanzierung von OA-Publikationskosten an wissenschaftlichen Einrichtungen divers und ein Monitoring der entstehenden Kosten wenig professionalisiert ist.[15] Wissenschaftliche Einrichtungen binden im Rahmen des Publikationsmonitorings verschiedene Datenquellen ein, darunter (halb-)automatisierte Hochschulbibliografien, Publikationsdatenbanken und Forschungsinformationssysteme sowie tabellarische Listen, um den Anforderungen eines Berichtswesens im OA-Publizieren nachzukommen.[16] Dieses Berichtswesen, welches auf der Aggregation verschiedener offener und proprietärer Datenquellen aufbaut, zeichnet sich durch eine hohe Heterogenität aus, wodurch eine automatisierte Dokumentation der Publikationsaus-

6 Visser et al. (2020), Martín-Martín et al. (2021).
7 Ammar et al. (2018).
8 Pampel und Tullney (2017) 162.
9 Ebd.
10 Scholarly Publishing and Academic Resources Coalition (2023).
11 Fournier und Weihberg (2013) 239.
12 Schmeja und Tullney (2020) 203.
13 Fournier und Weihberg (2013) 239.
14 Ebd.
15 Pampel (2019) 58.
16 Barbers et al. (2020) 20.

gaben nicht oder kaum gegeben ist.[17] Insbesondere die Auswertung von Publikationsdaten aus externen Literatur- und Zitationsdatenbanken stellt wissenschaftliche Einrichtungen vor diverse Herausforderungen, was nicht zuletzt mit Verzerrungen und Fehlanreizen aus bibliometrischen Datenbanken zu begründen ist. Zum Beispiel erfassen kommerzielle Systeme wie WoS und Scopus einen unterproportionalen Anteil nicht-englischsprachiger Literatur sowie einen geringen Anteil bestimmter Publikationstypen wie etwa Konferenzveröffentlichungen.[18] Ferner werden in diesen Datenbanken selektiv Zeitschriften aus etablierten Verlagen sowie spezifischen Disziplinen indexiert.[19] Die Erhebung von Daten aus kommerziellen Systemen sowie aus nicht öffentlich einsehbaren internen Datenquellen wirft darüber hinaus die Frage nach der Transparenz des Publikationsmonitorings an wissenschaftlichen Einrichtungen auf. Zum Beispiel ist es problematisch, wenn institutionsübergreifend Daten gesammelt und ausgewertet werden sollen, etwa für eine Gruppe von Institutionen, ein Bundesland oder ein Land.[20] Weiterhin ergeben sich aus fehlenden oder inkorrekten Metadaten, zum Beispiel durch eine mangelhafte oder geringfügige Bereitstellung von Publikationsdaten seitens der Verlage in Publikations- und Zitationsdatenbanken, Hürden in der anschließenden Auswertung des eigenen Publikationsaufkommens.[21] Eine datenbankübergreifende Problematik bildet die Disambiguierung von Autoren- und Institutionsnamen. Nicht zuletzt trägt eine heterogene Publikationskultur dazu bei, dass Herausforderungen innerhalb des Publikationsmonitorings entstehen.

Es existieren bereits einige Studien im Bereich der Evaluation bibliografischer Datenbanken. Insbesondere sollten hier die Studien von Piwowar et al., Akbaritabar und Stahlschmidt sowie Jahn et al. hervorgehoben werden, die sich mit dem OA-Discovery-Service Unpaywall auseinandersetzen.[22] Studien zu OpenAlex wurden unter anderem von Priem et al. sowie Scheidsteger und Haunschild durchgeführt.[23] Harzing sowie Borrego et al. untersuchten die Services Dimensions und Crossref.[24] Analysen zu MAG, Scopus und WoS wurden unter anderem von Martín-Martín et al., Visser et al. sowie Harzing und Alakangas durchgeführt.[25] Zu Semantic Scholar existiert eine Studie von Kinney et al.[26] Diese befasst sich aber ausschließlich mit dem Datenbestand und nicht mit OA.

3 Methode

Daten aus Semantic Scholar wurden über die vom Service bereitgestellte Datasets API heruntergeladen.[27] Über die Datasets API werden regelmäßig Datenbanksnapshots angeboten. Die Snapshots umfassen dabei sämtliche Informationen aus Semantic Scholar zu einem fixen Zeitpunkt. Snapshots werden alle zwei Wochen veröffentlicht. Ein Snapshot enthält verschiedene Datensätze. Eine Auflistung der Datensätze sowie deren Beschreibung findet sich auf der Webseite des Services.[28] Für diese Arbeit wurde der *papers*-Datensatz sowie der *publication-venues*-Datensatz verwendet. Die genutzten Datensätze tragen den Release 2023-05-23.

Für das Herunterladen des Datensatzes wird ein API-Schlüssel benötigt. Dieser kann entgeltfrei beantragt werden.[29] Der Datensatz enthält mehrere Links zu Dateien, die mit einem Python-Skript ausgelesen und heruntergeladen werden können. Insgesamt umfasst der Datensatz 43 Gigabyte. Der Datensatz wurde nach dem Herunterladen in eine BigQuery-Instanz der Staats- und Universitätsbibliothek (SUB) Göttingen überführt. BigQuery ist ein Data Warehouse und eine Cloud-Plattform, mit deren Hilfe sich große Datensätze analysieren lassen. Im Kontext von Big Scholarly Data, definiert als große und schnell anwachsende wissenschaftliche Datensammlungen,[30] bildet BigQuery ein Standardinstrument bei der Analyse und Auswertung von Publikationsdaten an der SUB Göttingen,[31] weshalb es, aufgrund eines bestehenden Zugangs, für diese Arbeit verwendet wurde. Eine Verarbeitung der Daten war im Vorfeld nicht erforderlich. Der Datensatz ist im Newline Delimited JSON-Format, welches ohne großen Aufwand in BigQuery überführt und geladen werden kann. In BigQuery wurden zwei Tabellen erstellt (*papers* und *venues*), die mittels SQL abgefragt werden können. Die Tabellen enthalten sämtliche Spalten des heruntergeladenen Snapshots.

Für die Datenanalyse wurde Python verwendet. Der Code ist in einem Jupyter-Notebook hinterlegt, das auf

17 Vierkant et al. (2017) 158, Barbers et al. (2020) 20.
18 Schmeja und Tullney (2020) 209.
19 Weingart (2005) 123.
20 Schmeja und Tullney (2020) 209.
21 Akbaritabar und Stahlschmidt (2019), Borrego (2023).
22 Piwowar et al. (2018), Akbaritabar und Stahlschmidt (2019), Jahn et al. (2021).
23 Priem et al. (2022), Scheidsteger und Haunschild (2022).
24 Harzing (2019), Hendricks et al. (2020), Borrego (2023).
25 Harzing und Alakangas (2016), Martín-Martín et al. (2021), Visser et al. (2021).
26 Kinney et al. (2023).
27 https://api.semanticscholar.org/api-docs/datasets.
28 https://api.semanticscholar.org/api-docs/datasets#tag/Release-Data/operation/get_release.
29 https://www.semanticscholar.org/product/api.
30 Xia et al. (2017) 18.
31 Haupka et al. (2022).

