Ulrich Herb

Viele Daten, hohe Hürden: Eine Bilanz aus dem Projekt Open-Access-Statistik

Great numbers of data, high obstacles: Results of the project Open Access statistics

http://doi.org/10.1515/bd-2018-0034

Zusammenfassung: Dieser Artikel referiert die Ergebnisse des Projekts Open-Access-Statistik, dessen Ziel es war, standardisierte Nutzungszahlen für wissenschaftliche Dokumente zu erheben. Die gesammelten Nutzungsdaten sollten in erster Linie dazu dienen, Impact-Werte für Open-Access-Dokumente zu ermitteln. Das Projekt sah auch die Implementierung anspruchsvollerer Verfahren wie Netzwerk-Analysen vor, sah sich jedoch mit komplexen rechtlichen Anforderungen konfrontiert. Der Beitrag versucht überdies, Open-Access-Statistik und Altmetrics in Beziehung zu setzen.

Schlüsselwörter: Nutzungsdaten, Impact, Datenschutz

Abstract: This article presents the results of the project Open Access statistics which aimed to collect standardised usage numbers for scientific documents. The usage data were primarily collected to determine impact values for Open Access documents. The project also planned to implement more ambitious processes like network analyses, but was confronted with complex legal demands. Moreover, the article tries to relate Open Access statistics to altmetrics.

Keywords: usage data, impact, data protection

1 Einleitung

Es fehlt nicht an wissenschaftlichen Informations- und Kommunikationsinfrastrukturen oder Fachportalen, jedoch finden nicht alle von ihnen den erhofften

Dr. Ulrich Herb: u.herb@sulb.uni-saarland.de

Open Access. © 2018 Ulrich Herb, publiziert von De Gruyter. BY-NC-ND Dieses Werk ist lizenziert unter der Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 Lizenz.

Zuspruch der wissenschaftlichen Communities. Im Sinne eines Monitoring und der Erfassung dieses Zuspruchs stellt sich bei Fachportalen die Frage, wie man Nutzungszahlen erheben und Nutzungsstatistiken erstellen kann, um die Akzeptanz und – wenn man so formulieren mag – den Impact des Angebots zu bestimmen. Dieser Beitrag berichtet vom Projekt Open-Access-Statistik und seinen Versuchen, alternativen, nutzungsdaten-basierten Impact in einem Netzwerk verteilter Server bereitzustellen.

2 Die Motivation

Nimmt man die Erklärung der Budapest Open Access Initiative im Jahr 2001 als mediales Geburtsjahr des Open Access, war der entgeltfreie Zugang zu wissenschaftlichen Informationen 2006 gerade erstmal fünf Jahre alt. Er galt vielen Wissenschaftlern, die nicht an der eher akademischen Diskussion um die Befreiung des Wissens teilnahmen, zu dieser Zeit als nicht sonderlich erstrebenswerte Publikationsoption.

Tatsächlich konnte Open Access Wissenschaftlern seinerzeit vor allem eines nicht bieten: Impact. Impact, oder verallgemeinernd die Resonanz einer wissenschaftlichen Publikation, wurde damals ausschließlich über durch kommerzielle Datenbanken, wie das Web of Science oder Scopus, berechnete Zitationshäufigkeiten (z. B. mittels des Hirsch- oder h-Index) oder -raten (z. B. über den Journal Impact Factor) zu ermitteln versucht. Open Access jedoch war von der zitationsbasierten Impactmessung ausgeschlossen: Open-Access-Repositorien, da sie in den Selection Criteria der Datenbanken keine Erwähnung fanden (und finden), Open-Access-Journale, da sie meist Neugründungen waren und von Datenbanken wegen mangelnder Publikationshistorie und Zitationshöhen ignoriert wurden. Die Scopus- und Web of Science-Alternative Google Scholar indizierte Open-Access-Material zwar, war allerdings erst seit 2004 als Beta-Version verfügbar und galt zu dieser keinesfalls als Konkurrent der etablierten Datenbanken. Besonders der Scope des indizierten Materials galt als kaum definierbar. Mayr & Walter[1] etwa konstatierten 2007: „[Google Scholar] can be recommended only with some limitation due to a lot of inconsistencies and vagueness (...) in the data“. Da Wissenschaftler aber (damals wie heute) bei der Entscheidung, ob sie Open Access oder Closed Access publizieren, weitgehend indifferent sind und vor allem die

1 Mayr, P., & Walter, A.-K. (2007). An exploratory study of Google Scholar. *Online Information Review*, *31*(6), pp. 814–830, p. 828. http://doi.org/10.1108/14684520710841784 [Zugriff: 20.12.2017].

sich in Impact (vulgo Zitationen) manifestierende Reputation eines Journals das prominenteste Entscheidungskriterium ist, war Open Access keine sehr attraktive Wahl.

