Probleme der Erhebung von Nutzungsstatistiken im Open Access

Pascal-Nicolas Becker, The Library Code GmbH
Yannick Paulsen, The Library Code GmbH

Zusammenfassung

Nutzungsstatistiken fordern in der Regel, dass automatisierte Zugriffe gefiltert und nur Zugriffe in der Statistik erfasst werden, die von einer Person initiiert wurden. Im Closed Access wird der Zugriff auf bestimmte identifizierte Gruppen beschränkt, zum Beispiel durch Freigabe spezieller IP-Adressräume oder Single-Sign-On-Lösungen. Viele Bots haben im Closed Access also keinen Zugriff oder können anhand der Authentifizierung erkannt werden. Open Access sichert zu, dass Inhalte ohne technische und andere Barrieren genutzt werden können. Bei offenen Angeboten ist das Erkennen von Bots nicht manipulationssicher möglich. Hinzu kommt, dass Inhalte im Open Access auch durch andere verbreitet werden dürfen und so weitere Kopien von Arbeiten auf anderen Servern auftauchen, die sich der Zählung entziehen. Eine belastbare Nutzungsstatistik für Inhalte im Open Access lässt sich daher nicht führen. Der exemplarische Fokus im Artikel liegt auf Open-Access-Repositorien; die zugrundeliegenden Probleme mit Nutzungsstatistiken gelten allerdings für alle Open-Access-Publikationsarten, inklusive Verlagspublikationen im Open Access.

Summary

Usage statistics generally require that automated accesses are disregarded and only accesses initiated by a person are counted for the statistics. For closed access content, access is restricted to certain identified groups, for example by using designated IP address ranges or single sign-on solutions. Many bots therefore have no access in closed access or can be easily recognized based on authentication. Open Access ensures that content can be used without technical or other barriers. Therefore, for Open Access content, bots cannot be identified in a tamper-proof manner. In addition, content in Open Access can also be distributed by others, which means that there may be further copies of the work on other servers that cannot be counted. It is therefore not possible to keep reliable usage statistics for Open Access content. This article focuses on Open Access repositories as an example, but the underlying problems with usage statistics apply to all kind of Open Access publications, including publisher publications in Open Access.

Schlagwörter: Nutzungsstatistik; Open Access; Dokumentenserver; COUNTER Code of Practice

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/6109

Autorenidentifikation: Pascal-Nicolas Becker, ORCID: 0000-0003-2169-1261,
Yannick Paulsen, ORCID: 0000-0003-2677-7056

1. Problembeschreibung

Die Bewertung von Inhalten und Services anhand quantitativer Parameter ist eine stark etablierte Praxis. Die Aussagekraft ist dabei oft davon abhängig, die Zahlen in das richtige Verhältnis zu setzen und zu interpretieren. Es gibt aber Sachverhalte, bei denen schon die Erfassung nicht zuverlässig funktioniert. Im Bereich von Open Access kann keine manipulationssichere Nutzungsstatistik geführt werden. In diesem Artikel wird dies exemplarisch am Anwendungsfall Open-Access-Repositorien diskutiert.[1] Wie im Laufe des Textes deutlich werden wird, betreffen die zugrunde liegenden Probleme bei der Erfassung von Nutzungsstatistiken auch Open-Access-Verlage. In der Aufbereitung des Themas werden technische Hintergrundinformationen zur Erfassung von Nutzungsstatistiken und dem vermeintlichen Herausfiltern von automatisierten Zugriffen erläutert, aber auch wissenschaftspolitische Argumente dargelegt.

Bei der Einführung (oder Migration) von Repositorien ist die statistische Erfassung von Nutzungen bzw. Zugriffen oft ein wichtiger Punkt in Ausschreibungen und bei der Softwarewahl. Das ist nachvollziehbar, da zum einen Träger der Bibliotheken in der Regel in solchen Zahlen einen direkten Weg sehen, um zu bewerten, wie gut das jeweilige System angenommen wird und ob sich die finanziellen sowie personellen Investitionen lohnen. Zum anderen bieten Nutzungsstatistiken den einzelnen Autor*innen Informationen über die Sichtbarkeit und die Nutzung ihrer Veröffentlichungen, was die Attraktivität der Nutzung von Repositorien deutlich erhöht. Personen, die diese Statistiken betrachten, haben in der Regel kein Verständnis dafür, dass die Daten nicht repräsentativ, nicht verlässlich und daher nicht hilfreich sind. Die DINI-Arbeitsgruppe Elektronisches Publizieren schrieb bereits 2013:

> „Aus Expertensicht schneidet COUNTER am besten ab, etwaige technische Unzulänglichkeiten spielten dabei eine nur untergeordnete Rolle. Interessanterweise schienen sich die Experten dabei weniger von detaillierten Fakten leiten zu lassen, sondern eher von ihrem Empfinden. Es schienen also nicht die tatsächlichen Fähigkeiten eines Systems im Vordergrund zu stehen, sondern seine Verbreitung und sein Image in den entsprechenden Communities."[2]