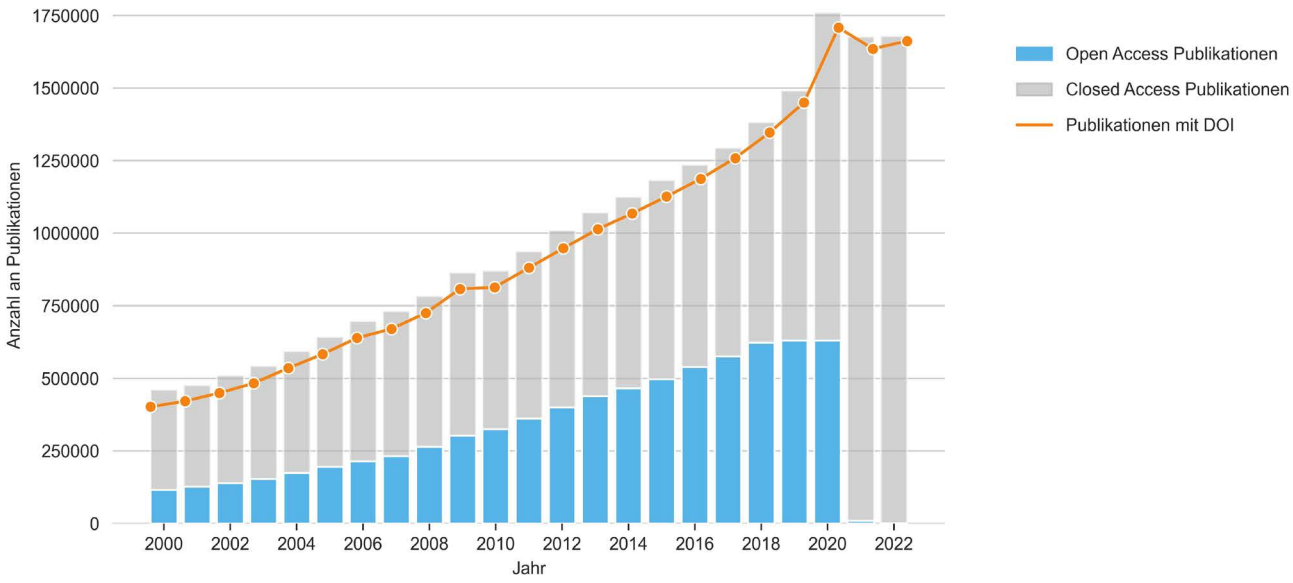

Abb. 1: Verbreitung von Open Access im Publikationsaufkommen[37]

GitHub zugänglich ist.[32] Bei der folgenden Analyse werden Zeitschriftenartikel der Publikationsjahre 2000 bis 2020 ausgewertet. Für das Publikationsjahr wurde das Feld *publicationDate* verwendet. Im späteren Verlauf der Arbeit werden Ergebnisse aus Semantic Scholar mit anderen bibliografischen Datenbanken verglichen. Hierbei handelt es sich um die Services Unpaywall, Crossref, OpenAlex und WoS. Es wurden jeweils komplette Dumps verwendet. Daten aus Unpaywall stammen von März 2022. Daten aus Crossref und OpenAlex sind von April 2023. Der Abgleich mit WoS-Daten ist innerhalb der KB-Infrastruktur (Kompetenznetzwerk Bibliometrie) erfolgt.[33] Diese stammen von Februar 2023. Zusätzlich wurde die Semantic Scholar API benutzt. Dabei wurde auf einen inoffiziellen Python-Client zurückgegriffen.[34] API-Anfragen wurden im Mai 2023 durchgeführt. OA-Quoten wurden, wenn nicht anders angegeben, über Informationen aus Semantic Scholar berechnet.[35] In dieser Arbeit wurde ein listenbasiertes Verfahren genutzt, um Paratexte auszuschließen. Dieses Verfahren ist dabei angelehnt an die Vorgehensweise des Services Unpaywall.[36] Paratexte sind Texte, die einen wissenschaftlichen Artikel begleiten, beispielsweise ein Inhaltsverzeichnis oder ein Titelcover einer Zeitschrift und oft dennoch als Dokumententyp Artikel vorliegen.

4 Ergebnisse

4.1 Publikationsaufkommen, Publikationstypen und Fachgebiete

Zum Erhebungszeitpunkt im Mai 2023 enthält Semantic Scholar 212 251 613 Publikationen (alle Publikationsjahre). Davon liegt für 116 545 907 Publikationen (55 %) ein DOI vor. Publikationen werden in zwölf Publikationstypen unterschieden, nämlich Zeitschriftenartikel, Reviews, Konferenzberichte, Studien, Fallberichte, Kommentare, Editorials, klinische Studien, Bücher, News, Metaanalysen und Datensets. In Semantic Scholar werden Publikationen mehreren Publikationstypen zugeordnet. Das bedeutet, dass eine Veröffentlichung beispielsweise sowohl als Zeitschriftenartikel als auch als Review klassifiziert werden kann. Über 157 Mio. Publikationen haben keine Zuordnung. Von den Publikationen mit zugeordnetem Typ machen Zeitschriftenartikel den größten Anteil in Semantic Scholar aus (60 %). Danach folgen Reviews (22 %) und Konferenzberichte (6 %). Bücher (0,4 %) und Datensets (0,001 %) haben eine relativ geringe Anzahl. Semantic Scholar deckt Publikationen von 1708 bis 2023 ab. Publikationen werden bis zu drei Fachdisziplinen zugeordnet. Von diesen existieren 23 in Semantic Scholar (s. Tab. 1). Für 2000 bis 2022 werden 23 026 821 Publikationen als Zeitschriftenartikel identifiziert. Davon haben 21 815 701 einen DOI (95 %).

32 https://github.com/naustica/MA.
33 https://bibliometrie.info.
34 https://pypi.org/project/semanticscholar/.
35 Hierfür wurde das Feld isopenaccess verwendet.
36 https://support.unpaywall.org/support/solutions/articles/44001894783-what-does-is-paratext-mean-in-the-api-.
37 Betrachtet werden Zeitschriftenartikel der Publikationsjahre 2000 bis 2022 in Semantic Scholar. Der OA-Anteil von Zeitschriftenartikeln wird in Blau dargestellt. Die orange Linie zeigt den Anteil aller Zeitschriftenartikel mit einem DOI.