Um diesen evidenten Impact- und Attraktivitätsnachteil des Open Access ausgleichen, verfielen die späteren Partner des Projekts Open-Access-Statistik auf die Idee, Nutzungsstatistiken und -daten wissenschaftlicher Dokumente als Impact-Indikatoren nutzen zu wollen. Initial hierfür waren Studien, denen zufolge Open-Access-Publikationen häufiger heruntergeladen und zitiert wurden als Closed-Access-Dokumente (z. B. Brody, Harnad, & Carr[2]; Moed[3]). Noch elaboriertere Anwendungsszenarien führten Johan Bollen und seine Kollegen vor Augen: Bollen, Luce, Vemulapalli, & Xu[4] stellten dar, inwiefern Nutzungsdaten in der Lage sind, zukünftige Forschungstrends vorherzusagen. Bollen, Van De Sompel, Smith, & Luce[5] arbeiteten heraus, dass Nutzungsdaten auf eine zu Zitationen komplementäre Art Resonanz messen, da sie das Verhalten von Lesern erfassen, wohingegen Zitationen Weiterverwendungen durch Autoren beschreiben. Bollen et al.[6] belegten, dass mittels durch Nutzungsdaten gewonnener Clickstreams und sozialer Netzwerkanalyse die Bedeutung einzelner wissenschaftlicher Zeitschriften für ihre Disziplin ermittelt werden kann. Clickstreams bezeichnen Abfolgen von Seitenaufrufen eines Nutzers im WWW und soziale Netzwerkanalyse beschreibt soziale Beziehungen anhand von Interaktionen (z. B. Kommunikation). In der Kombination beider Verfahren modellieren Bollen et al.[7] Journale als Entitäten, deren Beziehungen durch Clickstreams beschrieben werden. Im Ergebnis berechnen die Forscher Kennziffern, die z. B. die Zen-

2 Brody, T., Harnad, S., & Carr, L. (2006). Earlier Web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, *57*(8), pp. 1060–1072. http://doi.org/10.1002/asi.20373 [Zugriff: 20.12.2017].

3 Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, *56*(10), pp. 1088–1097. http://doi.org/10.1002/asi.20200 [Zugriff: 20.12.2017].

4 Bollen, J., Luce, R., Vemulapalli, S., & Xu, W. (2003). Detecting Research Trends in Digital Library Readership (pp. 24–28). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-540-45175-4_3 [Zugriff: 20.12.2017].

5 Bollen, J., Van De Sompel, H., Smith, J. A., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management*, *41*(6), pp. 1419–1440. http://doi.org/10.1016/j.ipm.2005.03.024 [Zugriff: 20.12.2017].

6 Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., & Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. *PloS One*, *4*(3), e4803. http://doi.org/10.1371/journal.pone.0004803 [Zugriff: 20.12.2017].

7 Wie Anm. 6.

tralität eines Journals in einem aus Clickstreams gebildeten Netzwerk aus Dokumenten-Downloads beschreiben.

Überdies wiesen Bollen, Van De Sompel, Hagberg, & Chute[8] anhand einer Hautpkomponentenanalyse nach, dass Zitationen vergleichsweise wenig Einfluss auf wahrgenommenen Impact haben: „Our results indicate that the notion of scientific impact is a multi-dimensional construct that can not be adequately measured by any single indicator, although some measures are more suitable than others. The commonly used citation Impact Factor is not positioned at the core of this construct, but at its periphery, and should thus be used with caution." Den erwähnten Netzwerk- und Clickstream-Analysen bescheinigte man im selben Paper mehr Aussagekraft: "Usage-based measures such as Usage Closeness centrality may in fact be better 'consensus measures'." Besser noch: Bollen & Van De Sompel[9] beschrieben sogar, wie eine Architektur zur Sammlung und Aufbereitung von Nutzungsdaten aussehen könne.