Etliche institutionelle Träger drängen auf die Bereitstellung von Nutzungsstatistiken. Auch wenn die genannten Probleme der Filterung maschineller Zugriffe bekannt sind, werden diese oft dennoch bereitgestellt. Die Filterung maschineller Zugriffe verfälscht die Statistiken, weil dabei menschliche Zugriffe fälschlicherweise als maschinelle Zugriffe ausgeschlossen werden, während andere maschinelle Zugriffe nicht als solche erkannt werden. Auch in der Open-Access-Community werden Nutzungsstatistiken oft eingefordert. So heißt es im DINI-Zertifikat von 2022:

> „Das Vorhalten und die Anzeige von (offenen) Metriken kann sowohl qualitativ und quantitativ als auch technologisch die Basis für die Bewertung eines Dienstes sein."[3]

1 Der exemplarische Fokus auf Repositorien bietet sich zum einen aufgrund der beruflichen Hintergründe der Autoren an, zum anderen existiert hier in der Regel eine direktere Einbeziehung und Verantwortlichkeit von bibliothekarischem Personal. Die Probleme sind jedoch für alle Publikationsformen und -dienste identisch.

2 DFG-Projekt „Open-Access-Statistik"; DINI-Arbeitsgruppe „Elektronisches Publizieren": Standardisierte Nutzungsstatistiken für Open-Access-Repositorien und -Publikationsdienste, 2013 (DINI Schriften 13-de). https://doi.org/10.18452/1497.

3 DINI AG Elektronisches Publizieren (E-Pub): DINI-Zertifikat für Open-Access-Publikationsdienste 2022, 2022 (DINI Schriften 3-de). https://doi.org/10.18452/24678.

Zwar können „(offene) Metriken“ nicht auf Nutzungsstatistiken beschränkt werden, das Führen einer eigenen konsistenten Zugriffsstatistik wird jedoch im DINI-Zertifikat als eine Mindestanforderung für das Zertifikat aufgezählt. Weiter fordert das Zertifikat im Falle der öffentlichen Bereitstellung, dass automatisierte Zugriffe herausgefiltert werden müssen. Darunter fallen Crawler z.B. von Suchmaschinen, Bots für das Text- und Data-Mining, Zugriffe von Browserplugins oder sonstiger Software, aber potenziell auch bewusste Manipulationen. Möchte man die direkte menschliche Nutzung messen, müssten alle diese Aktivitäten aus den Nutzungsstatistiken entfernt werden, ohne dabei zugleich Zugriffe durch Lesende fälschlicherweise zu filtern. Im Kontext von Closed-Access-Veröffentlichungen ist dies sehr gut möglich, etwa wenn eine Anmeldung vorausgesetzt wird oder Zugriffe nur aus bestimmten Netzwerken zugelassen sind. Aus Verträgen mit Verlagen sind diese Szenarien den Bibliotheken bekannt, liefern doch viele Verlage solche Statistiken, wenn es um den digitalen Zugriff auf von Bibliotheken lizensiertes Material geht. Diese Erfahrung wird auf das gesamte Publikationswesen verallgemeinert und führt zu der Erwartung, dass Nutzungsstatistiken auch für Open-Access-Inhalte geführt werden und die o.g. zusätzlichen Zugriffsformen dabei herausgefiltert werden können. Bei Open Access ist die Sachlage jedoch eine andere – anonyme direkte Zugriffe auf die Dokumente sind elementarer Bestandteil des Konzepts. Die Budapest Open Access Initiative führt explizit aus:

> „By ‚open access‘ to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.“[4]

Maschinelle Aktivitäten müssten also beim Erstellen einer Nutzungsstatistik auf andere Art und Weise von menschlichen Zugriffen unterschieden und ggf. entfernt werden.

2. COUNTER Code of Practice

Mit der Non-Profit-Organisation „COUNTER“ und ihrem Code of Practice gibt es seit 2003 einen etablierten Standard zur Berichterstattung von digitalen Nutzungszahlen. Bibliotheken werden hier konkret als Zielgruppe benannt, um sie bei Lizenzierungsentscheidungen zu beraten und vergleichbare Statistiken bereitzustellen.[5] Damit entspringt das Projekt zwar dem Anwendungsfall „Closed Access“, passt sich nun aber auch der Anwendung im Open-Access-Bereich an. Das hängt insbesondere damit zusammen, dass kein vergleichbares Pendant für Nutzungsstatistiken von freien digitalen Inhalten existiert.[6] So wird COUNTER etwa auch im DINI-Zertifikat als Standard bei der Veröffentlichung solcher Statistiken festgelegt[7] und setzte sich in einer Expertenumfrage vom Projekt Open-Access-Statistik gegen Alternativen durch.[8]

4 Read the Declaration, Budapest Open Access Initiative, https://www.budapestopenaccessinitiative.org/read/, Stand: 09.06.2024.
5 About Counter, Counter Metrics, https://www.countermetrics.org/about/, Stand: 03.01.2024.
6 Greene, Joseph W.: Developing COUNTER standards to measure the use of Open Access resources, in: Qualitative and Quantitative Methods in Libraries 6 (2), 2017, S. 315–320. https://www.qqml-journal.net/index.php/qqml/article/view/410, Stand: 03.01.2024.
7 DINI AG Elektronisches Publizieren (E-Pub): DINI-Zertifikat, 2022, S. 44f.
8 DFG-Projekt „Open-Access-Statistik”; DINI-Arbeitsgruppe „Elektronisches Publizieren“: Standardisierte Nutzungsstatistiken, 2013.