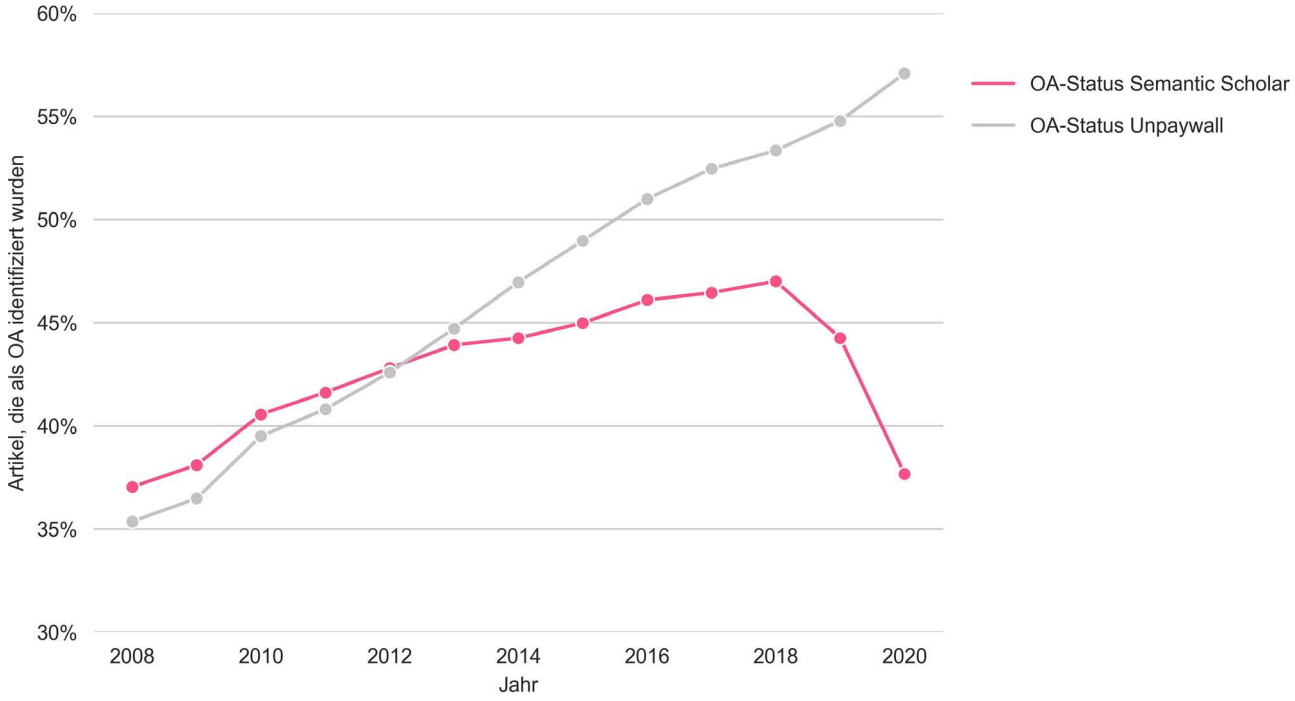

Abb. 2: Vergleich von OA in Semantic Scholar und Unpaywall[39]

4.2 Open-Access-Anteil im Publikationsaufkommen

Eine Publikation wird in Semantic Scholar als OA klassifiziert, wenn ein frei zugänglicher Volltext auf arXiv, ACL Anthology oder PubMed Central gefunden wird.[38] Zudem wird eine Publikation als OA klassifiziert, wenn die entsprechende Publikation in Unpaywall als OA gekennzeichnet wurde. Der OA-Status einer Publikation kann über das Feld *isopenaccess* ausgelesen werden. Im folgenden Teil wird der OA-Anteil von Zeitschriftenartikeln im Publikationsaufkommen von Semantic Scholar analysiert. Abb. 1 bildet das Wachstum der Anzahl von Zeitschriftenartikeln zwischen 2000 und 2022 ab. Die Anzahl der Veröffentlichungen von wissenschaftlichen Artikeln steigt hierbei kontinuierlich. Mit 2020 erreicht die Zahl der Publikationen ihren Kulminationspunkt und sinkt daraufhin leicht. Analog zum Publikationswachstum steigt die Anzahl und der Anteil von OA-Artikeln unter Zeitschriftenartikel bis 2018 stetig. Während 2000 etwa 25 % der Zeitschriftenartikel OA sind, sind es 2018 45 %. Allerdings stagniert ab 2019 die Anzahl von OA-Artikeln und folglich nimmt das Verhältnis von OA-Publikationen im Vergleich zu CA-Publikationen rapide ab. 2019 sind es 42 % und 2020 36 %. Zwischen 2021 und 2022 werden kaum OA-Publikationen klassifiziert. Der Anteil reduziert sich auf 0,5 und schließlich auf 0,05 %. Fast alle Zeitschriftenartikel von 2000 bis 2022 besitzen einen DOI. 2000 haben 87 % der Artikel einen DOI. 2022 sind es 99 %.

Aufgrund der signifikanten Abnahme von OA-Publikation zwischen 2019 und 2022 in Semantic Scholar werden im weiteren Fortgang Zeitschriftenartikel mit OA-Informationen des OA-Discovery-Service Unpaywall verknüpft, womit eine bessere Vergleichbarkeit und Abdeckung hergestellt werden soll. Bei der Kollation zwischen Semantic Scholar und Unpaywall werden die Publikationsjahre 2008 bis 2020 betrachtet, da der in dieser Arbeit verwendete Unpaywall-Snapshot nur Publikationen ab 2008 beinhaltet.[40] Im Gegensatz zum OA-Status in Semantic Scholar, dessen Ausprägung teilweise auf Daten aus Unpaywall beruht, wird der OA-Status von Publikationen in Unpaywall über frei erhältliche Volltexte von Verlagsseiten und Repositorien bestimmt.[41][42] Von 15 013 553 Zeitschriftenartikeln konnten 14 033 976 Artikel mit Publikationen aus Unpaywall über ein DOI-Matching zusammengeführt werden. Das entspricht 93,48 %. Da Unpaywall nur Publikationen mit einem DOI indexiert, kommt es zwangs-

38 Lo et al. (2020) 4972.

39 Dargestellt werden Zeitschriftenartikel der Publikationsjahre 2008 bis 2020, die mit einem DOI in Unpaywall gefunden wurden.

40 S. auch Haupka et al. (2022).

41 https://support.unpaywall.org/support/solutions/articles/44002140100-what-counts-as-an-open-access-location-.

42 https://support.unpaywall.org/support/solutions/articles/44001777288-what-do-the-types-of-oa-status-green-gold-hybrid-and-bronze-mean-.

Tab. 1: OA-Anteile von Zeitschriftenartikeln der Publikationsjahre 2008 bis 2020 in verschiedenen Disziplinen in Semantic Scholar[44]

Fachdisziplin	Anzahl der Artikel	OA-Anteil	Publikationsanteil
Medicine	10 960 460	47,07 %	39,92 %
Biology	4 732 805	53,49 %	17,24 %
Computer Science	3 117 030	32,63 %	11,35 %
Chemistry	1 939 725	42,40 %	7,06 %
Psychology	1 480 217	46,34 %	5,39 %
Materials Science	794 862	40,45 %	2,89 %
Engineering	777 208	29,96 %	2,83 %
Mathematics	657 956	40,73 %	2,40 %
Physics	648 514	47,75 %	2,36 %
Environmental Science	476 254	46,62 %	1,73 %
Agricultural And Food Sciences	392 199	42,09 %	1,43 %
Business	391 901	29,44 %	1,43 %
Political Science	265 460	45,89 %	0,97 %
Education	205 397	34,18 %	0,75 %
Sociology	156 207	34,04 %	0,57 %
Economics	150 709	40,35 %	0,55 %
Geography	85 381	56,10 %	0,31 %
Geology	56 962	49,09 %	0,21 %
Art	55 724	30,00 %	0,20 %
History	34 386	33,24 %	0,13 %
Philosophy	34 042	34,82 %	0,12 %
Law	22 893	31,81 %	0,08 %
Linguistics	21 170	47,67 %	0,08 %

läufig zu einem Ausschluss von Zeitschriftenartikeln ohne DOI. Diese machen 4,6 % der Zeitschriftenartikel zwischen 2008 und 2020 in Semantic Scholar aus. Ein Vergleich des OA-Status aus Semantic Scholar und Unpaywall wird in Abb. 2 dargestellt. Zwischen 2008 und 2013 ist der Anteil der OA-Veröffentlichungen in beiden Datenbanken annähernd gleich. Ab 2014 nimmt die Differenz in der OA-Kategorisierung beider Datenbanken zu. Während in Semantic Scholar nur ein leichter Anstieg des OA-Anteils gemessen werden kann, welcher ab 2019 erheblich abfällt, steigt der Anteil von OA-Publikationen laut Unpaywall beständig. Der Pearson-Korrelationskoeffizient des OA-Status beider Systeme beträgt 0,62. Semantic Scholar identifiziert zwischen 2008 und 2020 43 % der Artikel als OA (Unpaywall: 47 %).

4.3 Open-Access-Status nach Fachdisziplinen

Für die in Semantic Scholar verwendeten Fachdisziplinen wird im folgenden Teil der OA-Anteil bestimmt. Disziplinen werden zunächst einzeln betrachtet (s. Tab. 1). Tab. 2 gruppiert die verschiedenen Disziplinen in die Kategorien Lebenswissenschaften & Biomedizin, Naturwissenschaften, Sozialwissenschaften und Kunst- & Geisteswissenschaften. Die Einteilung erfolgt nach Bosman und Kramer.[43] OA-Anteile werden für die entsprechenden Tabellen für Zeitschriftenartikel der Publikationsjahre 2008 bis 2020 berechnet. Die jeweiligen OA-Anteile werden ausschließlich über Semantic Scholar bestimmt.