Nutzungsinformationen erschienen daher aus dreierlei Gründen für die Open Access Community interessant:

- Wenn Downloadzahlen Zitationshäufigkeiten vorhersagen können (Brody, Harnad, & Carr[10]), so die erste Überlegung, dann erfassen sie Impact in gleicher Art wie Zitationen (nur eben früher) und können als eigene (im Idealfall kostenfreie) Impact-Information genutzt werden. Kurzum: Nutzungsimpact kann als alternative Impactquelle für wissenschaftliche Dokumente genutzt werden und so das Open-Access-Reputationsdefizit, das sich aus dem fehlenden Zitationsimpact ergibt, ausgleichen.
- Vertrauenswürdige Download-Statistiken, die höhere Nutzungszahlen von Open Access im Vergleich zu Closed Access ausweisen und so höhere Zitationsraten in Aussicht stellen, könnten Wissenschaftler dazu verführen, Open Access zu publizieren, um ihren Zitationsimpact zu steigern – selbst, wenn sie Downloadinformationen als Impact-Informationen geringschätzen.
- Die von Bollen at al.[11] geschilderten Vorzüge der Netzwerkanalyse von Nutzungsdaten skizzierten sogar Möglichkeiten, ausgefeilte Impact-Maße zu

8 Bollen, J., Van De Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. PloS One, 4(6), e6022. http://doi.org/10.1371/journal.pone.0006022 [Zugriff: 20.12.2017].

9 Bollen, J., & Van de Sompel, H. (2006). An architecture for the aggregation and analysis of scholarly usage data. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries – JCDL '06* (p. 298). New York, New York, USA: ACM Press. http://doi.org/10.1145/1141753.1141821 [Zugriff: 20.12.2017].

10 Wie Anm. 2.

11 Wie Anm. 6.

modellieren, z. B. in Form der oben erwähnten Zentralitätsmaße. Schließlich basierten diese Verfahren anders als Zitationszählungen oder Download-Counts nicht auf dem Zählen banaler absoluter Häufigkeiten, sondern waren diesen methodisch überlegen.

3 Das Projekt

Die Perspektive, die Akzeptanz von Open Access durch die Ermittlung nutzungsdatenbasierter Impact-Verfahren zu fördern, bewog im Jahr 2006 die Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB Göttingen), die Universitätsbibliothek Stuttgart (UBS), die Saarländische Universitäts- und Landesbibliothek (SULB) und den Computer- und Medienservice (CMS) der Humboldt-Universität zu Berlin, in die Planung eines Drittelmittelprojekts zur Gewinnung von standardisierten Nutzungsstatistiken einzusteigen. Nach Ausräumen einiger Bedenken zur Impact-Messung im Allgemeinen und zur nicht-zitationsbasierten Impact-Messung im Besonderen, startete im Mai 2008 das von der Deutschen Forschungsgemeinschaft (DFG) geförderte Projekt Open-Access-Statistik,[12] dessen offizielle Bezeichnung „Vernetzte Repositorien: Dienste und Standards für international vergleichbare Nutzungsstatistiken“ lautete. An die erste, im Dezember 2010 endende Projektphase schloss eine weitere an, die von April 2011 bis November 2013 reichte, und in der ein neuer, vorrangig mit Technik betrauter Projektpartner hinzukam, die Verbundzentrale des Gemeinsamen Bibliotheksverbundes GBV (VZG).

3.1 Phase eins: Ziele

Ziel der ersten Projektphase dieses Projekts war vorrangig die Entwicklung und Etablierung eines einheitlichen Standards zur Ermittlung von Zugriffszahlen und Nutzungsstatistiken für Publikationen sowohl in Open-Access-Repositorien als auch in Closed-Access-Angeboten, z. B. subskribierten E-Journals. Nutzungsdaten zu Letzteren sollten durch die Auswertung von Linkresolver-Logs gewonnen werden. Linkresolver sind Anwendungen, die bei einer Recherche, z. B. in einer Datenbank oder Suchmaschine, automatisch prüfen, ob ein Dokumentenzugriff aus einem Campus-Netz möglich ist.