Die zwei etabliertesten Ansätze zur Generierung von Nutzungsdaten sind laut dem Release 5.1 des COUNTER Code of Practice „Log File Analysis“ und „Page Tagging“. Log File Analysis nutzt die Logdateien, die bei Transaktionen mit der Website entstehen. Die Methode ist browserunabhängig, beinhaltet allerdings auch die Daten von Bot-Aktivitäten. Webseiten, die im Cache geladen sind, werden nicht mit erfasst.[9] Page Tagging übermittelt die Nutzung beim Laden einer Seite im Browser und umfasst somit auch Seiten aus dem Cache. Dafür wird Page Tagging von manchen Browsern blockiert, sodass in diesen Fällen die jeweilige Nutzung komplett unberücksichtigt bleibt. Die Methode arbeitet üblicherweise mit der Vergabe von Cookies.[10]

Bei der Verarbeitung der durch diese beiden Methoden vorliegenden Nutzungsdaten werden unter anderem maschinelle Aktivitäten herausgefiltert. Für „Internet Robots“ und Crawler wird zu diesem Zweck eine Liste von Bots[11] geführt, die als Ausschlussliste gehandhabt wird.[12] Ein weiterer Fall von maschinellen Zugriffen ist das Text- und Data-Mining (TDM). Die eigentlichen Prozesse des TDM finden nach dem Download des Volltextes statt. Um aber die damit verbundenen Seitenaufrufe und -downloads identifizieren zu können, ist man auch laut COUNTER auf vorherige Absprachen mit den Institutionen angewiesen, die das TDM durchführen wollen.[13] Insbesondere mit dem Training von Large Language Modells, wie zum Beispiel ChatGPT, gab es Diskussionen und Gerichtsverfahren zum Abruf der Daten durch Robots.[14] Des Weiteren sollen maschinelle Zugriffe z.B. von Massen-Download-Tools oder Literaturverwaltungssoftware von den COUNTER-Reports ausgeschlossen werden.[15]

Das Reduzieren der Nutzungsdaten auf Aktivitäten von menschlichen Nutzenden ist somit auf die Kenntnis der jeweiligen Bots, Softwareprodukte und Institutionen angewiesen. Eine tatsächliche Trennung von Menschen und Maschinen gibt es bei der Datenerfassung nicht, unabhängig davon, ob es sich um Open-Access-Inhalte in Repositorien oder bei Verlagen handelt.

3. Handhabung in Open-Access-Repositorien

In der Praxis wird die Analyse und Aufbereitung von Nutzungsdaten in der Regel nicht durch die Manager*innen der Repositorien durchgeführt. Stattdessen gibt es entweder integrierte Lösungen in der jeweiligen Software oder es wird auf externe Produkte und Dienstleister zurückgegriffen.[16]

9 6.1 Log File Analysis, COUNTER Code of Practice Release 5.1, https://cop5.projectcounter.org/en/5.1/06-logging/01-log-file-analysis.html, Stand: 03.01.2024.

10 6.2 Page Tagging, COUNTER Code of Practice Release 5.1, https://cop5.projectcounter.org/en/5.1/06-logging/02-page-tagging.html, Stand: 03.01.2024.

11 COUNTER-Robots, GitHub, https://github.com/atmire/COUNTER-Robots, Stand: 03.01.2024.

12 7.8 Internet Robots and Crawlers, COUNTER Code of Practice Release 5.1, https://cop5.projectcounter.org/en/5.1/07-processing/08-internet-robots-and-crawlers.html, Stand: 03.01.2024.

13 7.10 Text and Data Mining, COUNTER Code of Practice Release 5.1, https://cop5.projectcounter.org/en/5.1/07-processing/10-text-and-data-mining.html, Stand: 03.01.2024.

14 „New York Times“ verklagt OpenAI und Microsoft wegen ChatGPT, Süddeutsche Zeitung, 27.12.2023. https://www.sueddeutsche.de/wirtschaft/kuenstliche-intelligenz-new-york-times-verklagt-openai-und-microsoft-wegen-chatgpt-dpa.urn-newsml-dpa-com-20090101-231227-99-420552, Stand: 09.06.2024.

15 7.9 Tools and Features that Enable Bulk Downloading, COUNTER Code of Practice Release 5.1, https://cop5.projectcounter.org/en/5.1/07-processing/09-tools-and-features-that-enable-bulk-downloading.html, Stand: 03.01.2024.

16 Shearer, Kathleen; Nakano Koga, Silvia Mirlene; Rodrigues, Eloy u.a.: Current State and Future Directions for Open Repositories in Europe, 2023. https://doi.org/10.5281/zenodo.10255559.