Es lässt sich erkennen, dass Disziplinen mit einem relativ hohen Publikationsoutput auch hohe OA-Publikationsraten haben. Die meisten Publikationen in Semantic Scholar sind Lebenswissenschaften und Biomedizin zuzuordnen (s. Tab. 2). In diesem Bereich werden zudem die meisten OA-Publikationen identifiziert (49 %). Die Naturwissenschaften, darunter Chemie, Mathematik, Physik und Informatik, machen den zweitgrößten Anteil am Publikationsaufkommen in Semantic Scholar aus. Der OA-Anteil der Naturwissenschaften misst 38 %. Der kleinste OA-Anteil in den Naturwissenschaften ist in den Ingenieurwissenschaften (30 %) und der Informatik (33 %) zu messen. In der Physik (48 %), Geologie (49 %) und in den Umweltwissenschaften (47 %) wird der größte OA-Output innerhalb der Naturwissenschaften ermittelt. In den Kunst- und Geisteswissenschaften ist der OA-Output am geringsten (32 %). Das liegt vor allem an der niedrigen OA-Identifikationsrate im

43 Bosman und Kramer (2018).

44 Der OA-Status wird aus Semantic Scholar übernommen. Artikel können in bis zu drei Disziplinen auftreten. Berechnet wurde der Publikationsanteil mittels der Summe aller Publikationen (Publikationen werden mehrmals gezählt, wenn diese in mehreren Disziplinen klassifiziert werden).

Tab. 2: Einteilung von Fachdisziplinen in Kategorien[45]

Kategorie	Fachdisziplinen	Anzahl der Artikel	OA-Anteil	Publikationsanteil
Lebenswissenschaften und Biomedizin	Biology, Agricultural and Food Sciences, Medicine	16 085 464	48,84 %	58,58 %
Naturwissenschaften	Chemistry, Computer Science, Engineering, Physics, Materials Science, Environmental Science, Mathematics, Geology	8 468 511	38,04 %	30,84 %
Sozialwissenschaften	Psychology, Business, Education, Economics, Political Science, Sociology, Law, Linguistics, Geography	2 779 335	42,19 %	10,12 %
Kunst- und Geisteswissenschaften	History, Art, Philosophy	124 152	32,22 %	0,45 %

Bereich Kunst (30 %), welcher mit den Rechtswissenschaften (32 %) und Betriebswirtschaft (29 %) die kleinsten OA-Anteile im Datensatz aufweist. Die Kunst- und Geisteswissenschaften machen andererseits nur 0,45 % der Publikationen unter den Zeitschriftenartikeln zwischen 2008 und 2020 in Semantic Scholar aus. In den Sozialwissenschaften ist der OA-Anteil in den Disziplinen Erziehungswissenschaften (34 %), Soziologie (34 %), Rechtswissenschaften (32 %) und Betriebswirtschaft (29 %) am geringsten. Am höchsten ist der OA-Anteil in den Sozialwissenschaften in den Fächern Politikwissenschaften (46 %), Psychologie (46 %) und Geografie (56 %). Der OA-Anteil in den Sozialwissenschaften liegt laut Semantic Scholar bei 42 %.

4.4 Open Access nach Verlagen

Das untersuchte Datenset enthält keine expliziten Verlagsinformationen. Entsprechende Informationen können nur mithilfe weiterer externer Datenquellen angereichert werden. Eine gängige Methode hierbei ist es, Informationen mittels der International Standard Serial Number (ISSN) aus dem ISSN International Portal zu extrahieren.[46] Informationen aus dem ISSN International Portal sind allerdings kostenpflichtig, weshalb Informationen aus Crossref nachgenutzt worden sind. Für die in Semantic Scholar enthaltenen Zeitschriftenartikel der Publikationsjahre 2008 bis 2020 wurden Verlagsnamen aus Crossref mit einem ISSN-Matching zusammengeführt. Bei der Zuweisung von Artikeln zu Verlagen wurden mögliche Imprint-Wechsel vernachlässigt, die durch die zeitliche Distanz zwischen den gewählten Snapshots vorkommen könnten. OA-Anteile der jeweiligen Verlage wurden daraufhin mit den vorliegenden Daten aus Semantic Scholar berechnet (s. Tab. 3). Insgesamt konnten Verlagsinformationen zu 13 841 906 Zeitschriftenartikeln gefunden werden (92 %). Artikel, die keiner ISSN zugeordnet waren, werden in der Analyse in diesem Abschnitt ausgeschlossen.

Daten aus Tab. 3 werden folgend in Abb. 3 grafisch dargestellt. Der Verlag Elsevier ist im vorliegenden Datenset am häufigsten vertreten. Insgesamt bemisst der Anteil der Zeitschriftenartikel des Verlages Elsevier 34 % am Publikationsaufkommen der Jahre 2008 bis 2020 in Semantic Scholar. Darauf folgt der Verlag Wiley mit knapp 14 %. Der Anteil von Zeitschriftenartikeln des Verlages Springer liegt bei 8 %. Der Verlag SAGE sowie der Verlag Informa UK Limited machen kumulativ 2 % des Publikationsvolumens im Datenset aus. Damit werden den „Big Five Publisher“[47] insgesamt 58 % der Artikel in Semantic Scholar zugeordnet. Dieses Ergebnis deckt sich mit den Ergebnissen von Stoy et al., die den Anteil dieser Verlage am Publikationsaufkommen auf über 55 % bestimmen (basierend auf Daten aus WoS).[48]

Große OA-Verlage sind mit knapp 10 % im Publikationsaufkommen vertreten. Davon gehen 4 % an den Verlag Multidisciplinary Digital Publishing Institute (MDPI), 3 % an den Verlag Public Library of Science (PLOS) und 2 % an Frontiers Media. Obwohl diese Verlage reine OA-Zeitschriften veröffentlichen, liegt die entsprechende OA-Quote im Datenset bei 93 bis 98 %. Diese Beobachtung kann möglicherweise auf die unzureichende OA-Datenlage in Semantic Scholar zurückgeführt werden. Kleinere Verlage, die unter dem Term Other zusammengefasst werden, machen kumulativ 16 % der Zeitschriftenartikel im untersuchten Datensatz aus. Der OA-Anteil dieser Verlage liegt mit 53 % unter dem Mittel (67 %).

Die OA-Quote ist unter den großen fünf Verlagen unterschiedlich ausgeprägt. Der Verlag Elsevier hat einen OA-Anteil von 22 %. Der OA-Anteil des Verlages Wiley liegt bei 42 %. Springers OA-Anteil bemisst 82 %. Informa UK Limited hat einen OA-Anteil von 74 %. Die OA-Quote des Verlages SAGE liegt bei 54 %. Medknow und Ovid Technologies haben unter den zwanzig am häufigsten vorkommenden Verlagen

45 Betrachtet werden Zeitschriftenartikel der Publikationsjahre 2008 bis 2020 in Semantic Scholar. OA-Anteile wurden mit Informationen aus Semantic Scholar berechnet.

46 https://www.issn.org.