12 https://dini.de/projekte/oa-statistik (DFG Projektnummer 72662563) [Zugriff: 20.12.2017].

Beide Datentypen sollten in einer Datenbank gesammelt werden, um eine Vergleichbarkeit der Zugriffsinformationen zu ermöglichen. Dies setzte eine Homogenisierung der Daten voraus, die von sehr unterschiedlichem Informationsgehalt waren.

Die Zugriffe auf die Open-Access-Dokumente sollten durch Auswertung der Webserver-Logfiles ermittelt werden. Die Reichhaltigkeit dieser Daten ist *prinzipiell* sehr hoch, da Zugriffsinformationen sehr detailliert geloggt werden können. Zudem liegen die Daten zumeist bei den Repositorien selbst vor. Diese Logs haben jedoch unterschiedliche Gestalt, zudem hängt die Granularität *faktisch* stark von lokalen Systemkonfigurationen ab. Diese, in den Webserver-Logs anfallenden Daten sollten mit den Linkresolver-Logs vereinheitlicht werden. Diese Logs wiederum sind allerdings nicht nur in ihrer Gestalt recht unterschiedlich, sie unterscheiden sich auch stark von den Webserver-Logs. Zudem werden die Linkresolver teils von Universitätsbibliotheken selbst gehostet, teils von kommerziellen Anbietern. Im zweiten Fall waren daher Abmachungen nötig, um auf die Logs zugreifen zu können. Ob eine Aggregation von Webserver- und Linkresolver-Daten machbar sein würde, wurde von den Projektpartnern selbst als Forschungsfrage betrachtet.

Absicht war es, durch die Kombination dieser beiden Verfahren eine größtmögliche Vollständigkeit der Dokumentnutzung zu erfassen. Die Webserver- und Linkresolver-Logs sollten aus den Diensten der Projektpartner (den Data Providern) gewonnen und in einer zentralen Datenbank (dem Service Provider) zusammengeführt werden. Dies erforderte die Definition und Schaffung von Schnittstellen zwischen den Data Providern und dem Service Provider. Die Software, mittels derer die Dokumentzugriffe auf einem Data Provider protokolliert, gespeichert und an den Service Provider ausgeliefert werden sollte, sollte generisch gestaltet sein, so dass sie in möglichst unterschiedlichen Systemen mit möglichst geringem Aufwand weiterverwendet werden können sollte. Der Service Provider selbst sollte mehrere Dienste und Funktionen erfüllen, u. a.

1. Dubletten-Erkennung (Zugriffe auf identische Dokumente bei verschiedenen Data Providern sollten summiert werden),
2. Nutzungsanalyse (Dokumentzugriffe als reine Häufigkeiten sowie Clickstream-Daten),
3. Identifikation von Benutzern (als Bedingung der Clickstream-Analyse).

Aufbauend auf diesen Funktionen sollten in einer zweiten Projektphase Mehrwertdienste entwickelt werden. Punkt drei verweist auf ein zentrales Arbeitspaket des Projekts, den Datenschutz. Ebenfalls in Projektphase eins sollte eine Sichtung möglicher Standards der Zugriffsmessung auf Online-Quellen erfolgen, um festzustellen, welche Daten die Data Provider bereitstellen sollten, um diese

Standards oder einen eigenen, auf Basis der ermittelten Referenz-Verfahren entwickelten Standard zu bedienen.

3.2 Die Standards

Die Projektgruppe identifizierte drei Referenz-Standards:
- COUNTER: Ein von Wissenschaftsverlagen entwickeltes Verfahren zur Messung von Zugriffen auf kostenpflichtig lizenzierte Dokumente,[13]
- LogEc: Ein vom Server-Netzwerk RePEc (Research Papers in Economics) entwickeltes Verfahren zur Messung der Zugriffe auf Open-Access-Dokumente in RePEc,
- IFABC: Ein von der Werbeindustrie entwickeltes Verfahren zur Reichweitenmessung von Online-Werbung.

Diese unterschieden sich vor allem durch die Definition von Doppelclick-Intervallen und der Verfahren zur Eliminierung maschineller Zugriffe. Keiner der Standards war auf eine Nutzer-Identifikation zur Erstellung von Clickstreams oder der De-Duplizierung von Dokumenten ausgelegt.