Die in Deutschland verbreitete Repositoriensoftware OPUS kann zur Analyse der eigenen Webserver-Logs AWStats[17] oder ePusta[18] nutzen.[19] Das Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) und der Kooperative Bibliotheksverbund Berlin-Brandenburg (KOBV) entwickeln OPUS zusammen und bieten breitflächig Hostings an. Nach Aussage des BSZ werden Suchmaschinen bei den Nutzungsstatistiken herausgefiltert.[20] AWStats nutzt hierfür eine Datenbank von robotstxt.org.[21] Auf der Webseite zu OPUS gibt der KOBV an, dass ePusta für OPUS 4 angepasst wurde und nach dem COUNTER-Standard arbeitet.[22] Demnach werden vermutlich auch die Filterlisten des COUNTER-Standards genutzt.

In der weltweit meistgenutzten Repositoriensoftware DSpace[23] hat man die Wahl zwischen zwei Arten der Erfassung von Nutzungsdaten. Zum einen gibt es eine DSpace-interne Variante über Solr-Statistics. Das Herausfiltern von Spider-Bots findet hier ebenfalls über einen Abgleich bestehender Listen statt.[24] Die von DSpace standardmäßig genutzten Listen wurden zuletzt vor mehreren Jahren aktualisiert, was Zweifel an der Zuverlässigkeit dieser Methode bestärkt.[25] Zum anderen gibt es die Möglichkeit der Aktivierung von Google Analytics. Dieses bietet „Known bot-traffic exclusion" über eine Kombination von eigener Forschung und der kostenpflichtigen „International Spiders and Bots List".[26] Google Analytics wird als zuverlässigstes Werkzeug für Nutzungsstatistiken bewertet mit Datenschutzbedenken als hauptsächlichen Abstrich.[27]

Besonders prominent werden Nutzungsstatistiken im institutions- und fachunabhängigen Repositorium Zenodo positioniert. Die inzwischen über vier Millionen Open-Access-Dokumente umfassende Plattform positioniert die Zahlen für die Downloads und die Views des einzelnen Items in der Record-Ansicht direkt neben dem Titel. Die Nutzungsstatistiken sind nicht deaktivierbar und werden in aggregierter Form an die Dienste OpenAIRE Usage Counts[28] und DataCite Sashimi[29] weitergereicht. Das Herausfiltern von maschinellen Zugriffen findet auch hier über den Abgleich mit den Listen des COUNTER-Projekts und von „Make Data Count"[30] statt.[31]

17 What is AWStats, AWStats official web site, http://www.awstats.org/, Stand: 03.01.2024.
18 ePuSta-logfileparser, GitHub, https://github.com/gbv/ePuSta-logfileparser sowie ePuSta-Server, GitHub, https://github.com/gbv/ePuSta-Server, Stand: 03.01.2024.
19 OPUS 4 – Repository Software, Kooperativer Bibliotheksverbund Berlin-Brandenburg, https://www.kobv.de/entwicklung/software/opus-4/, Stand: 03.01.2024.
20 Statistik, BSZ-Wiki, https://wiki.bsz-bw.de/display/OPUS/Statistik, Stand: 03.01.2024.
21 AWStats logfile analyzer 7.4 Documentation, 14.07.2015. https://awstats.sourceforge.io/docs/awstats.pdf, Stand: 03.01.2024.
22 OPUS 4 – Repository Software.
23 OpenDOAR Statistics, OpenDOAR, https://v2.sherpa.ac.uk/view/repository_visualisations/1.html, Stand: 03.01.2024.
24 SOLR Statistics, DSpace 7.x Documentation, zuletzt geändert am 15.07.2024, https://wiki.lyrasis.org/display/DSDOC7x/SOLR+Statistics#SOLRStatistics-ConfigurationsettingsforStatistics, Stand: 19.11.2024.
25 IP Addresses of Search Engine Spiders, https://www.iplists.com, Stand: 29.11.2024.
26 [GA4] Known bot-traffic exclusion, Analytics Help, https://support.google.com/analytics/answer/9888366?hl=en, Stand: 03.01.2024.
27 Obrien, Patrick; Arlitsch, Kenning; Mixter, Jeff u.a.: RAMP – the Repository Analytics and Metrics Portal. A prototype web service that accurately counts item downloads from institutional repositories, in: Library Hi Tech 35 (1), 2017, S. 144–158. https://doi.org/10.1108/LHT-11-2016-0122 ;
Perrin, Joy M.; Yang, Le; Barba, Shelley u.a.: All that glitters isn't gold. The complexities of use statistics as an assessment tool for digital libraries, in: The Electronic Library 35 (1), 2017, S. 185–197. https://doi.org/10.1108/EL-09-2015-0179.
28 UsageCounts Service by OpenAIRE, https://usagecounts.openaire.eu/, Stand: 03.01.2024.
29 sashimi, GitHub, https://github.com/datacite/sashimi, Stand: 03.01.2024.
30 Make Data Count, https://makedatacount.org/, Stand: 03.01.2024.
31 Frequently Asked Questions, Zenodo, https://help.zenodo.org/faq/#statistics, Stand: 03.01.2024.