47 Nishikawa-Pacher (2022), Butler et al. (2022).

48 Story et al. (2019) 9.

Tab. 3: Verlage in Semantic Scholar

Verlag	Publikationsanzahl	OA-Anteil	Publikationsanteil
Elsevier BV	2 354 894	21,70 %	34,03 %
Wiley	938 952	42,44 %	13,57 %
Springer Science and Business Media LLC	565 917	82,33 %	8,18 %
Ovid Technologies (Wolters Kluwer Health)	436 826	19,60 %	6,31 %
MDPI AG	264 089	96,99 %	3,82 %
Public Library of Science (PLOS)	232 331	93,25 %	3,36 %
Institute of Electrical and Electronics Engineers (IEEE)	174 227	33,53 %	2,52 %
Frontiers Media SA	158 632	97,40 %	2,30 %
Oxford University Press (OUP)	113 967	53,50 %	1,65 %
Medknow	79 427	5,25 %	1,15 %
BMJ	67 698	77,50 %	0,98 %
Informa UK Limited	64 367	74,37 %	0,93 %
Bentham Science Publishers Ltd	53 427	20,71 %	0,77 %
SAGE Publications	51 857	54,00 %	0,75 %
FapUNIFESP (SciELO)	50 597	68,35 %	0,73 %
Optica Publishing Group	42 667	41,71 %	0,62 %
The Optical Society	42 667	41,71 %	0,62 %
Hindawi Limited	42 410	30,14 %	0,61 %
American Society for Biochemistry & Molecular Biology (ASBMB)	42 228	96,24 %	0,61 %
International Union of Crystallography (IUCr)	39 015	86,49 %	0,56 %
Other	1 104 758	52,57 %	15,96 %

die niedrigsten OA-Quoten. Medknow hat einen OA-Anteil von 5 %. Ovid Technologies hat einen OA-Anteil von 20 %. Insgesamt werden in Semantic Scholar 1793 Verlage gezählt, die zwischen 2008 und 2020 mit mindestens einem Zeitschriftenartikel verknüpft werden.

Kontrastiert man dieses Ergebnis mit der ESAC Market Watch, so fällt auf, dass die relativen OA-Anteile einzelner Verlage in Semantic Scholar höher ausfallen.[49] Möglicherweise ist dies der Tatsache geschuldet, dass eine andere OA-Klassifizierung vorgenommen wurde sowie die Datengrundlage eine andere ist (WoS). Die ESAC Market Watch informiert über Trends im wissenschaftlichen Publikationsmarkt und wird von Bibliotheken und Forschungseinrichtungen für strategische Entscheidungen genutzt.

4.5 Open-Access-Nachweise in Semantic Scholar

Jede OA-Publikation in Semantic Scholar wird mit einem Link zu einem frei zugänglichen Textdokument verknüpft. Zusätzlich wird eine OA-Kategorisierung vorgenommen. Diese Informationen sind im Abstracts-Datensatz von Semantic Scholar enthalten, welcher jedoch nicht in dieser Arbeit berücksichtigt und analysiert wurde. Stattdessen wurden Daten aus der Graph API verwendet.[50] Um die Verteilung von OA-Kategorien sowie OA-Nachweisen in Semantic Scholar zu prüfen, wurde ein Datensample mit 5000 zufällig gewählten Zeitschriftenartikeln der Publikationsjahre 2008 bis 2020 erstellt und mit Informationen aus der Graph API angereichert. Für die Graph-API-Anfrage wurde das Python-Package *semanticscholar* verwendet.[51] Semantic Scholar nutzt die OA-Kategorien Gold OA, Green OA, Bronze OA und Hybrid OA. Diese werden auch in Unpaywall verwendet, weshalb sich ein Vergleich zwischen diesen beiden Anbietern aufdrängt. Abb. 4 zeigt einen Vergleich zwischen Semantic Scholar und Unpaywall hinsichtlich der OA-Kategorisierung. Für die Kategorisierung mit Unpaywall wurden sämtliche Zeitschriftenartikel der Publikationsjahre 2008 bis 2020 aus Semantic Scholar genutzt, die mit einem DOI in Unpaywall aufgefunden worden sind. Der entsprechende OA-Status wurde aus Unpaywall entnommen. Somit werden 5 000 Artikel mit OA-Informationen aus Semantic Scholar mit 14 Mio. Zeitschriftenartikeln mit OA-Informationen aus Unpaywall verglichen.

Die OA-Kategorisierung ist bei beiden Systemen ähnlich. Durchschnittlich werden 16 % der Publikationen in Semantic Scholar als Gold OA identifiziert, während in Unpaywall 19 % der Artikel als Gold OA klassifiziert werden. Der Anteil von Green OA liegt in Semantic Scholar durchschnittlich

49 https://esac-initiative.org/market-watch/ (Stand: Dezember 2023).

50 Dass der Abstract-Datensatz OA-Informationen enthält, wurde erst nach der Einreichung des Artikels entdeckt.

51 https://github.com/danielnsilva/semanticscholar.

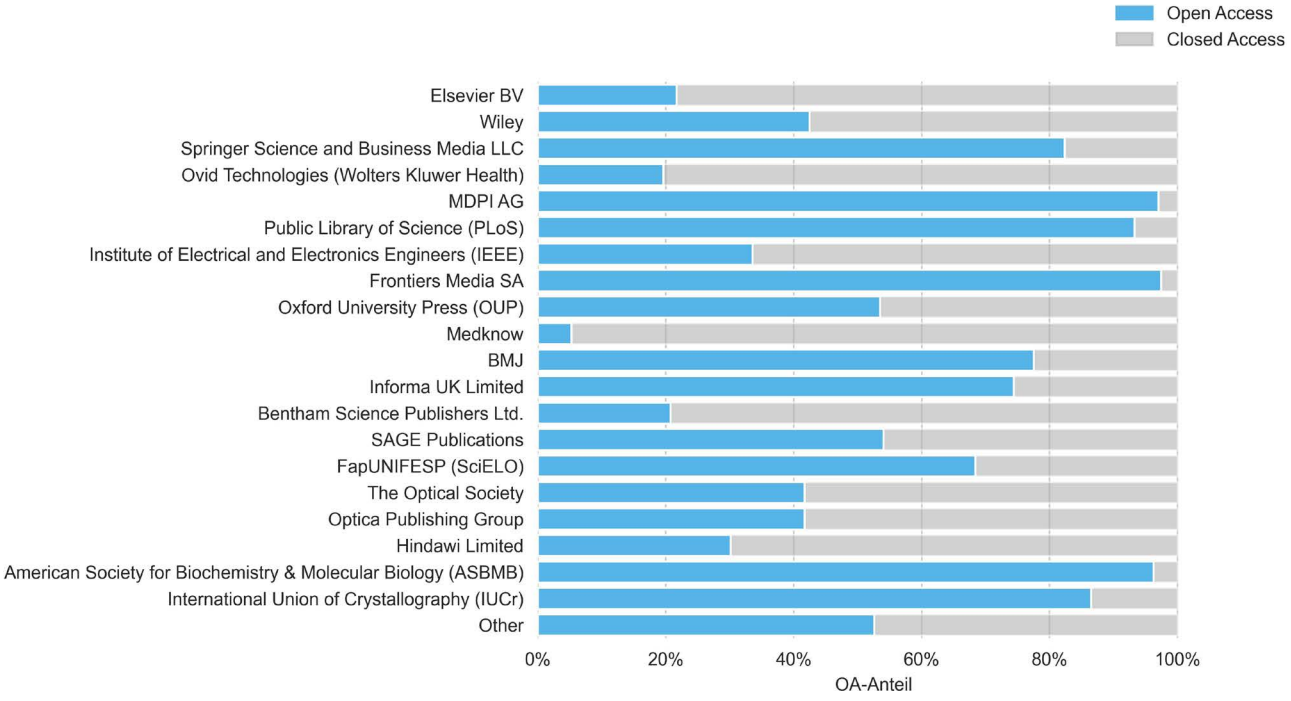

Abb. 3: OA-Anteile der zwanzig größten Verlage in Semantic Scholar

bei 13 %. In Unpaywall werden durchschnittlich ebenso 13 % der Artikel als Green OA klassifiziert. Als Hybrid OA werden in Semantic Scholar im Mittel 4 % gemessen und in Unpaywall 5 %. Für Bronze OA werden in Semantic Scholar im Schnitt 9 % ermittelt und in Unpaywall 11 %. In beiden Datenquellen lassen sich ähnliche Trends identifizieren. So wird etwa ein Anstieg von Gold OA in beiden Datenbanken festgestellt. Dieser geht mit einem kleinem Rückgang von Green OA einher. Bronze OA fällt leicht von 2008 bis 2020. Der Anteil von Hybrid OA steigt dezent. Von den 5 000 Artikeln des Samples sind 2 076 Veröffentlichungen im OA (41,52 %). Dieser Anteil liegt in Unpaywall leicht höher (46,46 %).