Was die Webserver-Logs anging, erwies es sich als recht einfach – eine passende Konfiguration vorausgesetzt – die Vorgaben der drei genannten Standards zu erfüllen. In der Folge wurden die technischen Voraussetzungen geschaffen, um die zur Anwendung der Standards benötigten Daten in den Webserver-Logs zu sammeln, sie daraus zu extrahieren und lokal zu speichern, zugleich wurden Schnittstellen entwickelt, um die Daten mit einem von der SUB Göttingen betriebenen Service Provider im Testbetrieb austauschen zu können. Parallel musste das Vorgehen von den Datenschutzbeauftragten der beteiligten Hochschulen bzw. Länder genehmigt werden.

3.3 Die technischen und rechtlichen Hürden

Leider zeigte sich früh, dass in Deutschland – anders als in den USA, in denen Bollen und Kollegen tätig sind – Zugriffe auf lizenziertes Material nicht in erheblichem Ausmaß über Linkresolver erfolgen: Es konnten nur wenige Daten aus den Linkresolvern gewonnen werden, auch war eine Nutzer-Identifikation mittels der

13 Zur Tauglichkeit COUNTERs für die Messung von Zugriffen auf Open-Access-Dokumente finden sich weitere Informationen im Abschnitt „Evaluierung von Standards“.

anfallenden Informationen nicht immer machbar. Zudem war es in den Fällen, in denen die Linkresolver bei Dienstanbietern und nicht beim Projektteilnehmer liefen, nicht ohne Weiteres möglich, Zugriff auf die Logs zu erhalten. Grund hierfür war der Datenschutz: Die Dienstanbieter hatten in ihren Nutzungslizenzen den Fall des Loggings und der Weitergabe von Nutzungsdaten nicht vorgesehen und scheuten davor zurück, Kunden durch diese sensible Thematik aufzuschrecken bzw. die Vermarktung ihres Produktes durch Einbeziehen von Datenschutzstellen zu verkomplizieren.

Allerdings erschwerte der Datenschutz auch die Verarbeitung der Webserver-Logs. Während einige der Datenschutzbeauftragten lediglich auf einer Pseudonymisierung der IP-Adressen inklusive Salting und Hashing[14] bestanden, machten andere weitergehende Vorgaben und diskutieren selbst die Pseudonymisierung sehr kontrovers. Insbesondere die Zentrale Datenschutzstelle der baden-württembergischen Universitäten (ZENDAS) hinterfragte das Projekt sehr kritisch, mit der Folge, dass die letztliche Umsetzung der Datenschutzvorgaben zu Ende der Projektphase eins noch in der Schwebe war.

3.4 Zwischenfazit

Die Ergebnisse des ersten Projektabschnittes waren:

- Man verzichtete auf die Integration von Linkresolver-Logs in die Architektur, u. a. wegen des geringen Volumens und der teils nicht gegebenen Verfügbarkeit.
- An allen Standorten war Software entwickelt, um lokal Daten zu erfassen, zu verwalten, zu pseudonymisieren und auszuliefern, die die erwähnten Standards bedienen konnten.
- Der im Testbetrieb angebotene Service Provider nahm Test-Daten entgegen und verarbeitete sie.
- Die Nutzungsbedingungen der lokalen Server erlaubten die Erhebung der Daten.
- Software und rechtliche Policies der lokalen Repositorien standen zur Nachnutzung bereit.
- Welche Informationen aus den Logfiles erfasst, gespeichert und weitergegeben werden konnten – ja, ob sie überhaupt weitergegeben werden durften – unterlag noch der Prüfung.

14 In diesem Szenario wird die IP mittels eines Strings (des Salt) erweitert und durch einen verschleierten Wert (den Hash) ersetzt.

3.5 Phase zwei: Ziele

Projektphase zwei hatte primär die folgenden Ziele ausgerufen:
- Erhöhung der Akzeptanz von Open Access bei Autoren und Rezipienten von wissenschaftlichen Publikationen durch Metriken und Mehrwertdienste,
- Kooperationen für international vergleichbare Nutzungsstatistiken,
- Anbieten einer nachhaltigen Service-Infrastruktur.