Es zeigt sich, dass die verwendeten Listen zwar zum Teil variieren, aber ein Abgleich mit bereits bekannten Bots bis jetzt der einzige etablierte Weg zum Herausfiltern maschineller Aktivitäten ist. Insbesondere der häufige Bezug auf COUNTER, ein Standard, der explizit für Closed-Access-Content konzipiert ist[32], verdeutlicht die methodische Leerstelle, die hier herrscht. Diese Art der Erkennung setzt voraus, dass der Bot die entsprechende User-Kennung sendet. In der Vergangenheit gab es Hinweise darauf, dass sich nicht alle Bots daran halten und zum Teil auch die robots.txt ignorieren, eine Datei, die es Serverbetreibern ermöglichen soll, Zugriffe von Bots zu steuern.[33] Firmen, die maschinelle Zugriffe verursachen, haben ein Interesse daran, dass diese nicht identifizierbar sind. Wären zum Beispiel die Zugriffe von Suchmaschinen leicht zu erkennen, wäre es möglich, der Suchmaschine andere Informationen bereitzustellen als menschlichen Nutzenden. Das würde eine entsprechende Manipulation der Suchindizes ermöglichen, die die Suchmaschinenbetreiber verhindern müssen, um sich vor Spam zu schützen und eine hohe Relevanz ihrer Suchergebnisse zu erhalten. Daher dürften die meisten Suchmaschinen mindestens Stichproben durchführen, bei denen sie möglichst genau Abrufe „durchschnittlicher" Nutzender simulieren und die Ergebnisse mit den Abrufen ihrer Bots vergleichen (Stichwort Cloaking).[34] Die tatsächliche Erkennung solcher Zugriffe ist wahrscheinlich nur durch Heuristiken oder mit Machine-Learning-Praktiken auf große Mengen an Metadaten über Datenverkehr möglich, auf denen man Vergleiche durchführen kann. Über die entsprechenden Informationen und technischen Möglichkeiten zur Erhebung und Analyse von Nutzungsstatistiken verfügen vermutlich nur wenige sehr große Tech-Firmen, mit denen die Zusammenarbeit für europäische Institutionen in der Regel datenschutzrechtliche Bedenken und Hürden aufwirft.

4. Datenschutzprobleme bei der Nutzung externer Dienstleister für Nutzungsanalysen

Eine häufig genutzte Alternative zu den leicht manipulierbaren und schwierig aktuell zu haltenden Abgleichlisten ist der Service Google Analytics.[35] Google gehört zu den Firmen, die über große Mengen an Metadaten zum Datenverkehr im Internet verfügen[36], um Analysen auf Basis von Verkehrsdaten über viele verschiedene Websites hinweg durchführen zu können. Nach eigenen Angaben erkennt Google Analytics Bots „mithilfe einer Kombination aus Forschungsdaten von Google und der „International Spiders and Bots List" des Interactive Advertising Bureau (IAB)".[37]

Vor einer Integration von Google Analytics stellen sich zum einen rechtliche und insbesondere datenschutzrechtliche Fragen, welche es separat zu behandeln gilt. Zum anderen schließt sich eine Nutzung von Google Analytics aus inhaltlichen Gründen aus. Datentracking und das dahinter liegende Geschäftsmodell beschäftigt Bibliotheken übergreifend in verschiedenen Bereichen. Dies liegt an der

32 Perrin, Joy M.; Yang, Le; Barba, Shelley u.a.: All that glitters isn't gold, 2017.

33 Googlebot und andere Google-Crawler prüfen, https://developers.google.com/search/docs/crawling-indexing/verifying-googlebot?hl=de, Stand: 22.07.2024; Weiß, Eva-Maria: Crawler ohne Grenzen. Perplexity ignoriert robots.txt, heise online, 20.06.2024, https://www.heise.de/news/Crawler-ohne-Grenzen-Perplexity-ignoriert-robots-txt-9770336.html, Stand: 25.07.2024.

34 Cloaking, https://de.wikipedia.org/wiki/Cloaking, Stand: 09.09.2024.

35 Obrien, Patrick; Arlitsch, Kenning; Mixter, Jeff u.a.: RAMP – the Repository Analytics and Metrics Portal, 2017.

36 Müller, Bernd: USA prüfen radikale Schritte gegen Google-Dominanz, Telepolis, 21.11.2024, https://www.telepolis.de/features/Google-Imperium-vor-dem-Zerfall-Chrome-und-Android-auf-der-Verkaufsliste-10082271.html, Stand: 10.12.2024.

37 [GA4] Ausschluss des Traffics von bekannten Bots, Google Analytics-Hilfe, https://support.google.com/analytics/answer/9888366?hl=de, Stand: 09.06.2024.

Attraktivität von Informationen über wissenschaftliche Aktivitäten und das Verhalten einzelner Wissenschaftler*innen sowie dem Zugang, den wissenschaftliche Bibliotheken über die Systeme, die sie anbieten, ermöglichen können. So wird die Transformation von großen Wissenschaftsverlagen zu „Data Analytics Businesses" vielfach beschrieben – unter anderem in einem Informationspapier der Deutschen Forschungsgemeinschaft (DFG).[38]

Das Ausmaß des Datentrackings und der damit verbundenen Geschäfte wird durch ein öffentlich gewordenes Dokument von Microsofts Xandr im Ansatz deutlich. Xandr gehört zu den weltweit größten Datenmarktplätzen und präsentiert in einer Angebotsliste über 650.000 Kategorien, in die Menschen einsortiert werden. Netzpolitik.org schreibt in einem ausführlichen Artikel zum veröffentlichten Dokument und den daraus zu gewinnenden Informationen:[39] So sei Datenerfassung ein allgegenwärtiger Zustand, welcher allerdings in seinem Vorgehen sehr undurchsichtig ist. Die Angebotsliste von Xandr zeigt nun, wie detailliert die gesammelten Informationen sind, welche jeden möglichen Lebensaspekt umfassen und aus mehreren Quellen zusammengeführt werden.