4.6 Semantic Scholar im Vergleich mit anderen Datenbanken

Im Folgenden soll der Bestand an Zeitschriftenartikeln mit einem DOI von 2008 bis 2020 in Semantic Scholar mit den Datenbanken Unpaywall, WoS, OpenAlex und Crossref verglichen werden (siehe Abb. 5). Hierzu wurden zunächst alle Zeitschriftenartikel der entsprechenden Jahre in Semantic Scholar sowie in den respektiven Datenbanken gezählt. Aus typologischer Perspektive ist der Vergleich mit Kompromissen behaftet. So kann es vorkommen, dass die verschiedenen Datenbankanbieter unterschiedliche Ansätze bei der Typologisierung von Zeitschriftenartikeln verfolgen. Gezählt werden zunächst Zeitschriftenartikel in Semantic Scholar mit einem DOI. Anschließend fand ein DOI-Matching mit den genannten Datenbanken statt. Daraufhin wurde für die respektiven Datenbanken die Anzahl der Publikationen mit dem Typ Zeitschriftenartikel gezählt.

Die meisten Publikationen aus Semantic Scholar werden in OpenAlex abgebildet. So sind in OpenAlex 99 % der Artikel aus Semantic Scholar enthalten. 59 056 Artikel wurden nicht in OpenAlex gefunden. Eine ähnliche Abdeckung wird in Unpaywall beobachtet. Insgesamt werden 98 % der Artikel aus Semantic Scholar in Unpaywall aufgefunden. 27 195 329 Artikel sind ausschließlich nur in Unpaywall zu finden. In Crossref werden 62 % der Zeitschriftenartikel aus Semantic Scholar aufgefunden. Über 5 Mio. Artikel sind dabei ausschließlich in Semantic Scholar zu finden. Da Crossref als DOI-Registrierungsagentur fungiert, ist davon auszugehen, dass in Semantic Scholar auch Publikationen mit DOIs enthalten sind, die nicht über Crossref vergeben worden sind. In WoS werden 53 % der Zeitschriftenartikel der Publikationsjahre 2008 bis 2020 mit einem DOI aus Semantic Scholar gefunden. Knapp 10 Mio. Zeitschriftenartikel sind nur in WoS enthalten, womit die geringste Abdeckung unter den Anbietern erreicht wird. Im Vergleich mit Unpaywall, Crossref und OpenAlex werden in WoS die wenigsten Artikel mit einem DOI aufgefunden.

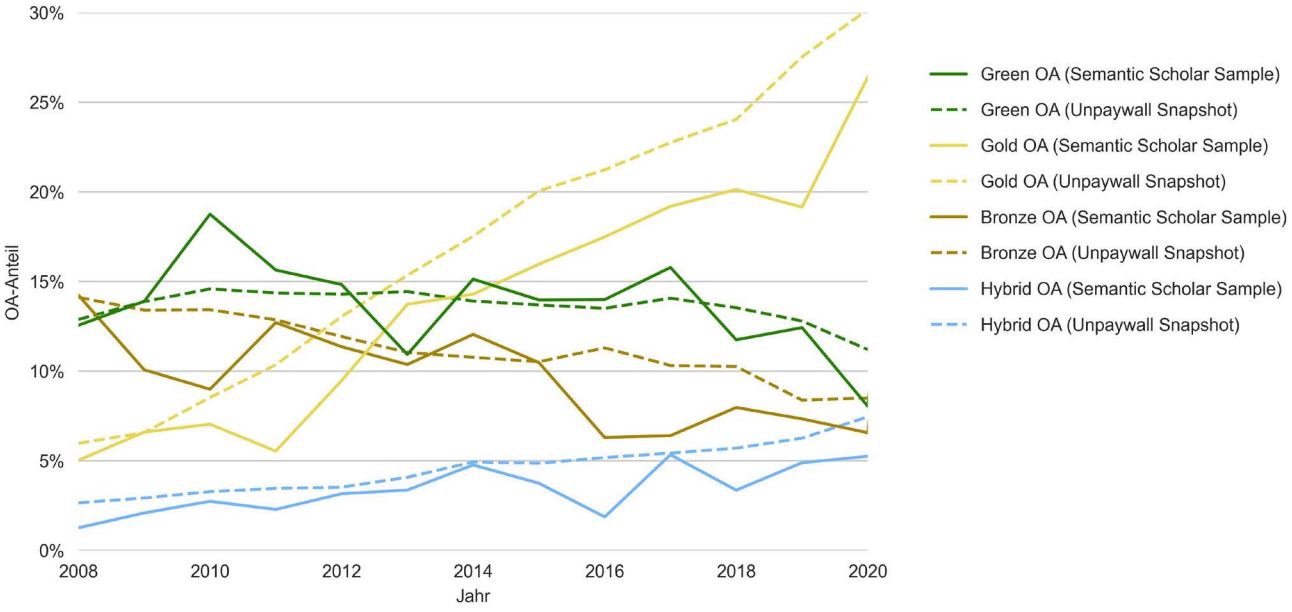

Abb. 4: Verbreitung von OA-Kategorien in Semantic Scholar und Unpaywall[52]

5 Diskussion

Aus der Analyse geht hervor, dass Semantic Scholar sich zum jetzigen Zeitpunkt nur bedingt für das OA-Publikationsmonitoring an Bibliotheken und Hochschulen eignet. Unter anderem konnten für die Publikationsjahre 2021 und 2022 Dateninkonsistenzen festgestellt werden (siehe Abb. 1), die eine adäquate Nachnutzung des OA-Status in Semantic Scholar zum gegenwärtigen Zeitpunkt ausschließen. Möglicherweise liegt eine Verzögerung bei der Indexierung von Publikationen vor (indexing lag). Auch wäre es möglich, dass ein Problem bei der OA-Klassifizierung vorliegt, wie es Stand November 2023 auch in OpenAlex der Fall ist.[53] Die Services OpenAlex und Unpaywall werden beide von der Non-Profit-Organisation OurResearch entwickelt. Allerdings fällt das Klassifizierungsproblem in OpenAlex geringer aus. In der Folge sollten Publikationen aus Semantic Scholar mit OA-Informationen aus anderen Datenquellen angereichert werden. Weiterhin ist unklar, wie Semantic Scholar bei der Klassifikation von OA-Artikeln vorgeht, da eine entsprechende Dokumentation fehlt. Außerdem fehlen ausführliche Informationen zur Klassifizierung von OA-Nachweisen, etwa, ob eine Publikation auf einem Repositorium oder einer Verlagswebseite vorzufinden ist. Das Feld *venue* enthält zwar vage Informationen zum Fundort einer Publikation (zum Beispiel: arXiv, Nature, Biochemistry), allerdings kann aufgrund von Namensambiguitäten sowie fehlender Klassifizierungen nicht mit diesen gearbeitet werden. Hierzu würde es eindeutige Identifier zu diesen Fundorten benötigen, die aktuell nicht in Semantic Scholar enthalten sind. Es sind lediglich ISSNs enthalten, mit denen Artikel zu Zeitschriften zugeordnet werden können. Verlagsinformationen fehlen gänzlich.