Punkt eins umfasste im Wesentlichen das Anbieten elaborierter Metriken oder auch eher handfester Mehrwertdienste wie z. B. Clickstream-basierter Recommender. Für beide Funktionen mussten Daten aus den verteilten Systemen aggregiert werden. Man ging davon aus, dass die datenschutzrechtliche Klärung Clickstream-Analysen (und die dazu benötigte Pseudonymisierung) ermöglichen würde, zumindest in einem gewissen Ausmaß. Sollte die Pseudonymisierung rechtlich nicht möglich sein, hatte man andere Anwendungsszenarien entwickelt, z. B.
- GeoIP-Auswertungen: Pro Dokument sollte es z. B. möglich sein, darzustellen, wo man sich denn für dessen Inhalt interessiert.
- Eine rein auf Nutzungshäufigkeiten basierende standardisierte Anzeige der Dokumentdownloads.

Nicht prominent erwähnt, aber sehr wichtig war die Rolle des GBV als neuem Projektpartner, der den von der SUB Göttingen als Testsystem betriebenen Service Provider operativ anbieten sollte.

3.6 Evaluierung von Standards

Die Evaluierung der Standards erfolgte durch Experten-Interviews. Das Ergebnis spiegelt einen gewissen resignierten Pragmatismus wider. Von den drei zur Diskussion gestellten Standards wurde das IFABC-Vorgehen als wenig tauglich für die Zugriffsmessung wissenschaftlicher Werke erachtet. Am besten *bewertet* wurde LogEc, das COUNTER sowohl in Sachen Doppelclick-Intervall, das bei COUNTER als zu kurz kritisiert wurde, als auch in Sachen Identifikation maschineller Zugriffe[15] überlegen war. Jedoch *empfahlen* die Experten die Anwendung

15 COUNTER setzt auf eine willkürlich erscheinende und kurze Robot-Liste, was auch dem Umstand geschuldet ist, dass der Standard entwickelte wurde, um Zugriffe auf Nicht-Open-Ac-

von COUNTER und nicht von LogEc, da letzterer als zu unbekannt und daher akzeptanzhemmend eingeschätzt wurde.

3.7 Datenschutz, die zweite

Das Ergebnis der datenschutzrechtlichen Prüfung[16] (Zentrale Datenschutzstelle der baden-württembergischen Universitäten ZENDAS, 2011) gab den Projektpartnern Empfehlungen zur Speicherung und Verarbeitung der gewonnen Daten und stellte sie vor eine größere Strategieproblematik. Bei vielen Aspekten existierte ein gewisser Interpretationskorridor zwischen einer sehr engen und voraussichtlich unantastbaren Auslegung der rechtlichen Normen und einer womöglich, aber nicht sicher, angreifbaren Interpretation. Erstere ermöglichte letztlich kaum innovative Auswertungen und Funktionalitäten, bedeutete aber weitgehende Rechtssicherheit; Zweitere lockte mit Features und reichhaltigen Metriken, setzte aber jede den Dienst nutzende Einrichtung rechtlichen Unwägbarkeiten aus. Die Projektgruppe entschied sich für die Option der Rechtssicherheit, in der Annahme, dass alleine die Vermutung rechtlicher Probleme die Attraktivität des Dienstes fundamental beschädigen könnte – ganz zu schweigen, von etwaigen konkreten Klagen. Faktisch bedeutete dies aber, dass u. a. folgende Ziele aufgegeben werden mussten:

1. die serverübergreifende Aggregation von pseudonymisierten Daten,
2. damit im Zusammenhang die Entwicklung von Clickstream-basierten Metriken und Recommendern,
3. die Eliminierung von mehrmaligen Zugriffen auf identische Dokumente auf verteilten Servern innerhalb des COUNTER-Doppelclick-Intervalls.
4. Weiterhin erschien laut ZENDAS-Gutachten auch die Auswertung anderer Informationen aus den Logs als angreifbar, z. B. des Referrer oder der GeoIP-Informationen, so dass weitere für eine Relevanzbewertung eines Dokumentes nützliche Informationen nicht ausgewertet werden konnten.