Hier stehen Bibliotheken in besonderer Verantwortung. Nutzende müssen darauf vertrauen können, die durch Bibliotheken bereitgestellten Systeme zu benutzen, ohne unbewusst getrackt und Teil eines groß angelegten Datenhandels zu werden, insbesondere wenn es sich um die Bibliothek des Arbeitgebers bzw. Dienstherrn handelt. Bei Forschenden gehören Rechercheportale, Publikationsplattformen und Informationssysteme zur beruflich notwendigen Infrastruktur. Sie sind den Verträgen, die unter anderem von Bibliotheken oder deren Trägern ausgehandelt werden, ausgeliefert. Dabei geht es um grundlegende Schutzrechte und teilweise Sicherheitsfragen mit übergeordneten Folgen, auch bezogen auf das private Leben der Forschenden. Wenn das europäische akademische Umfeld Datenschutz und Anti-Tracking fordert,[40] dann muss dies auch in den von ihren Institutionen angebotenen Infrastrukturen umgesetzt werden.

5. Technisches Fazit

Schon 2013 schrieb die DINI-Arbeitsgruppe Elektronisches Publizieren zusammen mit den Verantwortlichen des Projekts Open-Access-Statistik, dass man nur einen Teil der Roboter-Zugriffe präzise eliminieren kann und bei den übrigen nicht mit Sicherheit beurteilen kann, ob es sich nicht doch um einen menschlichen Zugriff handelt. Als Fazit wurde daraus gezogen, dass man bei der Filterung der Nutzungsdaten gleiche Kriterien verwenden sollte, um vergleichbare Statistiken zu haben.[41] Standardisierte Nutzungsstatistiken sind jedoch nicht automatisch vergleichbare Nutzungsstatistiken. Es gehört zum zentralen Nutzen solcher Statistiken, dass man die Zugriffe auf einen Artikel mit denen

38 DFG-Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme: Datentracking in der Wissenschaft. Aggregation und Verwendung bzw. Verkauf von Nutzungsdaten durch Wissenschaftsverlage. Ein Informationspapier des Ausschusses für Wissenschaftliche Bibliotheken und Informationssysteme der Deutschen Forschungsgemeinschaft. https://doi.org/10.5281/zenodo.5900759; Siems, Renke: Das Lesen der Anderen, in: O-Bib. Das Offene Bibliotheksjournal 9 (1), 2022, S. 1–25. https://doi.org/10.5282/o-bib/5797.

39 Dachwitz, Ingo: Microsofts Datenmarktplatz Xandr: das sind 650.000 Kategorien, in die uns die Online-Werbeindustrie einsortiert, netzpolitik.org, 08.06.2023, https://netzpolitik.org/2023/microsofts-datenmarktplatz-xandr-das-sind-650-000-kategorien-in-die-uns-die-online-werbeindustrie-einsortiert/#netzpolitik-pw, Stand: 03.01.2024.

40 Stop Tracking Science, https://stoptrackingscience.eu/, Stand: 10.06.2024.

41 DFG-Projekt „Open-Access-Statistik"; DINI-Arbeitsgruppe „Elektronisches Publizieren": Standardisierte Nutzungsstatistiken, 2013.

auf einen anderen Artikel, eines Monats mit denen eines anderen Monats und eines Repositoriums mit denen eines anderen Repositoriums vergleichen kann. Mit einer unklaren Menge an maschinellen Aktivitäten und keiner Methode der tatsächlichen Identifizierung kann man zwar auf vorhandene Nutzung schließen, sie aber nicht beziffern und somit auch nicht vergleichen.

Dazu kommen die oben dargestellten Versuche, maschinelle Zugriffe nicht identifizierbar zu halten (Stichwort Cloaking), die sich damit der Erkennung als maschinelle Zugriffe möglichst entziehen.[42] Auch der Bedarf, möglichst viele Daten möglichst schnell für Projekte zur Künstlichen Intelligenz, zum Beispiel für den Aufbau von Large Language Models zu sammeln, schafft schon jetzt Probleme beim Betrieb von Repositorien.[43] Es ist anzunehmen, dass das Verschleiern von Bots dabei technisch eher ausgefeilter werden wird.