Momentan kann davon ausgegangen werden, dass die in Semantic Scholar genutzte OA-Klassifikation von Unpaywall-Informationen abhängig ist. Das geht auch aus einem Artikel von Lo hervor, in dem die Integration von Unpaywall zur Bereitstellung des Semantic Scholar Open Research Corpus (S2ORC) beschrieben wird.[54] Durch die Nachnutzung von Unpaywall seitens Semantic Scholars ist davon auszugehen, dass verlagsbasiertes OA gegenüber repositorienbasiertem OA höher gewichtet wird,[55] sofern keine eigene OA-Klassifizierung seitens Semantic Scholar erfolgt. Semantic Scholar beschreibt in der offiziellen Dokumentation nicht, wie OA-Kategorien definiert werden. Es ist aber davon auszugehen, dass der von Unpaywall genutzte Algorithmus verwendet wird.[56]

Im Publikationsvolumen (Anzahl der indexierten Zeitschriftenartikel) wird Semantic Scholar von anderen Datenbankanbietern übertroffen (siehe Abschnitt 4.6). Jedoch

52 Betrachtet werden Zeitschriftenartikel der Publikationsjahre 2008 bis 2020. Das verwendete Semantic Scholar Sample enthält 5 000 Zeitschriftenartikel.

53 Jahn et al. (2023).

54 Lo (2020).

55 Jahn et al. (2021) 301.

56 https://support.unpaywall.org/support/solutions/articles/44001777288-what-do-the-types-of-oa-status-green-gold-hybrid-and-bronze-mean-.

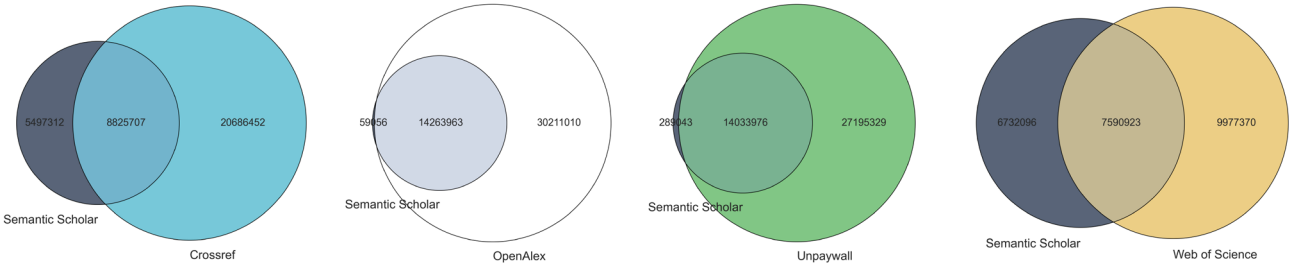

Abb. 5: Überlappung von Zeitschriftenartikeln mit einem DOI aus Semantic Scholar mit Crossref, OpenAlex, Unpaywall und Web of Science

existieren in Semantic Scholar eine Reihe an Publikationen ohne Publikationstyp, weshalb womöglich der Anteil von Zeitschriftenartikeln in Semantic Scholar in dieser Arbeit unterschätzt wurde. Allerdings gehen aus den Ergebnissen dieser Arbeit auch Stärken des Services hervor. Einerseits bietet Semantic Scholar eine Klassifikation von Publikationen nach Fachdisziplinen. Andererseits enthält Semantic Scholar Zitationsdaten, mit denen quantitative Zitationsanalysen durchgeführt werden können. Diese Informationen sind in anderen offenen Datenbanken (Unpaywall, Crossref) nur zum Teil verfügbar. Wie hoch die Zitationsabdeckung in Semantic Scholar ist, müsste allerdings in weiteren Studien untersucht werden. Auch die Qualität der Erschließung der Fachdisziplinen müsste noch mit anderen Datenbanken verglichen werden.

6 Fazit

In dieser Arbeit wurde die wissenschaftliche Suchmaschine Semantic Scholar untersucht. Dem Service wird ein breites Datenfundament konstatiert, das umfangreiche reproduzierbare Analysen im Kontext wissenschaftlicher Informationen erlaubt. Gleichzeitig wurden Probleme in der Ermittlung und Auswertung von OA-Publikationen identifiziert. Unter anderem führt eine fehlende OA-Dokumentation in Semantic Scholar zu Vergleichbarkeitsproblemen, da nicht klar ist, wie Semantic Scholar bei der OA-Klassifikation vorgeht. Weiterhin führen Datenfehler in dem in dieser Arbeit untersuchten Datenbanksnapshot zu ungenauen Aussagen über das wissenschaftliche Publikationsverhalten. Insgesamt wurden über 14 Mio. Zeitschriftenartikel von 2008 bis 2020 analysiert, wobei für die Forschungsfelder Lebenswissenschaften und Biomedizin ein OA-Anteil von 49 %, für die Naturwissenschaften 38 %, für die Sozialwissenschaften 42 % und für die Kunst- und Geisteswissenschaften ein OA-Anteil von 32 % gemessen wurde. Im Mittel wurde ein OA-Anteil von 42 % im Publikationsaufkommen für die Jahre 2008 bis 2020 gemessen. OA-bezogene Ergebnisse wurden in dieser Arbeit mit dem OA-Discovery-Tool Unpaywall verglichen. Dabei wurde herausgefunden, dass wesentliche OA-Informationen aus Unpaywall und Semantic Scholar bis zu einem gewissen Grad deckungsgleich sind. Semantic Scholar beinhaltet im Vergleich zu anderen großen bibliografischen Datenquellen (Unpaywall, WoS, OpenAlex, Crossref) weniger Zeitschriftenartikel. Nichtsdestotrotz könnten Informationen aus Semantic Scholar vor allem im Hinblick auf die Klassifikation von Forschungsfeldern nachgenutzt werden. Derartige Informationen wurden bislang in Studien überwiegend kommerziellen Datenbanksystemen entnommen. Ferner führt eine defizitäre Dokumentation des Services zu einer intransparenten Datenbasis, was die Nachnutzung des Services im Hinblick auf das Publikationsmonitoring an Bibliotheken und Hochschulen erschwert.

Literaturverzeichnis

Akbaritabar, Aliakbar; Stahlschmidt, Stephan (2019): Merits and Limits: Applying open data to monitor open access publications in bibliometric databases. DOI:10.48550/ARXIV.1902.03937.

Ammar, Waleed et al. (2018): Construction of the Literature Graph in Semantic Scholar. DOI:10.48550/arXiv.1805. 02262.

Archambault, Éric; Amyot, Didier; Deschamps, Philippe; Nicol, Aurore et al. (2014): Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels – 1996–2013. Verfügbar unter https://digitalcommons.unl.edu/scholcom/8.

Barbers, Irene; Rosenberger, Sonja; Mittermaier, Bernhard (2020): Auf dem Weg zur Open Access Transformation. In: *Informationspraxis*, 6 (2). DOI:10.11588/IP.2020.2.73240.

Borrego, Ángel; Ardanuy, Jordi; Arguimbau, Llorenç (2023): Crossref as a bibliographic discovery tool in the arts and humanities. In: *Quantitative Science Studies*, 4 (1), 91–104. DOI:10.1162/qss_a_00240.

Bosman, Jeroen; Kramer, Bianca (2018): Open access levels: a quantitative exploration using Web of Science and oaDOI data. Preprint, PeerJ Preprints. DOI:10.7287/peerj.preprints.3520v1.