Nimmt man die bereits in Projektphase eins gestrichene Auswertung der Linkresolver-Logs hinzu, entfällt ein weiteres Projektziel: Der Vergleich der Nutzungs-

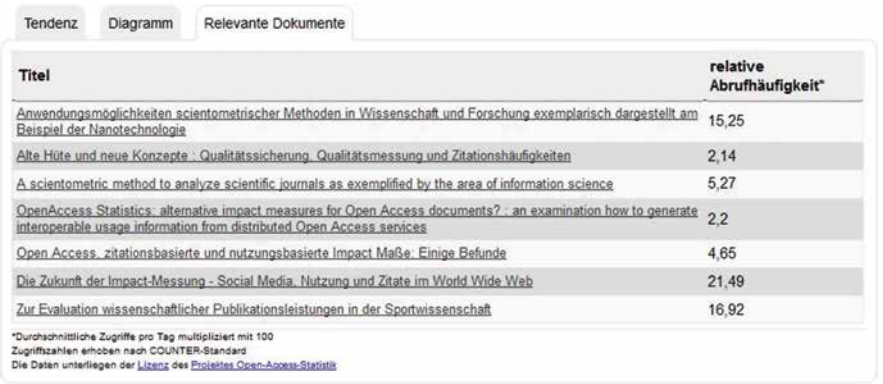
cess-Material zu messen. LogEc hingegen setzt neben einer ausgefeilteren Liste auch auf Data-Mining-Techniken zur Identifikation von nicht-menschlichen Zugriffen.

16 Zentrale Datenschutzstelle der baden-württembergischen Universitäten ZENDAS. (2011). *Datenschutzrechtliche Bewertung des Projekts „Open-Access-Statistik“*. https://dini.de/fileadmin/oa-statistik/gutachten/ZENDAS_Gutachten_2011.pdf [Zugriff: 20.12.2017].

zahlen von Open-Access- und Nicht-Open-Access-Dokumenten, der sich, kombiniert z. B. mit Zitationsinformationen, als für die szientometrische Forschung wertvoll hätte erweisen können.

3.8 Bilanz

In Phase zwei gelang es, neben der Evaluierung der Standards, den Service Provider beim GBV in Betrieb zu nehmen. Die lokalen Repositorien liefern anonymisierte Daten an den Provider, der sie nach COUNTER-Vorgaben aufbereitet und an die gebenden Repositorien zurückliefert. Dort werden sie als Metadaten abgespeichert und zusammen mit dem jeweiligen Dokument angezeigt. Der Limitierung der Robots-Liste von COUNTER wurde durch eine Erweiterung dieser in Abstimmung mit anderen Projekten entgegengewirkt. Teils wurden lokal Mehrwertdienste basierend auf den bereinigten Daten entwickelt, z. B. auf dem Repository SciDok[17] der SULB: Hier werden in einer Recommender-Funktion bei der Ansicht eines Dokumentes inhaltlich ähnliche Dokumente[18] inklusive deren Nutzungszahlen angezeigt (Abb. 1).


Tendenz | Diagramm | Relevante Dokumente

Titel	relative Abrufhäufigkeit*
Anwendungsmöglichkeiten scientometrischer Methoden in Wissenschaft und Forschung exemplarisch dargestellt am Beispiel der Nanotechnologie	15,25
Alte Hüte und neue Konzepte : Qualitätssicherung, Qualitätsmessung und Zitationshäufigkeiten	2,14
A scientometric method to analyze scientific journals as exemplified by the area of information science	5,27
OpenAccess Statistics: alternative impact measures for Open Access documents? : an examination how to generate interoperable usage information from distributed Open Access services	2,2
Open Access, zitationsbasierte und nutzungsbasierte Impact Maße: Einige Befunde	4,65
Die Zukunft der Impact-Messung - Social Media, Nutzung und Zitate im World Wide Web	21,49
Zur Evaluation wissenschaftlicher Publikationsleistungen in der Sportwissenschaft	16,92

*Durchschnittliche Zugriffe pro Tag multipliziert mit 100
Zugriffszahlen erhoben nach COUNTER-Standard
Die Daten unterliegen der Lizenz des Projektes Open-Access-Statistik

Abb. 1: Nutzungsdatenbasierte Empfehlungsfunktionen in SciDok.

17 http://scidok.sulb.uni-saarland.de [Zugriff: 20.12.2017].
18 Die Ähnlichkeitsbeziehung wird über benutzte Schlagworte ermittelt.