In der Softwareentwicklung werden seit Jahren automatisierte Tests mit den gängigen Browsern Chrome, Edge und Firefox durchgeführt. Die Bedienung ist dabei automatisiert.[44] Mit solchen Techniken lassen sich Zugriffe auf Websites ausführen, die sich von Zugriffen von Browsern, die von Menschen gesteuert werden, nicht unterscheiden lassen. Da viele Bots JavaScript aus Sicherheitsgründen nicht auswerten, wurde eine Zeit lang JavaScript für die Erstellung von Nutzungsstatistiken eingesetzt. Automatisch gesteuerte Browser, in denen JavaScript ausgeführt wird, können auch dieses Vorgehen nicht als automatisierte Zugriffe erkennen. Spammer verwenden automatisierte Browser heutzutage oft nicht, da der Betrieb zu aufwändig ist, um Spameinträge in Formulare einzufügen. Zur Manipulation von Nutzungsstatistiken reichen in der Regel aber Aufrufzahlen in geringer vierstelliger Höhe aus, signifikant weniger, als im Bereich von automatisiertem Spam benötigt wird. So hätten zum Beispiel im August 2024 knapp 2.000 Aufrufe gereicht, um ein Dokument zum meist abgerufenen Dokument im Econstore zu machen, einem fachspezifischen Repositorium der Wirtschaftswissenschaften. Laut Statistik wurde im August 2024 das am stärksten abgefragte Dokument im Econstore 1979 mal heruntergeladen.[45]

Problematisch sind daneben auch die Potenziale von gezielten Manipulationen: Repositorien, die besonders stark nachgefragte Inhalte hervorheben, geben damit womöglich einen Anreiz für Manipulationsversuche. Die schlechte Aussagekraft solcher Nutzungsstatistiken sollte also zumindest verdeutlicht werden. Mit den vorliegenden Erkenntnissen sollte dringend erwogen werden, auf die Erfassung von „Vollanzeigen von digitalen Einzeldokumenten“, darunter: „Vollanzeige von Einzeldokumenten auf dem institutionellen Repositorium“ in der Deutschen Bibliotheksstatistik DBS[46] zu verzichten.

42 Cloaking, https://de.wikipedia.org/wiki/Cloaking, Stand: 09.09.2024.

43 Sherrick, A. K.; Navarro, D. A. Pino: Creating a better balance. the need for tools and practices to combat AI harvests and resource flooding in repository environments, 2024. https://doi.org/10.5281/zenodo.12579304.

44 Selenium, https://www.selenium.dev/, Stand: 09.06.2024.

45 EconStor-Nutzungsstatistik: Gesamt-Downloads, https://www.econstor.eu/esstatistics/10419/0?year=2024&month=08, Stand: 09.09.2024.

46 DBS – Deutsche Bibliotheksstatistik, https://www.bibliotheksstatistik.de/, Stand: 09.06.2024.

6. Die grundsätzliche Fragwürdigkeit von Nutzungsstatistiken

Auch unabhängig von den technischen Schwierigkeiten, maschinelle Zugriffe herauszufiltern, gibt es Gründe, Nutzungsstatistiken kritisch zu bewerten. So finden Aufrufe und Downloadzahlen auf teilweise sehr unterschiedlichen quantitativen Ebenen statt, je nachdem welcher Disziplin die Publikation zugeordnet wird und welches Thema behandelt wird. Kleine Nischenfächer oder sehr spezifische Forschungsprojekte finden automatisch weniger Publikum. Es ist nicht zu vertreten, wenn diese Nutzungsstatistiken dann eine Argumentationsgrundlage bei der Leistungsbewertung von Repositorien werden, um etwa Services infrage zu stellen, die es kleinen Nischenfächern ermöglichen, ihre Inhalte im Open Access zu publizieren.

Neben der Präsentation von Nutzungsstatistiken für das Gesamtsystem sind auch immer wieder Publikationssysteme zu finden, welche Platz auf ihrer Startseite verwenden, um dort die Inhalte mit den meisten Zugriffen zum Beispiel des letzten Monats besonders hervorzuheben. Die Entscheidung, unabhängig von der wissenschaftlichen Relevanz und Qualität die Publikationen nach vorne zu stellen, die die größte Popularität erreichen, entspricht nicht maßgeblichen wissenschaftlichen Prinzipien und verführt dazu, Aufrufe und Downloads als zentrales Bewertungskriterium zu etablieren. Es verzerrt die Anreize und stellt vor allem bereits etablierte Wissenschaftler*innen von großen Fachbereichen nach vorne. Auf diese Art und Weise wird angemesseneren Präsentationsprinzipien, wie etwa neuen Einträgen im Repositorium, der Raum genommen.

Abgesehen von den bereits dargestellten Problemen, Nutzungsstatistiken für Inhalte im Open Access zu erheben, gibt es noch ein sehr grundsätzliches Argument. Ziel von Open Access ist es, Inhalte möglichst weit und frei zu verbreiten, vorzugsweise unter einer Creative-Commons-Lizenz Namensnennung (CC-BY), die es explizit zulässt, dass Inhalte durch Dritte weiterverbreitet werden. Wie viel Aussagekraft hat eine Nutzungsstatistik in einem Angebot, wenn der Inhalt auch auf anderen Servern zu finden ist und sich ein Teil der Nutzungszahlen damit der eigenen Statistik entzieht?