Butler, Leigh-Ann; Matthias, Lisa; Simard, Marc-André; Mongeon, Phil; Haustein, Stefanie (2022): The oligopoly's shift to open access publishing: How for-profit publishers benefit from gold and hybrid

article processing charges. In: *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*, August. DOI:10.29173/cais1262.

Deutsche Forschungsgemeinschaft (2020): Das DFG-Förderprogramm Open Access Publizieren – Bericht über die Förderung. Verfügbar unter https://www.dfg.de/dfg_profil/zahlen_fakten/evaluation_studien_monitoring/studien/studie_open_access/index.html, zugegriffen am 22.08.2023.

Fournier, Johannes; Weihberg, Roland (2013): Das Förderprogramm „Open Access Publizieren" der Deutschen Forschungsgemeinschaft. Zum Aufbau von Publikationsfonds an wissenschaftlichen Hochschulen in Deutschland. In: *Zeitschrift für Bibliothekswesen und Bibliographie*, 60 (5), 236–43. DOI:10.3196/186429501360528.

Harzing, Anne-Wil (2019): Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? In: *Scientometrics*, 120 (1), 341–49. DOI:10.1007/s11192-019-03114-y.

Harzing, Anne-Wil; Alakangas, Satu (2016): Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. In: *Scientometrics*, 106 (2), 787–804. DOI:10.1007/s11192-015-1798-9.

Haupka, Nick; Jahn, Najko; Hobert, Anne (2022): Praxisbericht Big Scholarly Data an der SUB Göttingen. In: *LIBREAS. Library Ideas*, (41). Verfügbar unter https://libreas.eu/ausgabe41/haupka/, zugegriffen am 22.08.2023.

Hendricks, Ginny; Tkaczyk, Dominika; Lin, Jennifer; Feeney, Patricia (2020): Crossref: The sustainable source of community-owned scholarly metadata. In: *Quantitative Science Studies*, 1 (1), 414–27. DOI:10.1162/qss_a_00022.

Hicks, Diana; Wouters, Paul; Waltman, Ludo; Rijcke, Sarah de; Rafols, Ismael (2015): Bibliometrics: The Leiden Manifesto for research metrics. In: *Nature*, 520 (7548), 429–31. DOI:10.1038/520429a.

Jahn, Najko; Haupka, Nick; Hobert, Anne (2023): Analysing and reclassifying open access information in OpenAlex. Verfügbar unter https://subugoe.github.io/scholcomm_analytics/posts/oalex_oa_status/, zugegriffen am 22.12.2023.

Jahn, Najko; Hobert, Anne; Haupka, Nick (2021): Entwicklung und Typologie des Datendiensts Unpaywall. In: *BIBLIOTHEK – Forschung und Praxis*, 45 (2), 293–303. DOI:10.1515/bfp-2020-0115.

Kinney, Rodney; Anastasiades, Chloe; Authur, Russell; Beltagy, Iz; Bragg, Jonathan; Buraczynski, Alexandra; Cachola, Isabel et al. (2023): The Semantic Scholar Open Data Platform. DOI:10.48550/ARXIV.2301.10140.

Lo, Kyle; Wang, Lucy Lu; Neumann, Mark; Kinney, Rodney; Weld, Daniel (2020): S2ORC: The Semantic Scholar Open Research Corpus. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4969–83. Online: Association for Computational Linguistics. DOI:10.18653/v1/2020.acl-main.447.

Martín-Martín, Alberto; Thelwall, Mike; Orduna-Malea, Enrique; Delgado López-Cózar, Emilio (2021): Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. In: *Scientometrics*, 126 (1), 871–906. DOI:10.1007/s11192-020-03690-4.

Mittermaier, Bernhard (2021): Die Rolle des Open Access Monitor Deutschland bei der Antragstellung im DFG-Förderprogramm „Open-Access-Publikationskosten". In: *o-bib. Das offene Bibliotheksjournal*, 8 (4), 1–14. DOI:10.5282/o-bib/5731.

Nishikawa-Pacher, Andreas (2022): Who are the 100 largest scientific publishers by journal count? A webscraping approach. In: *Journal of Documentation*, 78 (7), 450–63. DOI:10.1108/JD-04-2022-0083.

Pampel, Heinz (2019): Open Access an wissenschaftlichen Einrichtungen in Deutschland. Ergebnisse einer Erhebung im Jahr 2018. Bericht. Helmholtz Open Science Koordinationsbüro. DOI:10.2312/OS.HELMHOLTZ.005.

Pampel, Heinz; Tullney, Marco (2017): 3b. Open-Access-Publikationsfonds. In: *Praxishandbuch Open Access*, hg. von Konstanze Söllner und Bernhard Mittermaier, 162–72. De Gruyter Saur. DOI:10.1515/9783110494068-019.

Piwowar, Heather; Priem, Jason; Larivière, Vincent; Alperin, Juan Pablo et al. (2018): The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. In: *PeerJ*, 6 (Februar), e4375. DOI:10.7717/peerj.4375.

Pranckutė, Raminta (2021): Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. DOI:10.3390/publications9010012.

Priem, Jason; Piwowar, Heather; Orr, Richard (2022): OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv. Verfügbar unter http://arxiv.org/abs/2205.01833, zugegriffen am 30.08.2023.

Robinson-Garcia, Nicolas; Costas, Rodrigo; Leeuwen, Thed N. van (2020): Open Access uptake by universities worldwide. In: *PeerJ*, 8 (Juli), e9410. DOI:10.7717/peerj.9410.

Scheidsteger, Thomas; Haunschild, Robin (2022): Comparison of metadata with relevance for bibliometrics between Microsoft Academic Graph and OpenAlex until 2020. DOI:10.5281/zenodo.6975102.

Schmeja, Stefan; Tullney, Marco (2020): Publikationsmonitoring. DOI:10.34657/3498.

Scholarly Publishing and Academic Resources Coalition (2023): Campus Open Access Fund FAQ. Verfügbar unter https://sparcopen.org/our-work/oa-funds/faq/, zugegriffen am 22.08.2023.

Stoy, Lennart; Morais, Rita; Borrell-Damián, Lidia (2019): Decrypting the Big Deal Landscape: Follow-up of the 2019 EUA Big Deals Survey Report. Verfügbar unter https://www.eua.eu/resources/publications/889:decrypting-the-big-deal-landscape.html, zugegriffen am 22.08.2023.

Vierkant, Paul; Siegert, Olaf; Deinzer, Gernot; Gebert, Agathe et al. (2017): Workflows zur Bereitstellung von Zeitschriftenartikeln auf Open-Access-Repositorien – Herausforderungen und Lösungsansätze. In: *o-bib. Das offene Bibliotheksjournal*, 4, 151–69. DOI:10.5282/O-BIB/2017H1S151-169.

Visser, Martijn; Eck, Nees Jan van; Waltman, Ludo (2021): Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. arXiv. DOI:10.48550/arXiv.2005.10732.

Weingart, Peter (2005): Impact of bibliometrics upon the science system: Inadvertent consequences? In: *Scientometrics*, 62 (1), 117–31. DOI:10.1007/s11192-005-0007-7.

Xia, Feng; Wang, Wei; Bekele, Teshome Megersa; Liu, Huan (2017): Big Scholarly Data: A Survey. In: *IEEE Transactions on Big Data*, 3 (1), 18–35. DOI:10.1109/TBDATA.2016.2641460.

Nick Haupka
Georg-August-Universität Göttingen
Platz der Göttinger Sieben 1
D-37075 Göttingen
nick.haupka@sub.uni-goettingen.de
https://orcid.org/0009-0002-6478-6789