4 Eine Einordnung

Mit etwas Distanz fällt die Einordung des Projekts schwer: Einerseits wurde das im Projekt-Titel ausformulierte Ziel („Vernetzte Repositorien: Dienste und Standards für international vergleichbare Nutzungsstatistiken") ohne Abstriche erreicht. Es existiert ein funktionaler Dienst zur Ermittlung standardisierter Zugriffszahlen auf Repositorien.[19] Das im Titel des Projekts nicht ausformulierte, aber dennoch avisierte Ziel der Schaffung einer Infrastruktur zur Entwicklung Clickstream-basierter Metriken und Recommender konnte jedoch nicht erreicht werden. Ausschlaggebend hierfür war der Zwang, zwischen Funktionalitäten und Rechtssicherheit abzuwägen – hätte man sich für die Funktionalitäten und gegen die Rechtssicherheit entschieden, wäre der Dienst womöglich nicht mehr in Betrieb.

Angesichts der heute boomenden Altmetrics-Dienste ist Open-Access-Statistik auf ganz unterschiedliche Weise zu betrachten. Man kann darin einen sehr frühen Versuch sehen, Impact auf alternative Art zu messen.[20] Man mag das Projekt heute auch als etwas gestrig ansehen, da die Darstellung der Impact-Informationen auf den Repositorien sehr unterschiedlich ausfällt und gemessen an den poppigen Altmetrics-Visualisierungen, z. B. in Form des Metrics-Donuts des Anbieters Altmetric.com, recht hausbacken daherkommt.

Ganz sicher muss man dem Projekt zugutehalten, viele Informationen über die rechtlich-technische Machbarkeit der Datensammlung zu Zwecken der Impact-Bewertung gesammelt zu haben. Dies ist vielleicht das größte Verdienst des Projekts, auch wenn die Erkenntnisse ernüchternd sind: Die Datensammlung in Deutschland wäre für das Projekt z. B. rechtlich einfacher gewesen, hätte es ein kommerzielles Ziel verfolgt[21]. Zudem konnte man die Lehre daraus ziehen, die Daten am besten nicht selbst zu erheben, sondern externe Quellen zu benutzen, die – qua nationaler Rechtsprechung – geringere datenschutzrechtliche Barrieren kennen, so wie es die erwähnten Altmetrics-Dienste tun, wenn sie z. B. die API eines Webdienstes wie Twitter nutzen.

Zugleich war Open-Access-Statistik bescheidener als die Altmetrics-Dienste, war es doch erklärtes Ziel, nur standardisierte Daten bereitzustellen, auf deren Basis Impact-Metriken ermittelt werden sollten und nicht eigene Metriken, deren

19 Aktuell nutzen 14 Repositorien den Dienst.

20 Z. B. nicht-zitationsbasiert sowie unabhängig von Dokumenttyp und -sprache.

21 § 96 des Deutschen Telekommunikationsgesetzes (TKG) erlaubt die Erfassung von Nutzungsdaten zur Verarbeitung von kundenspezifischen Informationen (z. B. Kundennummern), die bei Open-Access-Diensten obsolet sind.

methodisches Fundament im Falle von Altmetrics mehr als fragwürdig ist (Herb[22], Herb[23]). Überdies – das ist bei allem Chique der Altmetrics das Plus von Open-Access-Statistik – sind die bereitgestellten Daten standardisiert und frei zugänglich.

Dr. Ulrich Herb
Saarländische Universitäts- und Landesbibliothek
Postfach 15 11 41
66041 Saarbrücken
Deutschland
E-Mail: **u.herb@sulb.uni-saarland.de**

22 Herb, U. (2016). Altmetrics zwischen Revolution und Dienstleistung : Eine methodische und konzeptionelle Kritik. In H. Staubmann (Ed.), *Soziologie in Österreich – Internationale Verflechtungen. Kongresspublikation der Österreichischen Gesellschaft für Soziologie* (pp. 387–410). Österreichische Gesellschaft für Soziologie ÖGS. http://doi.org/10.15203/3122-56-7 [Zugriff: 20.12.2017].
23 Herb, U. (2016). Impactmessung, Transparenz & Open Science. *Young Information Scientist*, *1*. http://doi.org/10.5281/zenodo.153831 [Zugriff: 20.12.2017].