Eine Ebene der Zugriffsstatistiken wurde hier ausgespart: Für den Betrieb technischer Dienste sind Metriken heutzutage sehr hilfreich. Zum Beispiel kann die Gesamtzahl erfolgreicher Zugriffe auf einen Dienst pro Minute und Veränderungen dieser Zahl ein wichtiger Indikator für die Stabilität und Performanz der technischen Infrastruktur sein. Hierbei dürfen automatisierte und rein maschinelle Zugriffe aber nicht ausgefiltert werden. Und jeglichen Rückschluss auf die wissenschaftliche Bedeutung eines Services oder die Popularität einzelner Inhalte lassen diese Zahlen eben nicht zu.

Die aufgezeigten Probleme mit Nutzungsstatistiken im Open Access sind technischer und inhaltlicher Natur. Sie sind nicht spezifisch für den genutzten Publikationsdienst, sondern spezifisch für Open Access. Sie treffen auf Open-Access-Verlage genauso zu wie auf Repositorien, Journal Systems oder andere Dienste, die Inhalte im Open Access bereitstellen. Auf eine auf Zugriffszahlen basierende Argumentation und auf eine Veröffentlichung von Nutzungsstatistiken sollte verzichtet werden.

Literaturverzeichnis

AWStats logfile analyzer 7.4 Documentation, 14.07.2015. https://awstats.sourceforge.io/docs/awstats.pdf, Stand: 03.01.2024.

Dachwitz, Ingo: Microsofts Datenmarktplatz Xandr: das sind 650.000 Kategorien, in die uns die Online-Werbeindustrie einsortiert, netzpolitik.org, 08.06.2023, https://netzpolitik.org/2023/microsofts-datenmarktplatz-xandr-das-sind-650-000-kategorien-in-die-uns-die-online-werbeindustrie-einsortiert/#netzpolitik-pw, Stand: 03.01.2024.

DFG-Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme: Datentracking in der Wissenschaft. Aggregation und Verwendung bzw. Verkauf von Nutzungsdaten durch Wissenschaftsverlage. Ein Informationspapier des Ausschusses für Wissenschaftliche Bibliotheken und Informationssysteme der Deutschen Forschungsgemeinschaft. https://doi.org/10.5281/zenodo.5900759.

DFG-Projekt „Open-Access-Statistik"; DINI-Arbeitsgruppe „Elektronisches Publizieren": Standardisierte Nutzungsstatistiken für Open-Access-Repositorien und -Publikationsdienste, 2013 (DINI Schriften 13-de). https://doi.org/10.18452/1497.

DINI AG Elektronisches Publizieren (E-Pub): DINI-Zertifikat für Open-Access-Publikationsdienste 2022, 2022 (DINI Schriften 3-de). https://doi.org/10.18452/24678.

Greene, Joseph W.: Developing COUNTER standards to measure the use of Open Access resources, in: Qualitative and Quantitative Methods in Libraries 6 (2), 2017, S. 315–320. https://www.qqml-journal.net/index.php/qqml/article/view/410, Stand: 03.01.2024.

Müller, Bernd: USA prüfen radikale Schritte gegen Google-Dominanz, Telepolis, 21.11.2024, https://www.telepolis.de/features/Google-Imperium-vor-dem-Zerfall-Chrome-und-Android-auf-der-Verkaufsliste-10082271.html, Stand: 10.12.2024.

„New York Times" verklagt OpenAI und Microsoft wegen ChatGPT, Süddeutsche Zeitung, 27.12.2023. https://www.sueddeutsche.de/wirtschaft/kuenstliche-intelligenz-new-york-times-verklagt-openai-und-microsoft-wegen-chatgpt-dpa.urn-newsml-dpa-com-20090101-231227-99-420552, Stand: 09.06.2024.

Obrien, Patrick; Arlitsch, Kenning; Mixter, Jeff u.a.: RAMP – the Repository Analytics and Metrics Portal. A prototype web service that accurately counts item downloads from institutional repositories, in: Library Hi Tech 35 (1), 2017, S. 144–158. https://doi.org/10.1108/LHT-11-2016-0122.

OPUS 4 – Repository Software, Kooperativer Bibliotheksverbund Berlin-Brandenburg, https://www.kobv.de/entwicklung/software/opus-4/, Stand: 03.01.2024.

Perrin, Joy M.; Yang, Le; Barba, Shelley u.a.: All that glitters isn't gold. The complexities of use statistics as an assessment tool for digital libraries, in: The Electronic Library 35 (1), 2017, S. 185–197. https://doi.org/10.1108/EL-09-2015-0179.

Shearer, Kathleen; Nakano Koga, Silvia Mirlene; Rodrigues, Eloy u.a.: Current State and Future Directions for Open Repositories in Europe, 2023. https://doi.org/10.5281/zenodo.10255559.

Sherrick, A. K.; Navarro, D. A. Pino: Creating a better balance. the need for tools and practices to combat AI harvests and resource flooding in repository environments, 2024. https://doi.org/10.5281/zenodo.12579304.

Siems, Renke: Das Lesen der Anderen, in: O-Bib. Das Offene Bibliotheksjournal 9 (1), 2022, S. 1–25. https://doi.org/10.5282/o-bib/5797.

Weiß, Eva-Maria: Crawler ohne Grenzen.: Perplexity ignoriert robots.txt, heise online, 20.06.2024, https://www.heise.de/news/Crawler-ohne-Grenzen-Perplexity-ignoriert-robots-txt-9770336.html, Stand: 25.07.2024.